PRESCALE: EXACT RISK ACCOUNTING FOR FROZEN-BANK VERIFIER SEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

Verifier-guided Best-of-N search can improve a proxy score while replacing an incumbent with a target-worse candidate. We study this risk through offline auditing of a frozen labeled candidate bank, with target labels hidden from the selector. PRESCALE exactly integrates harmful replacement over uniform subsets sampled without replacement, including competition between harmful and benign winners. A complete score-gap vector yields a conditional risk curve without a Gaussian assumption; independent complete vectors from a specified observation law additionally support finite-sample budget calibration. Candidate-wise verifier reliability does not identify this curve. Matched-margin quadratic-Gaussian controls isolate dependence effects. On one simulated PDE trajectory, aligned score gaps yield forecast MAE of 1.11 percentage points, versus 19.68 after column shuffling preserves empirical marginals; same-information subset Monte Carlo yields 1.15. In a controlled matched-data comparison, direct-loss calibration admits budget 16 versus 4 for a marginal union-bound reduction; known Gaussian marginals also admit 16. This comparison includes union-bound slack. On MATH-500, the rank interface quantifies harmful replacements alongside accuracy gains and separates cross-bank forecast discrepancies from within-bank sampling effects. This operational audit does not establish a distinct answer-selection advantage. PRESCALE makes frozen-bank search risk computable while separating exact conditional accounting, dependence-sensitive estimation, and statistical calibration.

1 INTRODUCTION

For a frozen bank of candidate solutions, an auditor may need to know how often expanding verifier-guided search replaces an incumbent with a higher-scoring but target-worse candidate. This reference-relative event is not the selected-answer error rate: aggregate accuracy gains can coexist with harmful replacements. We study an offline auditor with target labels and a selector that cannot access them. The task is to evaluate this label-blind policy, not replace it with oracle filtering.

PRESCALE makes this audit computable through an exact finite-bank subset-risk functional (Figure 1). Given a complete score-gap vector and harmful mask, Proposition 2 integrates uniform without-replacement sampling exactly, including harmful–benign winner competition and fixed tie-breaking. Averaging this bounded functional over independent complete vectors from a specified observation law permits separate statistical calibration. A deterministic vector instead gives exact conditional accounting for that bank, not transfer to new tasks or hypothetical verifier realizations.

Candidate-wise crossing probabilities do not identify this risk curve. Fully shared and independent crossings can have identical marginals but different full-bank union probabilities (Proposition 1). This motivates retaining dependence information rather than assuming candidate-level reliability determines search-level behavior. Quadratic-Gaussian geometry provides a controlled illustration, not an assumption of the rank functional.

This is a different random experiment from IID-response ROC characterizations (Dorner et al., 2026): a frozen bank shares one verifier realization and is sampled without replacement. Marginal crossing probabilities still give conservative upper bounds. What they do not determine is the exact replacement-risk curve or its largest admissible budget.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

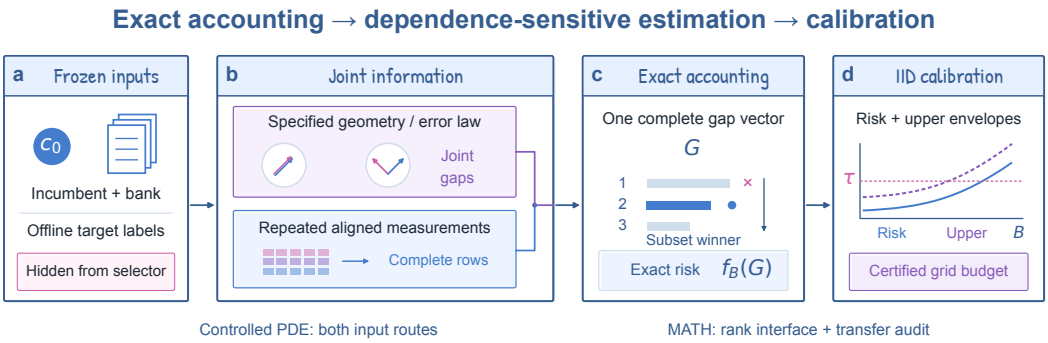

Figure 1: **Exact accounting, dependence-sensitive estimation, and calibration.** A specified geometry/error law or repeated raw measurements supplies complete aligned score gaps on one frozen bank. Proposition 2 integrates uniform subsets exactly; independent complete vectors support calibrated grid budgets. Icons and envelope curves are schematic. Offline labels remain hidden from the selector. MATH audits observed ranks and cross-bank transfer.

Fix an incumbent c_0 and a bank D of M challengers with offline target labels. At budget B, the label-blind selector chooses the best verifier score among c_0 and a uniform size-$(B-1)$ subset of D. The harmful-winner risk $\mathcal{R}_B$ averages over this subset and a shared verifier realization from an independently supplied or calibrated law. We seek the largest budget whose reference-relative risk remains below a chosen tolerance. For a predeclared grid $\mathcal{B} \subseteq \{1, \dots, M+1\}$ containing 1, define

$$\overline{\mathcal{R}}_B^{\mathcal{B}} = \max_{b \in \mathcal{B},\, b \le B} \mathcal{R}_b, \qquad \mathsf{Cap}_{\tau,\mathcal{B}}^{\mathrm{safe}} = \max\{B \in \mathcal{B} : \overline{\mathcal{R}}_B^{\mathcal{B}} \le \tau\}. \tag{1}$$

Every capacity is conditional on the reference, labeled bank, sampling policy, and error law; we suppress these fixed arguments. The full-integer definition sets $\mathcal{B} = \{1, \dots, M+1\}$. This envelope controls each evaluated fixed-budget risk, not the probability of any failure along a run or an arbitrary score-dependent stopping rule.

The independent homogeneous special case has classical Gaussian-extreme growth of order $\sqrt{\log B}$ (Boucheron et al., 2013; Vershynin, 2018); the finite-bank estimator retains non-homogeneous margins and dependence.

We make four contributions:

- **Exact conditional accounting.** Proposition 2 integrates uniform without-replacement subsets, harmful–benign winner competition, and fixed tie-breaking for any complete score-gap vector.
- **Dependence-sensitive risk estimation.** A non-identifiability construction and matched-margin controls show why marginals are insufficient; an aligned-observation ablation tests the resulting risk interface.
- **Finite-sample calibration.** Bounded row losses feed standard risk-control procedures. Matched data and tests quantify finite-sample certification under two risk reductions, without attributing the contrast solely to dependence.
- **Operational audits and limits.** PRM rankings expose harmful replacements and cross-bank transfer discrepancies. A CPU comparison measures conditional accounting cost; distinct policy benefit remains unestablished.

2 Related Work

Reward overoptimization and inference-time search. Proxy failures motivate reward-overoptimization analysis (Manheim & Garrabrant, 2018; Amodei et al., 2016; Gao et al., 2023) and reward-model ensembles (Coste et al., 2024). Table 1 distinguishes the random experiments and outputs of the closest analyses. Our distinction is incumbent-preserving, without-replacement selection from a frozen labeled bank, evaluated by reference-relative harmful-winner risk. The rank functional integrates subset randomness exactly; calibration over the gap law is separate. Hedging and rejection require policy-specific weights. The covariance in Afrin & Shihab (2026) is across

judges, not candidate–reference gaps; their conditional tail bounds also cover predictable adaptive proposals. Egashira et al. (2026) isolate systematic error patterns in RLVR training; our diagnostic instead fixes the bank and estimates inference-time selection risk.

Verification and risk control. Outcome and process verifiers improve mathematical reasoning (Cobbe et al., 2021; Uesato et al., 2022; Lightman et al., 2024; Wang et al., 2024); prover–verifier games and learned judges study related checking channels (Kirchner et al., 2024; Zheng et al., 2023). Selective classification chooses whether to answer (Geifman & El-Yaniv, 2017), whereas our diagnostic fixes the selector and varies its budget. Composing exact subset integration with bounded-loss calibration makes a certified budget grid computable while retaining harmful-versus-benign winner competition and dependence-sensitive risk. The ingredients—order-statistic counting, conditional expectation, and concentration—are classical; the audit interface specifies their inputs and the replacement event they control.

Verifiable environments and scientific search. Detectable reward-hacking environments (Roth et al., 2026) and symbolic equation discovery (Brunton et al., 2016; Rudy et al., 2017; Messenger & Bortz, 2021; Udrescu & Tegmark, 2020; Petersen et al., 2021; Cranmer, 2023) separate candidate correctness from proxy fit. Language models can also propose scientific structures (Xie et al., 2026; Du et al., 2026; Pekhale et al., 2026). We use an exactly labeled PDE bank to isolate the correlation mechanism. "Before execution" means before a fresh selection from a known bank, after its labels, sampling policy, and calibration data are available; it does not mean before generating or labeling that bank.

3 The PRESCALE Diagnostic

3.1 Problem setup and inputs

The geometry layer models joint score gaps; the rank layer also accepts observed score vectors directly, without a Gaussian assumption.

Fix a reference c_0, a challenger bank D, and a target loss T (lower is better). Offline labels define $D_0^- = \{c \in D : T(c) > T(c_0)\}$. They belong to the auditor, not the selector. Transfer from this bank to new unlabeled candidates is a separate calibration problem.

A shared realization ω defines scores on the full bank, independently of the uniformly sampled subset. Let P_G be the joint law of the complete gap vector $G = (G_c)_{c\in D}$, where $G_c = V(c;\omega) - V(c_0;\omega)$. The selector is

$$\widehat{c}_B = \arg\min_{c\in\{c_0\}\cup S_B} V(c;\omega), \qquad S_B \sim \mathrm{Unif}\{S \subseteq D : |S| = B-1\}. \tag{2}$$

Define the reference-relative harmful-winner event and its risk by

$$\mathcal{H}_B = \{V(\widehat{c}_B;\omega) < V(c_0;\omega),\ T(\widehat{c}_B) > T(c_0)\}, \quad \mathcal{R}_B(P_G; D, c_0) = \Pr(\mathcal{H}_B). \tag{3}$$

Table 1: Different search experiments require different risk information. Prior ensemble guarantees already allow dependent candidates; our distinction is the reference-relative winner event and its exact conditional accounting.

Analysis	Information and sampling	Output
Dorner et al. (2026)	Full ROC and prevalence; IID responses	Best-of-N accuracy
Khalaf et al. (2025)	Proxy–target pairs and empirical quantiles; IID generation	BoN/SBoN/BoP reward curves; HedgeTune parameters
Afrin & Shihab (2026)	Judge-error covariance and joint tail bounds; dependent candidates allowed	Finite-search selected overstatement and regret bounds
PRESCALE	Labeled frozen bank and joint gaps; uniform subsets without replacement	Exact subset-risk functional and calibrated budget grid

Here $1 \le B \le M+1$ counts the incumbent as well as the $B-1$ challengers. The maximized score in the introduction is $-V$. The incumbent wins score ties; challenger ties use a fixed index order independent of labels. Fresh per-query errors and score-adaptive proposals define different joint experiments.

Proxy improvement alone is not hacking; the selected candidate must be independently target-worse. No intentional behavior is required. We suppress P_G, D, c_0 when fixed.

3.2 Exact geometry reduction

Quadratic scores $V(c;e) = \|r_c - e\|_2^2$ and a shared $e \sim \mathcal{N}(0,\Sigma)$ specify one joint gap law. Here $r_c \in \mathbb{R}^p$ is a supplied ideal residual vector in whitened measurement coordinates; the PDE study constructs it from weak-form discrepancies (Section 4.1). Define the pairwise direction $\Delta_c = r_c - r_{c_0}$ and noiseless proxy gap $\gamma_c = \|r_c\|_2^2 - \|r_{c_0}\|_2^2$. Expanding the quadratic scores gives, for any reference,

$$V(c;e) < V(c_0;e) \iff \Delta_c^\top e > \tfrac{1}{2}\gamma_c. \tag{4}$$

The controlled PDE environment uses the special case $r_{c_0} = 0$, but neither Equation 3 nor Equation 4 requires it.

For every challenger, harmful or benign, set $s_c^2 = \Delta_c^\top \Sigma \Delta_c$. When $s_c > 0$, define

$$m_{c|0} = \frac{\gamma_c}{2s_c}, \qquad C_{cc'} = \frac{\Delta_c^\top \Sigma \Delta_{c'}}{s_c s_{c'}} \quad (s_c s_{c'} > 0). \tag{5}$$

If $s_c = 0$, the gap is deterministic, $G_c = \gamma_c$: it beats the incumbent iff $\gamma_c < 0$. We do not divide by zero or assign a fictitious correlation to this coordinate. Write $p_c = \overline{\Phi}(m_{c|0})$ for positive variance and $p_c = \mathbf{1}[\gamma_c < 0]$ otherwise. In a mixed bank, the full gap scales s_c and gaps γ_c, including benign candidates, determine the winner; standardized margins and correlations alone describe the all-harmful crossing reduction.

For a fixed evaluated set with distinct candidate gaps, the score gaps $G_c(e) = V(c;e) - V(c_0;e) = \gamma_c - 2\Delta_c^\top e$ are jointly Gaussian. The exact event is

$$\min_{c \in S \cap D_0^-} G_c(e) < \min\left\{0, \min_{c \in S \setminus D_0^-} G_c(e)\right\}. \tag{6}$$

Empty minima equal $+\infty$. Equation 6 applies when challenger gaps are distinct. With ties, a failure occurs iff the lexicographically best challenger has $G_c < 0$ and belongs to D_0^-. Equation 10 implements this rule, including competition from benign challengers.

In the all-harmful case $S \subseteq D_0^-$ with $s_c > 0$ for every $c \in S$, including our exhaustive PDE bank, the probability reduces to

$$\mathcal{R}(S,\Sigma) = \Pr\left(\max_{c \in S}\{Z_c - m_{c|0}\} > 0\right) = 1 - \Phi_{|S|}(m_S; C_S), \tag{7}$$

where $Z_S \sim \mathcal{N}(0, C_S)$ and Φ_k is a k-dimensional joint lower-tail probability. If an all-harmful set contains a deterministic negative gap, its risk is 1. Otherwise remove deterministic nonnegative gaps and apply the CDF to the positive-variance coordinates; if none remain, the risk is 0. Equations 6–7 are the key reduction: risk depends on the joint score-gap law and target labels, not on B in isolation.

Proposition 1 (marginals do not identify search risk). For K harmful challengers with common margin m, fully correlated projections give $\mathcal{R} = \overline{\Phi}(m)$, whereas independent projections give $\mathcal{R} = 1 - \Phi(m)^K$. Both have identical per-challenger crossing probability, but the latter approaches one with K. Thus marginal verifier quality, including a one-candidate ROC point, cannot identify Best-of-B risk. Appendix B gives the construction.

A directly computable certificate follows from the union bound:

$$\mathcal{R}(S,\Sigma) \le \min\left\{1, \sum_{c \in S \cap D_0^-} p_c\right\}, \tag{8}$$

where $\overline{\Phi} = 1 - \Phi$. The bound is deliberately one-sided: a small value certifies low displacement risk without assuming independent candidate errors. For uniform size-$(B-1)$ subsets, marginal information alone therefore gives $\mathcal{R}_B \le \min\{1, (B-1)M^{-1}\sum_{c\in D_0^-} p_c\}$. This can certify a conservative budget without identifying the exact risk curve.

3.3 Exact finite-bank rank identity

For any realized complete gap vector g, let

$$W_c(g) = \sum_{c' \in D} \mathbf{1}[(g_{c'}, i_{c'}) >_{\text{lex}} (g_c, i_c)] \tag{9}$$

be the number of challengers ranked below c. Lower fixed indices win challenger ties.

Proposition 2 (exact finite-bank subset-risk identity). Fix an incumbent, a frozen bank of M challengers, and an offline harmful mask. Sample $B-1$ challengers uniformly without replacement, independently of $G \sim P_G$. Let the incumbent win score ties and use a fixed, label-independent challenger tie order. For any complete gap vector, the conditional harmful-winner probability is the sum below; for any joint law P_G and $2 \le B \le M+1$,

$$\begin{aligned} \mathcal{R}_B(P_G; D, c_0) &= \mathbb{E}_{G\sim P_G}[f_B(G)], \\ f_B(g) &= \sum_{c\in D_0^-} \mathbf{1}[g_c < 0] \tbinom{W_c(g)}{B-2} \big/ \tbinom{M}{B-1}. \end{aligned} \tag{10}$$

Set $f_1 = \mathcal{R}_1 = 0$ and impossible binomial coefficients to zero. *Proof.* A harmful c wins exactly when $G_c < 0$, it is included, and the other $B-2$ sampled challengers come from its $W_c(G)$ lower-ranked alternatives. These disjoint winner events have the displayed combinatorial weights. Averaging over P_G proves the result (Appendix B).

The identity is distribution-free with respect to the score-gap law. Gaussian geometry is one way to generate complete gap vectors, not a requirement of rank accounting. With every challenger harmful, let $K(G) = \sum_c \mathbf{1}[G_c < 0]$; then $f_B(G) = 1 - \tbinom{M-K(G)}{B-1}\big/\tbinom{M}{B-1}$. Thus the all-harmful curve needs the crossing-count distribution, not a fitted full joint density: marginals determine $\mathbb{E}[K]$, but generally not the law of K. Mixed banks additionally require the harmful–benign winner ranking. One set of gap draws can evaluate any declared budget grid. A uniform permutation of D, with its first $B-1$ entries defining S_B, couples budgets as nested pools. The identity itself gives marginal risks and requires no such coupling. For an all-harmful bank risk is nondecreasing; in a mixed bank a benign winner can shield the incumbent, so the running maximum is necessary (Appendix I). A sparse grid does not check all intermediate budgets in a mixed bank.

Exact integration is also conditional variance reduction: if H_B is the binary winner event and $f_B(G) = \mathbb{E}[H_B \mid G]$, averaging f_B removes subset-sampling noise while retaining uncertainty across complete score vectors by classical Rao–Blackwellization (Appendix D.3).

3.4 Finite-sample budget calibration

Freeze the bank, labels, policy, and budget grid independently of the calibration sample. For IID complete gap vectors $G^{(1)}, \dots, G^{(n)} \sim P_G$, let $\widehat{\mathcal{R}}_B = n^{-1}\sum_i f_B(G^{(i)})$, where the integrand f_B in Equation 10 lies in $[0,1]$. For J nontrivial evaluated budgets, jointly with probability at least $1-\delta$,

$$\mathcal{R}_B \le U_B(\delta) := \min\left\{1, \widehat{\mathcal{R}}_B + \sqrt{\tfrac{\log(J/\delta)}{2n}}\right\} \ (B>1), \qquad U_1 = 0. \tag{11}$$

Replacing the risks in Equation 1 by these upper bounds gives a certified grid capacity, $\mathsf{Cap}^{\text{cert}}_{\tau,\mathcal{B}}$. Independence is across complete vectors, not across candidates or budgets within a vector. Selecting the bank or policy with these same errors is not covered by this bound. Appendix D gives the proof. Equation 11 is one simultaneous-band option. Table 2 instead uses prespecified increasing-budget KL fixed-sequence tests, which certify the passed grid prefix, not a simultaneous band (Appendix D.1).

Using the audit in practice. Freeze the incumbent, bank, audit labels, selection rule, and budget grid. For a deterministic scorer, one complete gap vector gives exact subset-sampling risk and a bank-conditional admissible grid budget through Proposition 2, without a Gaussian model. This is not a statistical guarantee over unobserved verifier variability or new task banks. Calibration instead requires independent complete rows from the same-bank observation law, bounded row losses, and a prespecified rule. Report that law, the admitted grid budget, tolerance, confidence level, and calibration cost. Different-question rankings assess cross-bank transfer, not repeated errors on one bank; new tasks, proposal policies, or verifiers require separate assessment.

Scope of the inputs. Gaussianity enters the geometric model, not the rank functional or bounded-loss calibration. Student-t_8 changes the supplied error law; the simulated measurement audit constructs gaps from noisy observations; PRM studies use observed rankings. These test distinct interfaces, not a fitted law for native PRM uncertainty. Sampling from a fitted model does not cover model-fitting error, and perturbations define a new observation law (Appendix H.4).

Independent homogeneous limit. The critical-margin scale $\sqrt{2 \log B}$ requires independent projections and common margins; it need not hold for correlated banks (Appendix B.4).

4 Evaluation

4.1 Controlled structural-label environment

The one-dimensional conservative PDE bank provides exact structural labels and 64-dimensional weak-form residual vectors. Canonicalization and admissibility checks leave the reference and 703 structurally unequal challengers, all harmful, with $r_{c_0} = 0$. A structurally wrong correction may agree at some metadata settings; the bank construction and worked example are in Appendices C and H.4. The geometry intervention preserves each candidate's residual norm, hence its marginal crossing probability, while changing joint dependence. The all-harmful bank isolates dependence effects; it does not demonstrate a gain in target quality from increasing search.

We impose isotropic Gaussian error $\Sigma = \sigma^2 I_{64}$, giving $m_{c|0} = \|r_c\|_2/(2\sigma)$. The reference scale $\sigma_0 = 0.0443171$ is an RMS over development weak residual coordinates; the construction and its distinction from independent calibration draws are given in Appendix H.5.

4.2 Two hundred seventy evaluation regimes

We cross ten evaluated budgets $B \in \{2, 4, 8, 16, 32, 64, 128, 256, 512, 704\}$ with nine noise multipliers $\sigma/\sigma_0 \in \{0.35, 0.5, 0.7, 1, 1.4, 2, 2.8, 4, 5.6\}$. Our decisive intervention isolates correlation. We evaluate three candidate geometries with exactly the same 703 candidate norms: the empirical residual bank, independently randomized directions, and directions clustered around one shared axis with mixing coefficient 0.9. Every one-candidate hacking probability is therefore identical across the three banks; only the joint correlation matrix in Equation 5 changes.

Across 246,753 candidate pairs, the median absolute projection correlation is 0.5868 (empirical), 0.0850 (randomized), and 0.8113 (clustered). Every correlation matrix has rank at most 64; independently sampled directions do not imply 703 independent projections. This randomized-direction control is distinct from the independent model of Equation 15.

For each regime, PRESCALE averages 20,000 simulated error draws with the schedule integrated exactly. The evaluation uses 50,000 independent error draws and, conditional on each draw, samples a Bernoulli event using the combinatorial hit probability. The resulting 13.5 million event draws across 270 regimes test Monte Carlo agreement under the same specified error law; they are not 13.5 million explicit subset-and-winner executions. A separate direct-subset oracle enumerates 19,440 subset selections over all 81 gap vectors in $\{-1, 0, 1\}^4$, agreeing with the integrated form to 1.11×10^{-16} (Appendix H.2).

Baselines. Appendix H.3 specifies the point-estimation ablations; certificates compare admissible budgets, not RMSE. Appendix D.3 fixes scores and labels for the same-information comparison.

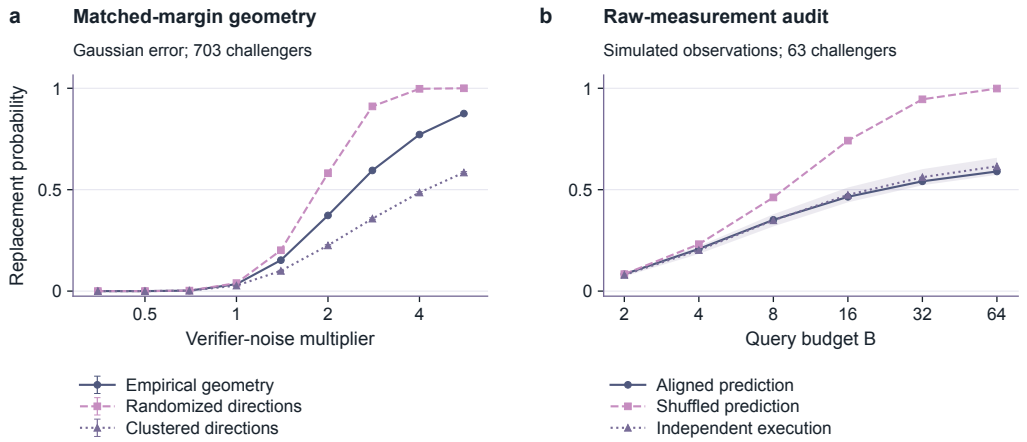

Figure 2: **Dependence matters at two distinct audit interfaces.** (a) Full-bank risk across all nine noise scales, holding 703 candidate margins fixed across geometries; bars are ± 1.96 event Monte Carlo standard errors. (b) Frozen aligned and column-shuffled forecasts versus fresh explicit selection on a separate 63-challenger bank. The band is a pointwise approximate 95% interval over 512 independent complete evaluation rows, conditional on the one simulated PDE trajectory. Prediction lines condition on 256 calibration rows; the shuffled line averages 32 column permutations.

5 Results

5.1 Dependence in geometry and measured score gaps

Figure 2a changes only angular correlation while fixing candidate norms, marginal hacking probabilities, budgets, and verifier-noise scales. At the full bank ($B = M + 1 = 704$) and $5.6\times$ noise, randomized directions reach simulated risk 1.000, the empirical geometry reaches 0.875, and clustered directions reach 0.585. PRESCALE predicts 1.000, 0.877, and 0.584, respectively. The panel shows all nine noise levels at the full bank; the complete 270-regime numerical check appears in Appendix Figure 5.

The empirical bank lies toward the high-risk end of these three controls, not the clustered endpoint. At reference noise, all three remain within 10% risk at the full budget (0.034, 0.039, and 0.027 for empirical, randomized, and clustered directions); the separation grows over the tested noise scales (Figure 2a). In Proposition 1's all-harmful construction, fully shared crossings are safer than independent ones at fixed margins. Mixed banks also involve benign-winner competition, so their risk cannot be ordered by crossing correlation alone.

Aligned observations preserve dependence-sensitive risk information. Figure 2b starts one level earlier: 1% noise is added to raw measurements of one fixed simulated PDE trajectory, then nonlinear weak moments and scores are recomputed on a preselected bank of 63 harmful challengers. From 256 complete calibration vectors, aligned-gap forecasts have 1.11 percentage-point MAE across six nontrivial budgets against fresh explicit selection on 512 evaluation realizations, versus 19.68 after column-wise shuffling preserves empirical marginals but breaks alignment. Same-information subset Monte Carlo gives 1.15 points: this contrast tests removal of alignment, not predictive superiority over Monte Carlo. The six budgets are correlated summaries of one audit; the 32 column permutations are not independent observation datasets. The additive certificate admits only $B = 1$. This audit tests the observation-to-risk interface on a simulated trajectory without fitting a Gaussian gap law (Appendix D.2).

Appendix Table 4 contrasts information loss: independent marginals discard dependence, budget-only calibration also loses noise-scale information, and the nearest-margin bank replaces every margin by the most vulnerable one. The union bound instead supplies a conservative certificate.

Table 2: Finite-sample certification at $2.8\times$ reference noise, $\tau = .1$ and $\delta = .05$. The first two rows share 256-row replication IDs, the budget grid, and KL fixed-sequence structure. Budget B includes the incumbent.

Risk reduction	Calibration information	Rule	Admitted grid budget
Exact conditional loss $f_B(G)$	Same 256 complete rows	KL fixed sequence	16 (32/32 runs)
Marginal union bound	Crossing mean, same 256 rows	KL fixed sequence	4 (32/32 runs)
Marginal union bound	Known Gaussian marginals	Known-law bound	16, deterministic

5.2 Finite-Sample Certification and Risk Reduction

Appendix Figure 4 reports point-estimated capacity at 10% risk on the evaluation grid, augmented by $B = 1$. At reference noise the empirical full bank is admissible; at multipliers 2, 2.8, and 5.6, the largest safe grid points are 64, 32, and 4. In this all-harmful bank, monotonicity also controls smaller integer budgets. Mixed banks require checking their intermediate risks. Certification has a separate cost: at $2.8\times$ noise and $n = 20{,}000$, point/reference capacity 32 becomes certified capacity 16.

Table 2 compares two risk reductions using the same 256-row replications, grid, confidence level, and increasing-budget KL fixed-sequence tests, stopping at the first failure. Direct calibration of the exact loss admits 16 in all 32 replications; the marginal union-bound reduction admits 4. Analytically known Gaussian marginals also admit 16. Direct-loss calibration thus matches the known-marginal comparator at this operating point; the contrast does not isolate dependence from union-bound slack. This post-hoc matched-rule reanalysis uses saved sufficient statistics, without resimulation. The marginal fixed-sequence and earlier upper-bound rules are dual representations, not independent replications (Appendix D.1).

5.3 Calibration, computation, and model sensitivity

Finite-sample estimation and the choice of confidence bound have distinct costs. Across eight replicates and 30 budget–noise regimes, mean risk RMSE decreases from 0.0137 at $n = 256$ to 0.0018 at $n = 20{,}000$ (Appendix D). Appendix Table 8 holds n fixed: bounded-loss tests following Angelopoulos et al. (2025) recover a nontrivial budget where the additive radius exceeds the tolerance. Table 2 instead compares risk reductions with matched data and tests. The independent reference risk at $B = 16$ is 0.04784, with 99% simultaneous interval $[0.04208, 0.05360]$ (Appendix D.1).

Computational value of exact integration. On the 276 cached MATH banks below, exact risk accounting takes 0.251 ms for all four budgets, versus 0.190 ms for one batched Monte Carlo subset per bank, whose task-risk RMSE is 7.49 percentage points. At 1,024 subsets, two direct-MC implementations take 114–160 ms and retain 0.22-point RMSE. This post-hoc CPU comparison isolates subset sampling variance; it excludes model generation and scoring (Appendix D.3).

Two further controlled studies probe the input law. Changing checkpoint-conditioned proposal support requires recalibration (Appendix E); under simulated Student-t_8 errors, matching-law integration gives RMSE 0.00185, versus 0.01493 for a Gaussian plug-in over 36 regimes (Appendix A).

5.4 Operational rank-layer audit on MATH-500

We test the rank layer on MATH-500 (Hendrycks et al., 2021; Hugging Face H4, 2024), using Qwen2.5-Math-7B-Instruct (Yang et al., 2024) and Skywork-o1-Open-PRM-Qwen-2.5-7B (He et al., 2024). Each question has 64 sampled solutions: the highest-PRM answer among the first eight is the reference, leaving 56 challengers.

Table 3: Matched-target MATH risk accounting (%). d_N is frozen on 224 development questions; c_N and p_N evaluate the same 276 confirmation banks. All rates use all-task denominators and the first-eight winner as reference.

N	Development d_N	Confirmation rank c_N	Prefix p_N
16	1.4027	.7795	.3623
32	2.0015	1.0491	1.0870
64	2.6786	.7246	.7246

We use 224 questions for development and 276 for confirmation; forecasts, selection code, and budget rules are frozen before confirmation evaluation. Generation budgets $N \in \{8, 16, 32, 64\}$ map to theoretical $B = 1 + (N - 8) \in \{1, 9, 25, 57\}$. Test correctness is used only for evaluation. Appendix F gives the protocol; the earlier 512-question GSM8K audit likewise finds harmful displacement while aggregate accuracy improves (Appendix G).

Reward, correctness, and replacement. Across the same 276 questions, mean selected PRM rises from 0.82346 at $N = 8$ to 0.85153 at 64, while correct answers rise from 247 to 250 and oracle availability from 261 to 268. That net gain comprises five corrections and two harmful replacements. At $N = 32$ there are three harmful replacements, falling to two at 64: benign winners can shield a mixed bank, so aggregate improvement and reference-relative risk coexist. The full replacement counts are in Appendix Table 14.

Separate transfer from the sampling target. Table 3 places the frozen development forecast d_N beside the confirmation-bank rank integral c_N and recorded-prefix risk p_N. The identity $d_N - p_N = (d_N - c_N) + (c_N - p_N)$ separates cross-bank transfer discrepancy from the within-bank subset-average versus prefix discrepancy. At $N = 64$ the entire remaining bank is selected, so $c_N - p_N = 0$ exactly: the residual forecast error is cross-bank. Confirmation labels enter only this evaluation, never budget selection. The rank, lookup, and Monte Carlo net-gain rules all select eight, whereas the point risk-capacity rule selects 64; this audit does not establish a distinct policy benefit.

6 Conclusion

PRESCALE provides exact conditional accounting of harmful replacement under uniform subset search on a frozen labeled bank. Aligned complete score gaps support dependence-sensitive estimation; independent observations from a specified law permit finite-sample calibration. Matched-margin controls and the simulated alignment ablation illustrate what marginals leave unresolved. At the tested sample size, direct-loss calibration admits a larger budget than the specified marginal union-bound reduction and matches the known-marginal comparator; this does not isolate dependence from bound slack. PRM audits reveal harmful replacements alongside aggregate gains, while distinct policy benefit remains unestablished. The contribution is a reusable audit functional with explicit information and calibration requirements.

Limitations

Capacity is conditional on the reference, labeled bank, sampling policy, target channel, and joint error law; changes call for recalibration. Scalar verifier scores do not identify latent covariance: aligned complete score-gap vectors require repeated verifier realizations on one frozen bank (Appendix H.4). All 703 PDE challengers are harmful; the LLM evaluation exercises mixed-bank winner competition but only the ranking layer. The LLM rates are operating-point measurements, not controlled comparisons of model strength. The additive certificate is conservative; fixed-sequence calibration retains power on the same controlled data (Appendix D.1). The held-out MATH comparison establishes neither accuracy superiority nor noninferiority. Appendix I details these conditions, the calibration cost, and the limits of transfer.

AI USE STATEMENT

Generative AI tools assisted in developing the conceptual framework, formulating and checking mathematical claims and derivations, refining hypotheses, designing methodology and experiments, implementing and testing methods, producing controlled synthetic data, processing result files, and interpreting results. They also assisted with literature discovery and summarization, software and figure editing, manuscript drafting and revision, reference formatting, and presentation structure. Explicit derivations, deterministic tests, held-out simulations, and archived numerical comparisons accompany the AI-assisted work. No human-subject qualitative analysis, interview design, transcription, or translation was performed. The authors take responsibility for the final text, claims, code, and artifacts.

ETHICS STATEMENT

This work uses synthetic, equation-derived, and public mathematical benchmark data and involves no human participants or personal information. The proposed diagnostic is intended to identify unsafe verifier-guided scaling before deployment. Because any fixed diagnostic can itself become a target for optimization, deployment should retain an independent target channel and restrict access to hidden evaluation details.

REPRODUCIBILITY STATEMENT

Appendix J maps each result to its packaged inputs and reproduction level. The default local checker reaggregates saved tables and recomputes decisions or risks where sufficient statistics or scores are retained. Historical runners document the generation protocols; they are not all runnable from the anonymous package. In particular, the original 703-by-64 residual bank is excluded, so its controlled results support table-level verification, not packaged regeneration. Revised exact baseline records remain separate from the original outputs. The earlier 300-task audit is retrospective.

For MATH-500, Appendix F specifies the development/confirmation split, prediction inputs, frozen decision rules, grading, and paired inference. Timestamped hash records separate rule freezing, label-free selection, and evaluation; archived candidate records and question-level summaries reproduce the reported tables and figures.

REFERENCES

Fariya Afrin and Ibne Farabi Shihab. Evaluator ensembles under reward hacking: Covariance geometry and finite-search guarantees. *arXiv preprint arXiv:2608.08002*, 2026. doi: 10.48550/arXiv.2608.08002. URL https://arxiv.org/abs/2608.08002.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi: 10.1214/24-AOAS1998.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, 2016.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. Reward model ensembles help mitigate overoptimization. In *International Conference on Learning Representations*, 2024.

Miles Cranmer. Interpretable machine learning for science with PySR and SymbolicRegression.jl. *arXiv preprint arXiv:2305.01582*, 2023.

Florian E. Dorner, Yatong Chen, André F. Cruz, and Fanny Yang. ROC-n-reroll: How verifier imperfection affects test-time scaling. In *International Conference on Learning Representations*, 2026.

Jinyang Du, Hao Ma, Xiaohu Shi, Bo Yang, Yanchun Liang, Heow Pueh Lee, and Chunguo Wu. LLM-PDESR: Robust PDE discovery via subdomain weighted residuals and LLM-guided symbolic hypothesis generation. *arXiv preprint arXiv:2607.10546*, 2026.

Kazuki Egashira, Mark Vero, Jasper Dekoninck, Florian E. Dorner, Robin Staab, and Martin Vechev. Delay, plateau, or collapse: Evaluating the impact of systematic verification error on RLVR. In *Conference on Language Modeling*, 2026. URL https://arxiv.org/abs/2605.02909.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, 2023.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 2017.

Jujie He, Tianwen Wei, Rui Yan, Jiacai Liu, Chaojie Wang, Yimeng Gan, Shiwen Tu, Chris Yuhao Liu, Liang Zeng, Xiaokun Wang, Boyang Wang, Yongcong Li, Fuxiang Zhang, Jiacheng Xu, Bo An, Yang Liu, and Yahui Zhou. Skywork-o1 Open Series, November 2024. URL https://doi.org/10.5281/zenodo.16998085.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html.

Hugging Face H4. MATH-500. Dataset repository, 2024. URL https://huggingface.co/datasets/HuggingFaceH4/MATH-500. Accessed September 16, 2026.

Hadi Khalaf, Claudio Mayrink Verdun, Alex Oesterling, Himabindu Lakkaraju, and Flavio du Pin Calmon. Inference-time reward hacking in large language models. In *Advances in Neural Information Processing Systems*, 2025.

Jan Hendrik Kirchner, Yining Chen, Harri Edwards, Jan Leike, Nat McAleese, and Yuri Burda. Prover-verifier games improve legibility of LLM outputs. *arXiv preprint arXiv:2407.13692*, 2024.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024.

David Manheim and Scott Garrabrant. Categorizing variants of Goodhart's law. *arXiv preprint arXiv:1803.04585*, 2018.

Daniel A. Messenger and David M. Bortz. Weak SINDy for partial differential equations. *Journal of Computational Physics*, 443:110525, 2021.

Sarang Manoj Pekhale, Amartya Roy, Rajat Sarkar, and Souvik Chakraborty. See, hypothesize, validate: Multimodal agentic framework for discovering governing PDEs. *arXiv preprint arXiv:2608.27869*, 2026.

Brenden Petersen, Mikel Landajuela Larma, Terrell N. Mundhenk, Claudio Santiago, Soo Kim, and Joanne Kim. Deep symbolic regression: Recovering mathematical expressions from data via risk-seeking policy gradients. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=m5Qsh0kBQG.

Amit Roth, Ankur Samanta, Matan Halevy, Yoav Levine, and Yonathan Efroni. Hack-verifiable environments: Towards evaluating reward hacking at scale. *arXiv preprint arXiv:2605.20744*, 2026.

Samuel H. Rudy, Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Data-driven discovery of partial differential equations. *Science Advances*, 3(4):e1602614, 2017.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, volume 32, pp. 2530–2540, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/8fb21ee7a2207526da55a679f0332de2-Abstract.html.

Silviu-Marian Udrescu and Max Tegmark. AI Feynman: A physics-inspired method for symbolic regression. *Science Advances*, 2020.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, 2024. doi: 10.18653/v1/2024.acl-long.510.

Zikai Xie, Wenmei Li, Man Luo, Jun Jiang, and Linjiang Chen. Language models guide symbolic equation discovery by controlling search. *arXiv preprint arXiv:2607.04156*, 2026.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-Math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, 2023.

A Empirical-Error Extension Beyond Gaussian Noise

We isolate the error-distribution assumption in a synthetic bank of 256 challengers represented in 16 dimensions. The bank combines a shared direction with candidate-specific components, and its norms span the transition region. We cross nine budgets with four noise scales under both Gaussian errors and independent standardized Student-t_8 coordinates. Each diagnostic uses 20,000 calibration draws; evaluation uses 50,000 independently seeded conditional Bernoulli event draws per regime, for 72 regimes and 3.6 million simulated event draws.

Under Student-t_8 error, the Gaussian approximation reaches RMSE 0.01493. Replacing Gaussian draws with empirical calibration errors in Equation 10 reduces RMSE to 0.00185 and maximum absolute error from 0.03048 to 0.00382. On the Gaussian control, the corresponding RMSE values are 0.00281 and 0.00315, so empirical sampling retains the accuracy of the closed form while adapting to the heavy tail. The schedule integration and capacity definition are unchanged; only the error sampler changes.

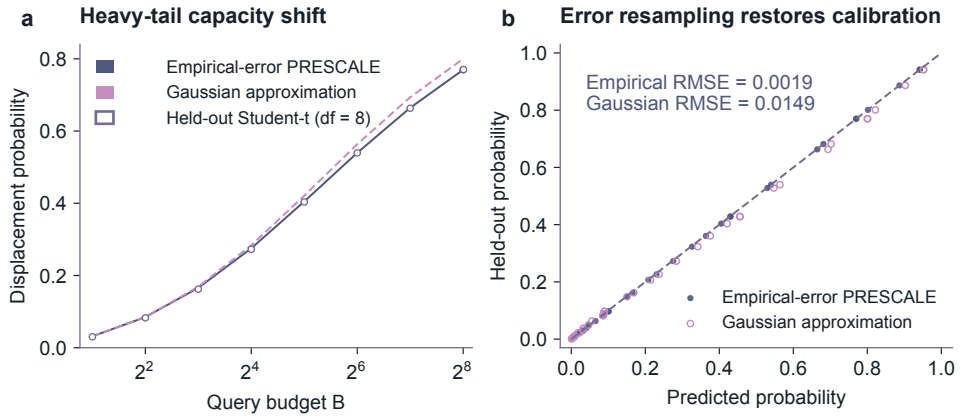

Figure 3: **Sample-based risk estimation under simulated heavy-tailed errors.** (a) At $1.4\times$ noise, standardized Student-t_8 errors shift the risk curve relative to the Gaussian approximation. Points are 50,000-event simulated estimates with ± 1 standard-error bars; the solid line uses 20,000 independent calibration errors. (b) Predicted versus held-out risk over all 36 Student-t_8 budget–noise regimes. Empirical-error PRESCALE reaches RMSE 0.0019 versus 0.0149 for the Gaussian approximation.

B Derivations

B.1 Pairwise displacement identity

Let $r_0 = r_{c_0}$ and $r = r_c$. Expanding the two quadratic scores in one chain gives

$$\begin{aligned} V(c; e) - V(c_0; e) &= (r - e)^\top (r - e) - (r_0 - e)^\top (r_0 - e) \\ &= \|r\|_2^2 - \|r_0\|_2^2 - 2(r - r_0)^\top e \\ &= \gamma_c - 2\Delta_c^\top e. \end{aligned} \tag{12}$$

Thus c beats c_0 exactly when $\Delta_c^\top e > \gamma_c/2$. Under $e \sim \mathcal{N}(0, \Sigma)$, the standardized projections

$$Z_c = \frac{\Delta_c^\top e}{\sqrt{\Delta_c^\top \Sigma \Delta_c}}$$

are jointly standard normal on positive-variance coordinates, with correlation matrix in Equation 5. Deterministic coordinates are handled as in Section 3.2. This proves Equation 7 in the all-harmful case. For mixed labels, a harmful candidate must beat both the incumbent and every sampled benign candidate. For distinct challenger gaps this yields Equation 6; with ties, compare (G_c, i_c) lexicographically, ordering the incumbent first.

B.2 Schedule integration

Fix an error draw and order the M challengers lexicographically by $(G_c(e), i_c)$, with smaller gap and then smaller fixed index preferred. If candidate c has W_c lower-ranked challengers, it is selected from a uniform sample of $B-1$ challengers precisely when (i) c is included and (ii) the other $B-2$ draws all come from those W_c candidates. The probability is $\binom{W_c}{B-2}/\binom{M}{B-1}$. The winner events are disjoint, so summing this rank weight over harmful candidates whose gap is negative and averaging over verifier error gives Equation 10. When every challenger is harmful, the event occurs iff the schedule hits at least one of the $K(e)$ negative-gap candidates, giving

$$\Pr(\mathcal{H}_B \mid e) = 1 - \frac{\binom{M-K(e)}{B-1}}{\binom{M}{B-1}}. \tag{13}$$

This form computes all budgets from the same vector of negative-gap counts.

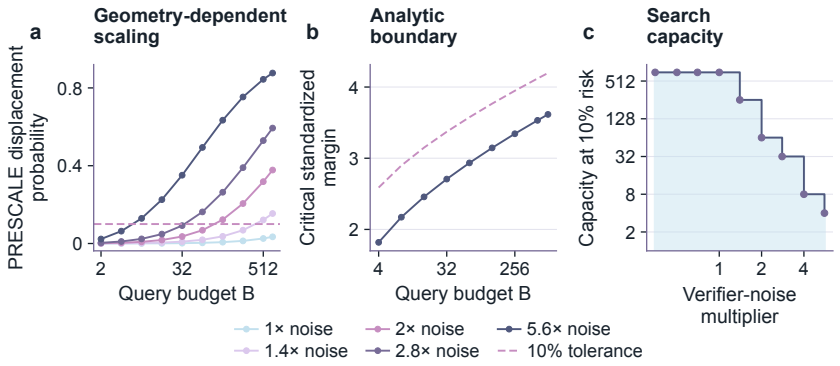

Figure 4: **Conditional capacity and the independent homogeneous limit.** (a) PRESCALE risk curves for five verifier-noise regimes; the dashed line is the 10% tolerance. (b) Exact homogeneous 10% phase boundary and its $\sqrt{2\log B}$ leading-order scaling. (c) Point-estimated grid capacity at the same tolerance for the empirical PDE bank. Panel (a) uses 20,000 verifier-error draws; Monte Carlo standard errors are at most 0.0036.

B.3 Marginal non-identifiability and all-harmful monotonicity

For Proposition 1, let $Z_1 = \cdots = Z_K = Z$ in the fully correlated system. Every challenger has crossing probability $p = \overline{\Phi}(m)$ and their union also has probability p. For IID Z_k, the same marginals give union probability $1-(1-p)^K = 1-\Phi(m)^K$, which converges to one for fixed $p > 0$. The two systems have identical marginal score-gap distributions but disagree at the selected-system level.

For $m > 0$, both constructions are realizable by the quadratic model: take $r_{c_0} = 0$ and $\Sigma = I_p$, with $p \geq K$. For shared projections set $r_c = 2mu_1$ for every challenger; for independent projections set $r_c = 2mu_c$, where $u_1, \ldots, u_K$ are orthonormal. Both banks have $\gamma_c = 4m^2$ and $s_c = 2m$, hence the same standardized margin m. Distinct candidate indices may have the same residual vector. The independent construction requires sufficient dimension as K grows; it is not the 703-direction, 64-coordinate control.

For all-harmful schedule monotonicity, draw one uniform random permutation of the M challengers and define budget B by its first $B-1$ entries. The event that this prefix contains one of the $K(e)$ negative-gap candidates is nested as B increases. Its conditional probability, and hence its expectation over e, is nondecreasing. Mixed labels break this nesting for the selected-winner event because a newly included benign candidate can outrank a harmful candidate.

B.4 Homogeneous boundary

For an all-harmful bank with $B-1$ independent projections, finite common margin m, and tolerance $0 < \tau < 1$,

$$\mathcal{R}_B = 1 - \Phi(m)^{B-1}, \qquad \mathsf{Cap}^{\mathrm{safe}}_{\tau,\{1,\ldots,M+1\}}(m) = \min\left\{M+1,\ 1 + \left\lfloor \frac{\log(1-\tau)}{\log \Phi(m)} \right\rfloor \right\}. \tag{14}$$

For $B > 1$ and $0 < \tau < 1$, the corresponding critical margin is

$$m_\tau(B) = \Phi^{-1}\left((1-\tau)^{1/(B-1)}\right) = \sqrt{2\log B} + o(\sqrt{\log B}). \tag{15}$$

The capacity above ranges over all integers and respects the finite-bank ceiling; on a sparse grid one instead selects its largest admissible point. The asymptotic statement concerns a sequence of independent homogeneous systems. In that regime, maintaining fixed risk requires margins to grow on the $\sqrt{\log B}$ scale. It is not a requirement for all correlated banks: the fully shared-error construction in Proposition 1 retains constant risk at fixed margins. The leading term is the classical Gaussian maximum scale, derived in Appendix B, rather than an additional geometric asymptotic claim.

With $B-1$ independent candidates and common standardized margin m, no displacement occurs with probability $\Phi(m)^{B-1}$. Solving $1-\Phi(m)^{B-1} = \tau$ yields the exact first equality in Equation 15.

Table 4: Prediction error across 270 budget–noise–geometry regimes.

Predictor	MAE	RMSE	Max. error
PRESCALE	.00070	.00132	.00574
Independent marginals (exact)	.04504	.12368	.56652
Budget only	.12589	.26193	.98148
Nearest-margin bank (exact)	.42175	.55909	.98179

Let $q_B = 1 - (1-\tau)^{1/(B-1)}$. Then $q_B \sim -\log(1-\tau)/(B-1)$. The Gaussian upper-quantile expansion

$$\Phi^{-1}(1-q) = \sqrt{2\log(1/q)} + O\left(\frac{\log\log(1/q)}{\sqrt{\log(1/q)}}\right)$$

therefore gives $m_\tau(B) = \sqrt{2\log B} + o(\sqrt{\log B})$. The statement concerns independent standardized projections; the full estimator does not replace a correlated bank by this asymptotic approximation.

B.5 DISTRIBUTION-SHIFT INEQUALITY

For a fixed search policy, let $h_B(x) \in [0,1]$ denote the conditional probability of the hacking event given a calibration unit x (task, incumbent, candidate bank, and verifier state). By the variational characterization of total variation,

$$|\mathbb{E}_{P_{\text{live}}} h_B - \mathbb{E}_{P_{\text{cal}}} h_B| \le \text{TV}(P_{\text{live}}, P_{\text{cal}}),$$

which proves Equation 20. This result is intentionally policy-specific: changing the generator changes h_B and the distribution over candidate banks, motivating recalibration.

C CONTROLLED CANDIDATE-BANK CONSTRUCTION

The candidate universe is a finite grammar of conservative one-dimensional PDE correction terms. Expressions are expanded, rational coefficients are canonicalized, and differential monomials are collected before deduplication. This produces 1,088 distinct normal-form classes. Analytic principal-part and positivity rules remove 384 ineligible classes; the oracle reference and 703 eligible non-reference classes remain. The target label is structural inequality to the reference normal form, exact under this canonical structural-equality criterion and independent of the numerical residual used for ranking. It is not a label of general numerical-solution accuracy.

Each candidate is evaluated under four interventions. Sixteen normalized weak moments per intervention produce a 64-coordinate residual vector. All moment numerators, denominators, normalizers, candidate identifiers, and the development estimate $\sigma_0 = 0.0443171$ are stored in the sealed evidence artifact. This construction makes three otherwise confounded quantities separately observable: target label, noiseless proxy margin, and verifier-error geometry.

The margin-preserving interventions start from the vector norms of this empirical bank. The randomized-direction control assigns independent random unit directions before restoring each original norm. The clustered control mixes each random direction with a shared unit direction at coefficient 0.9, renormalizes, and again restores the original norms. Hence all three banks have the same $\{m_c\}$ under isotropic error; only C changes. Random seeds are fixed before held-out evaluation.

C.1 NUMERICAL EVIDENCE AND HISTORICAL AUDIT

Across these 270 fixed regimes, PRESCALE has RMSE 0.00132, MAE 0.00070, and maximum absolute error 0.00574. Table 4 retains the full information-ablation comparison.

The equation-discovery application separates correctness from verifier error. Symbolic expressions are collected into a differential-polynomial normal form, so candidate equality is exact rather than evaluated at random points. The weak evidence retains all moments and normalizers needed to reconstruct each coordinate of r_c.

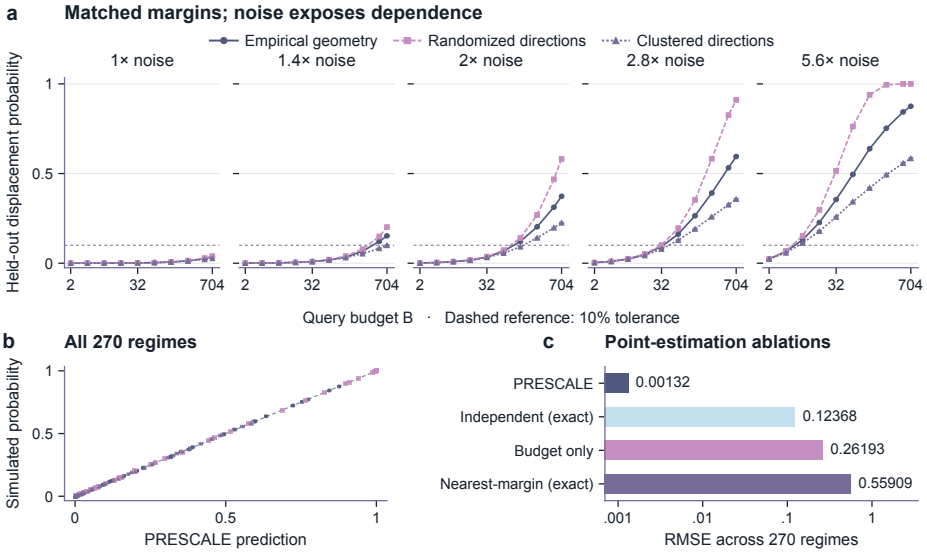

Figure 5: **Complete specified-law numerical checks.** (a) Risk curves at five of the nine noise scales compare three geometries with the same 703 margins. (b) Predictions versus 50,000 simulated event draws per regime across all 270 budget–noise–geometry regimes. (c) Removing dependence raises RMSE from 0.0013 to 0.1237; budget-only calibration reaches 0.2619. Bars are ± 1.96 Monte Carlo standard errors (50,000 evaluation and 20,000 diagnostic draws), often smaller than markers. The evaluation uses a Bernoulli draw from the conditional hit probability, rather than executing an explicit subset winner. These checks measure numerical agreement within the specified Gaussian law; the raw-measurement audit is reported separately in Figure 2b.

The archived summaries of the original audit report acceptance thresholds fitted on 36 policy tasks, a disjoint 72-task calibration screen, and 300 audit tasks. At budgets 4 and 32, answer rates are 0.2867 and 0.2900; accepted structural errors are $0/86$ and $1/87$. This selective-risk endpoint is downstream of the displacement mechanism and remains a separate deployment-policy measurement.

Table 5: Controlled evaluation inventory.

Object	Count
Canonical grammar classes	1,088
Analytically excluded classes	384
Eligible target-worse candidates	703
Residual coordinates per candidate	64
Margin-matched geometries	3
Budgets $\times$ noise levels per geometry	10×9
Simulated event draws per regime	50,000
Total simulated event draws	13,500,000

D Finite-Calibration Study

The principal controlled result supplies the true Gaussian family and uses 20,000 diagnostic draws. The finite-calibration experiment instead observes only IID error vectors drawn from that family. For each sample size $n \in \{256, 1024, 4096, 20000\}$, eight independent calibration sets are passed directly to the empirical-error form of Equation 10. A separate set of 50,000 conditional Bernoulli event draws defines the evaluation target over ten budgets and noise multipliers 1.0, 2.8, and 5.6.

The confidence statement in Equation 11 is conditional on a predeclared budget grid, this frozen candidate bank and harmful mask, the fixed sampling policy, and IID calibration errors from the

specified verifier-error law. The bank and policy are independent of the calibration sample. We use $\delta = 0.05$, $J = 10$ per fixed noise setting, and $U_1 = 0$. It makes no coverage claim under generator, verifier, or target-channel shift. For each budget, one-sided Hoeffding gives $\Pr\{\mathcal{R}_B - \widehat{\mathcal{R}}_B > t\} \le \exp(-2nt^2)$; setting $t = \sqrt{\log(J/\delta)/(2n)}$ and taking a union bound over J budgets proves simultaneous coverage. Applying the running maximum before thresholding preserves that coverage for certified safe capacity. Budget one has deterministic zero replacement risk and needs no statistical bound.

Table 6: Finite-calibration RMSE across eight independent replicates.

Error draws	Mean	Std. dev.	Minimum	Maximum
256	.01372	.00548	.00727	.02183
1,024	.00704	.00278	.00362	.01181
4,096	.00284	.00138	.00143	.00551
20,000	.00181	.00035	.00113	.00220

Table 7: Safe-capacity decisions at 10% risk over 24 replicate–noise comparisons per row. The shortfall counts positions on the evaluated budget grid relative to the 50,000-event reference.

Error draws	Point over.	Cert. over.	Cert. exact	Mean shortfall
256	.000	.000	.000	5.67
1,024	.000	.000	.333	.79
4,096	.000	.000	.667	.33
20,000	.000	.000	.667	.33

In actual budget units, the mean certified shortfalls are 245.67, 7.00, 5.33, and 5.33 for the four sample sizes. The reference capacities at noise multipliers $(1, 2.8, 5.6)$ are $(704, 32, 4)$. With $n = 256$, all three certified capacities are 1; with $n \ge 4096$, they are $(704, 16, 4)$. Thus a mean shortfall of only 0.33 grid positions still costs 16 queries at the middle noise level. At $n = 1024$, the middle capacity is 16 in five replicates and 8 in three, while the other capacities are 704 and 2. These are comparisons to a finite Monte Carlo reference, not exact known capacities or independent coverage trials.

The result distinguishes two sources of accuracy. Exact schedule integration removes randomness from which candidates happen to be sampled, but it does not remove uncertainty about the verifier error distribution. The shared 0/96 overrun count for point and certified capacities is therefore only an implementation and conservatism sanity check, not a coverage experiment. Simultaneous coverage follows from the bound above; the table records how much capacity that guarantee withholds at each calibration size.

D.1 Independent comparison of calibration rules

Exact subset integration and statistical calibration address different uncertainties. We isolate the latter by keeping the 703-candidate PDE bank, error law, and integrated loss $f_B(e) \in [0, 1]$ fixed and comparing three classical calibration rules. The protocol, seeds, and primary comparison were frozen before drawing new simulated errors. Each of 32 independent calibration replications supplies 4,096 Gaussian vectors; sample sizes 256 and 1,024 use prefixes of the same replication. We share these vectors across noise multipliers 1, 2.8, and 5.6 and across methods. The budget grid is $\{1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 704\}$, with risk tolerance $\tau = 0.1$ and error level $\delta = 0.05$ separately for each method, sample size, and noise setting.

The first rule is Equation 11. The second replaces its additive Hoeffding bound with the bounded-loss KL–Chernoff test and Bonferroni correction. The third uses the same test in a fixed sequence of increasing budgets, stopping at the first failed test, following the risk-control framework of Angelopoulos et al. (2025). For the null $\mathcal{R}_B \ge \tau$, a valid conservative p-value is

$$p_B = \exp\{-n\,\mathrm{kl}(\min(\widehat{\mathcal{R}}_B, \tau), \tau)\}, \qquad \mathrm{kl}(a, b) = a\log(a/b) + (1-a)\log((1-a)/(1-b)),$$

Table 8: All calibration settings: selected grid budget, mean (SD) over 32 replications. Each row uses the same calibration vectors across rules. Large mean budgets need not themselves be evaluated grid points.

n	Noise	Additive Hoeffding	KL–Bonferroni	Fixed sequence
256	1	1 (0)	666 (98.69)	698 (33.94)
256	2.8	1 (0)	9.5 (3.17)	16 (0)
256	5.6	1 (0)	2 (0)	2.125 (.49)
1024	1	704 (0)	704 (0)	704 (0)
1024	2.8	14 (3.52)	16 (0)	16 (0)
1024	5.6	2 (0)	4 (0)	4 (0)
4096	1	704 (0)	704 (0)	704 (0)
4096	2.8	16 (0)	16 (0)	18 (5.38)
4096	5.6	4 (0)	4 (0)	4 (0)

with continuous endpoint conventions. The tests use thresholds $\delta/10$ and δ, respectively. Budget one is structurally safe. In the fixed sequence, erroneously passing any unsafe budget requires rejecting the first true null; its probability is at most δ. This certifies the passed grid prefix, not an anytime stopping rule or a simultaneous confidence band. The bound is valid for fractional IID bounded losses; we do not interpret integrated losses as binomial counts.

An independent reference uses 131,072 simulated error vectors and simultaneous two-sided Hoeffding intervals over the 30 nontrivial noise–budget points at family error 0.01. This reference is a numerical estimate, not exact ground truth. In the predeclared primary setting, $n = 256$ and $2.8\times$ noise, fixed-sequence calibration selects $B = 16$ in all 32 replications, whereas the additive rule selects $B = 1$ in all 32. The mean paired difference is 15 queries (replicate standard deviation 0); the independent reference risk at $B = 16$ is 0.04784, with simultaneous interval $[0.04208, 0.05360]$. The additive radius is about 0.10173, exceeding τ even at zero estimated risk. Thus its fallback here reflects conservatism of that bound, not the absence of an admissible expansion. The fixed-sequence comparison reduces this calibration cost on the same supplied geometry; it neither measures native LLM verifier errors nor establishes an improvement in final-answer correctness.

All sample sizes, noise levels, methods, and replications are archived in `results/cpu-calibration-comparison-20260917/replicates.csv`, with an independent `reference.csv`, the frozen protocol, and source hashes. The resulting 864 decision rows share calibration samples and must not be treated as 864 independent coverage trials.

The primary-setting KL–Bonferroni rule selects 8 in 26 replications and 16 in six. The independent reference at 8 is .02334 with simultaneous 99% interval $[.01758, .02910]$. All 864 archived decisions were independently recomputed from the saved calibration means, and all 33 reference intervals were checked. Individual calibration vectors were not archived in that earlier run; this check does not replay unrecorded draws.

Matched marginal fixed-sequence test. For a frozen bank of M challengers, let

$$X_i = \frac{1}{M} \sum_{c \in D_0^-} \mathbf{1}[G_c^{(i)} < 0], \qquad \mu = \mathbb{E}[X_i], \qquad \widehat{\mu} = n^{-1} \sum_i X_i.$$

Uniform sampling independently of the gaps gives $\mathcal{R}_B \le \min\{1, (B-1)\mu\}$ by the union bound. For $B > 1$, set $q_B = \tau/(B-1)$ and test $H_{0,B}^{\mathrm{marg}} : \mu \ge q_B$ with

$$p_B^{\mathrm{marg}} = \exp\{-n \,\mathrm{kl}(\min(\widehat{\mu}, q_B), q_B)\}.$$

For IID $X_i \in [0,1]$, convexity gives $\mathbb{E}e^{\lambda X_i} \le 1 - \mu + \mu e^{\lambda}$. The lower-tail Chernoff bound makes this a valid conservative p-value without candidate independence. We use the same grid, $n = 256$, $\delta = .05$, $\tau = .1$, increasing-budget order, and first-failure stopping rule as the joint test. The first-true-null argument above controls the passed prefix separately for each replication and procedure.

In the all-harmful bank, $f_2(G^{(i)}) = X_i$ exactly, hence $\widehat{\mu} = \widehat{\mathcal{R}}_2$. This post-hoc comparator uses archived sufficient statistics for the identical 256-row sample in each of all 32 primary replications;

no errors are regenerated and no analytic tails enter either finite-data test. At $2.8\times$ noise, all marginal decisions are $B = 4$ and all joint decisions are 16. Marginal p_4 ranges from .00191 to .00460, whereas p_8 ranges from .15674 to .27209; joint p_{16} ranges from .00085 to .04104, and p_{32} from .34009 to 1. All p-values, tested-prefix indicators, decisions, replication IDs, and source hashes are archived in `results/matched-marginal-sequence-20260920/`. The earlier run saved means rather than individual vectors; this is a reproducible reanalysis of those statistics, not a replay of unrecorded draws.

Confidence-bound duality. Inverting the same bounded-row inequality at error level δ gives

$$U_\mu = \sup\{q \in [\widehat{\mu}, 1] : n\,\mathrm{kl}(\widehat{\mu}, q) \le \log(1/\delta)\}, \qquad U_B^{\mathrm{marg}} = \min\{1, (B-1)U_\mu\}, \quad U_1^{\mathrm{marg}} = 0.$$

The event $\mu \le U_\mu$ controls all budgets without candidate-wise or budget-wise Bonferroni. For $0 < \tau < 1$, $p_B^{\mathrm{marg}} \le \delta$ is equivalent to $q_B \ge U_\mu$, hence to $(B-1)U_\mu \le \tau$, with consistent non-strict boundary conventions. Thresholds q_B decrease with B, so fixed-sequence testing and the previous marginal upper-bound rule return the same budget. Their agreement is a dual representation, not independent evidence. Numerically, U_μ ranges from .01950 to .02269 and all 32 upper-bound decisions also equal 4 (`results/same-data-marginal-20260920/`). The statistic $\widehat{\mu}$ is invariant to independent column permutations. Matching tests therefore compares the exact joint loss with this alignment-free union-bound reduction; it does not separate joint-information value from union-bound slack or imply universal superiority.

Known-law marginal certificate. With uniform sampling, the marginal union bound is $\min(1, (B-1)M^{-1}\sum_c p_c)$, where $p_c = \Pr(G_c < 0)$ for each harmful challenger. Under the stipulated Gaussian law, these tails are available analytically. At $2.8\times$ noise, their mean is .00346029, giving bounds .05190 at $B = 16$ and .10727 at 32. The resulting certified grid budgets at multipliers 1, 2.8, and 5.6 are 704, 16, and 4. This comparator uses known marginals, not the same finite calibration data. It ties fixed-sequence calibration in the primary setting. Comparing it with the finite-data marginal procedure isolates the additional cost of estimating the crossing mean. The independent joint reference interval at 32 is $[.08544, .09696]$, illustrating conservative slack in the marginal bound without certifying 32 from the $n = 256$ data.

D.2 Measurement-level simulated observation audit

This separate experiment tests how a complete score-gap observation can be obtained from upstream measurements. It uses one fixed simulated trajectory and a new bank, not the 703-challenger Gaussian residual experiment. We choose the first eligible law in the deterministic catalog, retain its structural oracle as reference, and select the 63 structurally nonequivalent challengers with smallest noiseless weak residual scores. Thus every challenger is harmful under the structural target. Clean-score selection makes this a deliberately local difficulty subbank; it is not a sample of PDEs or a test of mixed-bank competition.

Raw measurement and scoring contract. The periodic domain has length 2π, 128 spatial points, and Fourier modes 1–8 in an initial condition with RMS 0.35. A spectral simulation uses time step 0.000625 for 128 steps and keeps every fourth snapshot, producing 33 observations over time 0.08. Fixed metadata are $(1/2, 1/3)$. Each fresh realization independently adds Gaussian noise to every raw space–time entry, with standard deviation $0.01\max(\mathrm{RMS}(u), 0.125) = 0.00340071$. The same noisy observation is shared by all 64 candidates. Existing weak-moment routines recompute nonlinear candidate moments using 16 sine/cosine probes and eight Simpson windows, giving 128 weak coordinates. The lower-is-better score is normalized RMS residual. Each coordinate's normalizer is frozen from the clean trajectory as the maximum of the window duration divided by eight, the absolute clean field increment, and the absolute clean oracle integral. This clean-oracle normalization is part of the simulated contract, not an estimated deployment input. Coefficients, reference, labels, and bank remain fixed; no bias centering, covariance fitting, or Gaussian gap approximation is applied.

Frozen stages and independent evaluation. Disjoint seeds produce 128 development, 256 calibration, and 512 evaluation rows. The protocol is written before development; calibration predictions are written before evaluation observations are generated. Local hashes and execution

order document this sequence, not external preregistration. All 896 complete rows are retained without filtering. Development rows do not enter the reported estimates. The declared grid is $B \in \{1, 2, 4, 8, 16, 32, 64\}$; the reference wins exact ties, and remaining ties use fixed candidate order. For each evaluation row, an independent routine explicitly draws 128 uniform subsets without replacement, selects the score winner, and checks its structural label. It does not call the rank formula. The independent sample size is 512 complete measurement rows, not 512×128 independent measurements.

Table 9: Measurement-level audit on the new 63-challenger bank. All entries are percentages; fresh-risk SE is computed across 512 complete measurement rows. Predictions use the same 256 calibration rows. Direct MC uses 128 subsets per row; the shuffled control averages 32 artificial column permutations, not 32 datasets.

B	Aligned gaps	Shuffled gaps	Direct MC	Fresh risk (SE)
2	8.352	8.352	8.395	7.968 (0.410)
4	20.808	23.121	20.926	20.222 (0.969)
8	35.118	46.186	35.141	34.889 (1.537)
16	46.486	74.186	46.536	47.418 (1.872)
32	54.134	94.554	54.077	56.151 (2.032)
64	58.984	99.866	58.984	61.523 (2.152)

The column-shuffled control exactly preserves each empirical calibration-gap marginal while destroying alignment between candidates. It is an information ablation, not a physically realized second verifier. Mean absolute error over the six nontrivial budgets is 1.115 percentage points for aligned gaps, 19.682 for shuffling, and 1.146 for same-information MC. At $B = 64$, prediction minus fresh risk is -2.539 points; a descriptive two-sample bootstrap resampling whole calibration and evaluation rows separately gives 95% interval $[-9.961, 4.883]$ (4,000 replicates). A separate post-run bootstrap of evaluation rows, conditional on the frozen predictions, gives a shuffled-minus-aligned MAE difference of 18.568 points, interval $[13.387, 20.397]$ (2,000 replicates). Neither interval measures variation across trajectories, banks, or measurement laws. The 1.115 versus 1.146 comparison does not establish predictive superiority over MC.

Calibration result and side-diagnostic correction. The predeclared additive rule uses $\tau = 0.1$, $\delta = 0.05$, and six nontrivial budgets. Its radius is 0.09670 and it selects only $B = 1$; no nontrivial certified expansion is demonstrated. An auxiliary archived field initially called a marginal bound computed $\mathbb{E}[\min(1, kK/M)]$, where K is the row crossing count. This is a valid conditional-union bound but still uses joint information. We preserve that record and separately report the post-run marginal-only correction $\min(1, k\mathbb{E}[K/M])$. A single Hoeffding upper bound on $\mathbb{E}[K/M]$ controls all budgets; its certified budget is also one. This correction neither changes primary predictions nor retroactively becomes a predeclared comparison.

Implementation checks and scope. An independently assembled, unfactored weak-moment scorer on a reconstructed noisy observation matches all 64 scores to 5.55×10^{-17}; an independent hypergeometric calculation matches rank integration to 2.22×10^{-16}. All 896 observation hashes are distinct. Scoring takes 23.58 seconds on a local CPU, excluding bootstrap and plotting. Scripts and frozen arrays are in `scripts/measurement-bridge-20260920/` and `results/measurement-bridge-20260920/`. The result supports an observation-to-risk interface under the specified simulated measurement process. It does not validate natural Gaussian gap geometry, native PRM uncertainty, or answer improvement.

D.3 Cached exact accounting versus direct subset simulation

Conditional variance. Let H_B be a harmful-winner indicator for a random subset and let G denote the complete vector of score gaps. Equation 10 computes $f_B(G) = \mathbb{E}[H_B \mid G]$. The classical law of total variance gives

$$\mathrm{Var}(H_B) = \mathrm{Var}(f_B(G)) + \mathbb{E}[\mathrm{Var}(H_B \mid G)].$$

Table 10: Cached CPU risk accounting over all 276 banks and four budgets. Means over 32 numerical repetitions; parenthesized values are repeat SDs. RMSE is risk error over 276 tasks and the two interior budgets $N = 16, 32$, in percentage points relative to exact integration. Times exclude loading, generation, and PRM scoring. Memory is peak traced allocation during the method, not total process RSS.

Method	Batched time (ms)	Per-task time (ms)	Batched RMSE	Peak MiB
Exact	.251 (.006)	—	0	.319
MC, $L = 1$	.190 (.024)	4.743	7.494 (1.276)	1.025
MC, $L = 8$	.937 (.043)	5.581	2.499 (.378)	2.609
MC, $L = 32$	3.453 (.093)	7.812	1.347 (.178)	8.125
MC, $L = 128$	13.738 (.444)	16.567	.638 (.119)	19.450
MC, $L = 1024$	160.322 (2.194)	114.388	.223 (.039)	20.057

With n IID complete vectors and L conditionally independent subsets per vector, the variance of direct MC is $\mathrm{Var}(f_B(G))/n + \mathbb{E}[f_B(G)(1 - f_B(G))]/(nL)$. Exact integration removes the second term, not calibration, task, or model uncertainty. This is Rao–Blackwellization of a specified finite-bank audit.

Cached recurrence and cost. For $k = B - 1 > 0$ challengers from a bank of size M, the probability that rank j (best is one) wins is $w_{1,k} = k/M$ and $w_{j+1,k} = w_{j,k}(M - j - k + 1)/(M - j)$ for $1 \le j \le M - k$. All later weights and all $k = 0$ challenger weights are zero. The implementation stores an immutable $O(JM)$ table keyed by bank size and grid; construction takes $O(JM)$ once. Sorting each complete gap row costs $O(M \log M)$, and summing the harmful, reference-beating ranks costs $O(JM)$. The isolated implementation removes an unused schedule table from validation; it does not change the archived experiment code. Twenty-one tests (15 kernel tests and six existing calibration tests), including fixed-seed Monte Carlo checks, cover enumeration, ties, endpoints, validation, caching, and calibration.

Post-hoc same-information computation benchmark. Every method receives the same cached 276 MATH score vectors and offline target labels. There are 56 challengers relative to each first-eight reference and $k \in \{0, 8, 24, 56\}$, corresponding to $N \in \{8, 16, 32, 64\}$. No new model calls, calibration observations, or budget decisions are made. The exact result matches all 1,104 archived task–budget risks to 1.11×10^{-16}. The benchmark compares two direct-selection implementations: per-task permutations and a stronger batched version, whose protocol was fixed after inspecting the first implementation. Both are retained, including their unfavorable outcomes. Each uses fresh uniform permutations, executes winner selection, and handles ties by the same fixed candidate order.

One batched subset is faster than exact accounting; at $L \ge 8$, exact is faster than both supplied MC implementations on this measured grid. This is an implementation-specific tradeoff, not a universal runtime ordering. At $L = 1024$, per-task MC is faster than batching and has task-risk RMSE .224 points. Across repetitions, batched aggregate-risk RMSE falls from .439 to .012 points as L grows from one to 1,024. Only 20 of the 552 interior task–budget cells have nonzero conditional variance; $N = 8$ and 64 are exact null-variance endpoints, included in all-budget timings and the complete risk arrays but excluded from both task-risk and aggregate-risk RMSE.

The benchmark uses an Apple M5 Pro (15 CPU cores, 48 GiB RAM), macOS 26.5.2, Python 3.12.13, and NumPy 2.3.5 linked against Apple Accelerate. One Python process executes each call, without a CPU-affinity or BLAS-thread constraint. Method order is randomized across 32 timing blocks; exact timing averages 50 calls per block, not 50 independent repetitions. Memory instrumentation runs separately in fresh processes. Cold weight construction, cache lookup, sorting, and summation respectively take 5.643, .059, 223.531, and 8.368 microseconds in separately timed phases. Raw times, seed-level estimates, memory measurements, source snapshots, and both protocols are archived in `results/cpu-rank-efficiency-20260920/`. Cached records contain 17,664 sampled candidates plus 276 greedy candidates, 11,877,359 output tokens, 2,261,480 prompt tokens, and 14,118,584 PRM tokens. Original generation/scoring wall time is unavailable in this benchmark, not zero. The CPU times therefore establish neither an end-to-end inference speedup nor forecasting superiority on unseen tasks.

E Pythia Checkpoint Protocol and Results

We use `EleutherAI/pythia-410m-deduped` at revisions `step0`, `step1000`, `step10000`, and `step143000`. Pythia is useful here because the architecture and tokenizer stay fixed while training state changes. For each canonical candidate string s_c and the fixed prompt "A candidate correction law for a partial differential equation is:", we compute

$$\ell_\theta(c) = \frac{1}{|s_c|} \sum_{j=1}^{|s_c|} \log p_\theta(s_{c,j} \mid s_{c,<j}, \text{prompt}). \tag{16}$$

The top 128 candidates by ℓ_θ define D_θ. This does not assert that likelihood is an optimal PDE generator; it supplies a reproducible, checkpoint-dependent proposal prior over an otherwise fixed and exactly labeled universe.

At $2.8\times$ noise, the initial curve permits $B = 64$ (predicted risk 0.0627), but that budget reaches 0.1230 at step 1,000. Recalibration selects $B = 32$ (predicted and simulated risk 0.0697). This evaluates uniform top-128 search, not checkpoint generation.

Table 11: Checkpoint-conditioned proposal support and point-estimated 10% grid capacity.

Checkpoint	Jaccard w/ final	$1.0\times$	$2.8\times$	$5.6\times$
step0	.076	129	64	4
step1000	.580	129	32	4
step10000	.829	129	129	8
step143000	1.000	129	129	8

Each checkpoint-conditioned prediction uses 20,000 verifier-error draws; each held-out estimate uses 50,000 independently seeded conditional Bernoulli event draws for seven budgets and three noise levels. Aggregated over 84 checkpoint–budget–noise regimes, checkpoint-conditioned RMSE is 0.00150. Broadcasting the step-0 prediction across all checkpoints gives RMSE 0.02969. The comparison isolates the fixed-landscape issue: even with an unchanged verifier, a moving proposal support invalidates a stale capacity estimate.

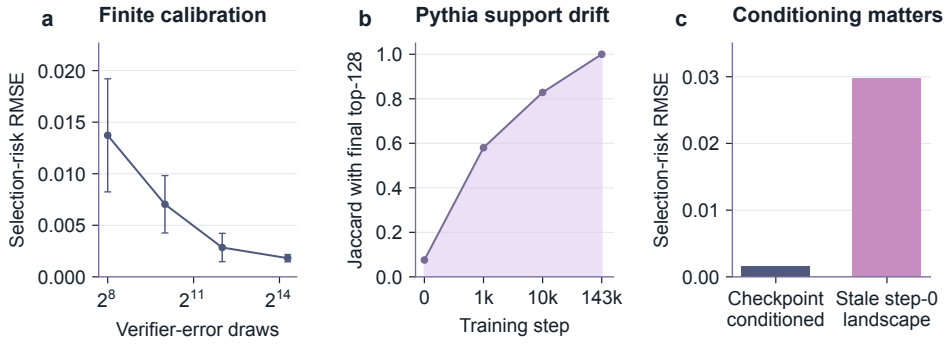

Figure 6: **Finite calibration error and checkpoint-conditioned proposal drift.** (a) Selection-risk RMSE across eight finite-error calibration replicates; bars show one standard deviation. (b) Overlap between each Pythia checkpoint's top-128 candidate support and the final support. (c) Prediction error when capacity is recomputed for each checkpoint versus transferred unchanged from step 0. Checkpoint-conditioned predictions use 20,000 error draws and evaluation uses 50,000 conditional Bernoulli event draws in each regime.

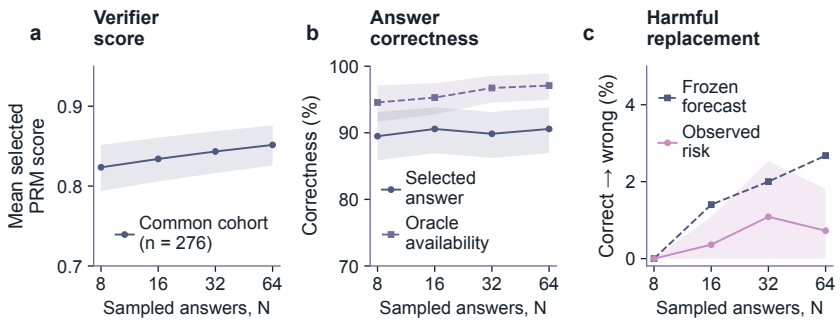

Figure 7: **Three consequences of scaling on 276 reserved MATH questions.** (a) Selected PRM score. (b) Selected-answer correctness and oracle availability. (c) Harm relative to the first-eight winner: frozen development rank forecast versus observed prefix risk, both over all tasks. Shading denotes 95% task-bootstrap intervals for confirmation means, not forecast calibration bounds. Forecasts integrate uniform subsets; evaluation uses the recorded prefixes. Appendix F separates these sampling targets.

Table 12: **Frozen rules on the confirmation set.** Harm uses all 276 tasks relative to $N = 8$; tokens are mean generated-output tokens per question, excluding development cost. Identical selections are grouped. The fallback is not a nontrivial certificate.

Policy (identical outputs grouped)	N	Acc. (%)	Harm (%)	Tokens (k)
Fixed-8; rank+net; lookup+net; MC+net; Hoeffding fallback	8	89.49	0.00	5.31
Fixed-16	16	90.58	0.36	10.64
Fixed-32	32	89.86	1.09	21.36
Fixed-64; risk-capacity point estimate	64	90.58	0.72	42.36

F MATH-500: Frozen Forecasts and Budget Decisions

F.1 Reward, replacement, and frozen-budget comparisons

Reward, correctness, and replacement. Across the same 276 questions, mean selected PRM rises from 0.82346 at $N = 8$ to 0.85153 at 64, while correct answers rise from 247 to 250 and oracle availability from 261 to 268 (Figure 7). That net gain comprises five corrections and two harmful replacements. At $N = 32$ there are three harmful replacements, falling to two at 64: benign winners can shield a mixed bank. Thus aggregate improvement and reference-relative risk coexist; higher reward alone describes neither outcome.

From a risk curve to a budget. Exact subset integration on development banks forecasts all-task harm of 1.40%, 2.00%, and 2.68% at $N = 16, 32, 64$; confirmation prefix rates are 0.36%, 1.09%, and 0.72%. At tolerance $\tau = 0.05$, the point-estimated capacity chooses 64. A separate rule maximizes predicted corrections minus harmful replacements among admissible budgets, breaking ties toward smaller budgets; it chooses 8, as do simple prefix lookup and 1,024-subset Monte Carlo using the same development data (Table 12).

The prespecified $N = 8$ minus $N = 64$ accuracy difference is -1.09 percentage points (95% paired bootstrap $[-2.90, 0.72]$; exact McNemar $p = 0.453$). Eight samples use 87.47% fewer output tokens in replay, but this interval does not establish noninferiority.

Exact schedule accounting therefore does not itself establish decision advantage. The earlier 1.5B-policy/1.5B-PRM GSM8K audit reports 3.44% harm conditional on 394 correct incumbents (Appendix G), unlike MATH's all-task denominator; different models, benchmarks, and references also prevent attributing the contrast to verifier strength.

F.2 Data, scoring, and the information boundary

We use the 500-question MATH-500 test set (Hendrycks et al., 2021; Hugging Face H4, 2024). A fixed seeded permutation (seed 20260915) assigns 224 questions to development and 276 to confirmation. This is a split of a public finite benchmark, not 500 IID draws from all future mathematical queries. Development outcomes were inspected during method design. Confirmation outcomes were evaluated only after the prediction values, budget rules, selection code, and primary comparison were frozen and hashed. Label-free selection accepts scores and candidate IDs, saves the chosen IDs, and only then loads gold labels for evaluation. Public benchmark answers are not claimed to have been cryptographically inaccessible before this analysis.

Qwen2.5-Math-7B-Instruct generates 64 stochastic solutions per question at temperature 0.7, top-$p = 0.95$, and maximum 3,072 new tokens. A separately generated greedy completion is retained in the archive but excluded from the main experiment and its budget N. Skywork-o1-Open-PRM-Qwen-2.5-7B is scored in FP32 using the mean of its step scores and a 4,096-token context limit. The frozen final-answer extractor and Math-Verify 0.9.0 supply correctness; an unparsable final answer counts as incorrect. Model and dataset revisions, generation records, scoring records, and per-file hashes are retained in the analysis manifest. The target is final-answer correctness, not validity of every intermediate reasoning step.

Hardware and numerical precision. MATH-500 generation and PRM scoring were run as separate stages, each using one NVIDIA A800 80GB PCIe GPU. The recorded environment was PyTorch 2.8.0 with CUDA 12.8. PRM scoring used FP32 with TF32 disabled for both CUDA matrix multiplication and cuDNN. These hardware details concern the MATH-500 collection; the subsequent cached-bank analyses do not require GPU inference.

Null or context-overflow scores are unselectable, not imputed as zero. Score ties use fixed candidate IDs independent of labels; a challenger must strictly beat the first-eight reference to replace it. If all first-eight scores are invalid, the predefined fallback keeps candidate 0 and stops at eight. No confirmation task invokes this fallback. All 276 questions remain in accuracy and risk denominators. Their selected scores are finite at all four main budgets, so the main reward curve uses the same cohort. For the auxiliary $N = 1, 2, 4, 8, 16, 32, 64$ reward curve, the common finite-score cohort has 271 questions; its means must not be mixed with all-task means.

The confirmation archive contains 17,940 records including greedy completions: 17,683 finite PRM scores, 257 context overflows, 446 final-answer parsing failures, and 459 length-capped generations. These categories can overlap. Both incorrect $N = 64$ winners that replace correct references have valid PRM scores, parsed answers, and normal end-of-sequence termination. A separate comparison of extracted endpoint answers with benchmark answers agrees with the frozen labels for all seven endpoint changes. This checks those fourteen endpoint labels, not every candidate in the archive.

F.3 Targets and fixed rules

Let $c_{i,8}$ be question i's first-eight PRM winner, and $c_{i,N}$ its recorded-prefix winner. Define all-task harmful replacement and correction indicators

$$h_i(N) = \mathbf{1}[T(c_{i,8}) = 0,\ T(c_{i,N}) = 1], \qquad g_i(N) = \mathbf{1}[T(c_{i,8}) = 1,\ T(c_{i,N}) = 0],$$

where $T = 0$ denotes a correct answer. Their difference satisfies $\mathrm{Acc}(N) - \mathrm{Acc}(8) = \bar{g}(N) - \bar{h}(N)$ exactly. The conditional harmful-replacement rate divides the same harmful count by the 247 correct confirmation references instead of all 276 tasks. At $N = 64$ these rates are $2/247 = 0.81\%$ and $2/276 = 0.72\%$, respectively. They are not the selected-answer error rate $26/276$.

For each development question, rank integration holds its first-eight reference fixed and averages over uniform subsets of size $N - 8$ from the remaining 56 candidates. The theoretical budget is consequently $B = 1 + (N - 8)$, not N: the four evaluated B values are 1, 9, 25, and 57. Integrating harmful and beneficial winner labels gives $h_i^{\mathrm{rank}}(N)$ and $g_i^{\mathrm{rank}}(N)$; averaging the 224 development banks produces the frozen task-average forecasts. This is one global budget per policy, not a per-question risk predictor inferred from eight unlabeled answers.

The capacity point estimate chooses the largest N on $\{8, 16, 32, 64\}$ whose evaluated prefix of mean predicted harm remains at most $\tau = 0.05$. The rank-plus-net rule maximizes $\bar{g}^{\mathrm{rank}}(N) - \bar{h}^{\mathrm{rank}}(N)$

Table 13: Secondary post-hoc matched-target accounting of harmful replacement, in percentage points (rounding may affect the last decimal). Original forecasts and budget decisions are unchanged. Confirmation labels enter only these evaluation functionals.

N	Forecast minus prefix $d_N - p_N$	Cross-bank $d_N - c_N$	Within-bank $c_N - p_N$
16	1.04036	.62318	.41718
32	.91455	.95243	−.03787
64	1.95393	1.95393	.00000

among those admissible budgets, breaking ties toward smaller N. Prefix lookup uses the observed development-prefix means in the same rule. MC-plus-net replaces exact subset integration with 1,024 sampled subsets per development bank and budget. All three receive the same labeled banks. Capacity selects 64; all three net rules select 8. Fixed-8/16/32/64 are prespecified comparators, not choices retuned after confirmation. The development rule allocates total $\delta = 0.05$ as $\delta_h = 0.025$ for harm and $\delta_g = 0.025$ for net gain. Its harm Hoeffding radius at $n = 224$, $J = 3$, $\delta_h = 0.025$ is 0.103375 (rounded), larger than τ; its output of eight is the zero-additional-search fallback, not a useful certified expansion. Because development data were used to design the rules, this exercise is not an independent certification trial.

Prefix evaluation versus subset integration. The primary evaluation uses recorded prefixes; the forecast integrates uniformly resampled remaining-bank subsets. These are distinct conditional targets for a realized bank. They match in expectation over exchangeable candidate generation conditional on the first eight, not pointwise for each question. A secondary confirmation audit applies the exact rank identity to its fully labeled cached banks: all-task harmful risks at $N = 16, 32, 64$ are 0.7795%, 1.0491%, and 0.7246%. At 64 the full-bank prefix and rank results coincide exactly. These secondary bank calculations use test labels for evaluation only; they are not the forecasts used to choose budgets.

Table 13 separates cross-bank transfer from the realized prefix–subset difference. Let d_N denote the frozen development rank forecast, c_N the post-hoc confirmation rank integral, and p_N the observed confirmation prefix rate. The accounting identity $d_N - p_N = (d_N - c_N) + (c_N - p_N)$ requires no new model fit. The first difference compares the same subset functional across the two question banks; the second compares different sampling targets on the same confirmation bank. Neither component alone proves population distribution shift. At the full bank the sampling-target difference is exactly zero, so it cannot explain the remaining forecast error.

F.4 Confirmation outcomes and uncertainty

The sole prespecified primary utility comparison is accuracy at $N = 8$ minus accuracy at $N = 64$. Of 276 paired tasks, 245 are correct under both budgets, 24 under neither, two only at eight, and five only at 64. The difference is $-3/276 = -1.0870$ percentage points, with 95% paired task-bootstrap interval $[-2.8986, 0.7246]$ and two-sided exact McNemar $p = 0.453125$. Bootstrap resampling uses 50,000 draws (seed 2026091604), keeping each question's whole bank together. This interval is not a noninferiority test; no noninferiority margin was specified. Pointwise curve intervals are descriptive, not simultaneous certificates.

Table 14: Confirmation prefix outcomes. All columns use 276 questions. Harm and gain are relative to the first-eight winner; oracle means at least one correct answer in the prefix.

N	Correct	Oracle	Harm	Gain	Mean PRM
8	247	261	0	0	.82346
16	250	263	1	4	.83398
32	248	267	3	4	.84332
64	250	268	2	5	.85153

Rank integration removes conditional subset-sampling variance, not uncertainty about new questions. High-sample Monte Carlo closely approximates it here, and the same-information net rules

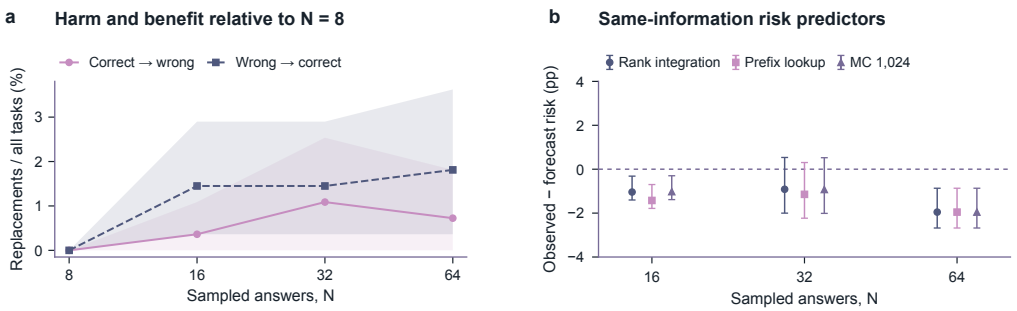

Figure 8: **Replacement accounting and same-information forecasts.** (a) Harmful and beneficial replacements relative to $N = 8$, over all 276 tasks. (b) Observed prefix harm minus each frozen development forecast, in percentage points. Bars and shading are pointwise 95% task-bootstrap intervals conditional on the frozen forecast; they exclude uncertainty from development-set estimation and method selection.

Table 15: Forecast error on confirmation prefixes: mean absolute error in percentage points over $N = 16, 32, 64$. These three budget points are correlated summaries, not independent trials.

Frozen development estimator	Harm	Gain	Net accuracy change
Exact rank integration	1.303	.317	1.416
Prefix lookup	1.508	.531	1.292
MC, 1,024 subsets	1.301	.311	1.430

produce identical decisions. The frozen rank forecast overestimates confirmation harm at all three expansion budgets. Its empirical risk-capacity choice of 64 has observed harm below 5%, but that observation does not convert a point estimate into a coverage guarantee. The Gaussian model was not fitted to these scalar PRM observations.

Development and pooled descriptions. On the inspected development set, $N = 8$ and 64 give 197/224 and 194/224 correct, with six harmful and three beneficial replacements. The corresponding confirmation counts are 247/276 and 250/276, with two harmful and five beneficial replacements. Pooled over all 500 tasks, both endpoints therefore give 444/500; the pooled equality is descriptive and does not replace the independent confirmation comparison. Auxiliary integer-prefix replay finds six confirmation tasks harmed at some $N \in \{8, \ldots, 64\}$, four of which recover by 64. This path event differs from endpoint harm and is not an anytime guarantee or a data-selected stopping policy.

Compute accounting and replication unit. The first-eight policy uses eight rather than 64 sampled candidate calls per question; mean output tokens are 5,307.5 versus 42,358.3, a replay reduction of 87.47%. The actual experiment generated the entire bank. These counts exclude development-bank collection, model loading, and batching overhead and are not measured GPU-time or monetary savings. The confirmation sample size is 276 question–bank pairs, not 17,940 candidate records. Task bootstrap intervals condition on these realized banks; they do not include all regeneration uncertainty or guarantee performance on another benchmark, model, or verifier.

G Fixed-Pool Operational Audit of LLM Best-of-N Selection

The fixed-pool operational audit uses the GSM8K test split (Cobbe et al., 2021). The saved 512-task manifest retains the original 64 pilot indices and adds 448 indices. The extension code samples uniformly from the complement and writes the manifest before generation. Qwen2.5-Math-1.5B-Instruct (Yang et al., 2024) produces one deterministic greedy completion and 16 stochastic challengers per task (temperature 0.7, top-$p = 0.95$, and a 384-token generation limit). We score all

17 responses with the final-step value of Skywork-o1-Open-PRM-Qwen-2.5-1.5B (He et al., 2024). Model and dataset revisions are pinned in the experiment script and recorded with the outputs.

The greedy response is the incumbent. A budget $B \in \{1, 2, 4, 8, 16\}$ adds $B - 1$ challengers sampled uniformly without replacement from the fixed per-task bank; the exposed PRM selects the highest-scoring response. An independent exact numeric checker supplies the target label. A target-worse replacement therefore occurs precisely when a correct greedy incumbent is displaced by an incorrect higher-PRM response. Conditional risk is reported over tasks with correct incumbents; selected-answer accuracy uses all tasks. Equation 10 integrates the finite schedule exactly, while 5,000 independently seeded schedules per task and budget provide a direct implementation audit. Confidence intervals bootstrap tasks, rather than treating the many resampled schedules as independent data.

Aggregate risk can conceal sparse task support, so we also report four concentration diagnostics at $B = 16$: the number of correct-incumbent tasks with nonzero risk, the shares contributed by the largest one and two tasks, the leave-largest-task-out mean, and the effective number of risk-bearing tasks $n_{\text{eff}} = (\sum_i r_i)^2 / \sum_i r_i^2$. These statistics distinguish a reproducible tail phenomenon from an average determined by one exceptional problem.

For task i, the rank identity gives cached-bank risk $r_i(B)$. The all-task mean is $512^{-1} \sum_i r_i(B)$; conditioning on a correct greedy incumbent divides the same numerator by 394. At $B = 16$, the two risks are 0.02649 and 0.0344, respectively; the latter's 95% task-bootstrap interval is $[0.0176, 0.0531]$. PRM-selected accuracy is 0.8762, versus greedy accuracy 0.7695, and oracle availability is 0.9399. Seventeen tasks carry nonzero risk; the largest contributes 7.4% of its total, and removing it leaves conditional mean risk 0.0320. Exact integration versus direct schedules has conditional-risk RMSE 0.00138 and accuracy RMSE 0.00230 over the four nontrivial budgets $B \in \{2, 4, 8, 16\}$; $B = 1$ is excluded from these aggregates. Oracle availability includes the greedy reference. There are 16 challengers plus that reference, so $B = 16$ samples 15 challengers and $B = 17$ includes the full bank.

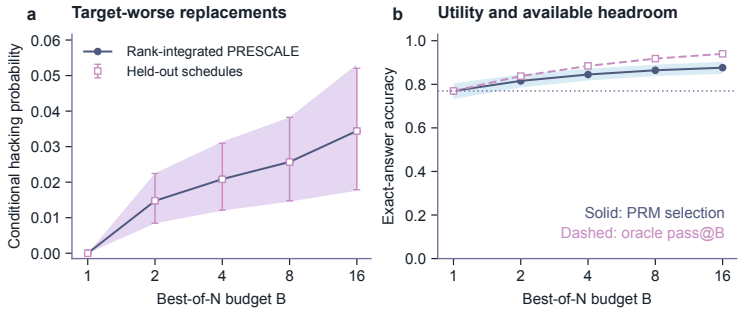

Figure 9: **The original GSM8K fixed-pool rank audit.** (a) Harmful replacement conditional on 394 correct greedy incumbents, from exact rank integration and direct schedules. (b) Selected-answer accuracy and oracle availability over 512 tasks. Bands and bars are 95% task-bootstrap intervals.

Table 16: Fixed-pool Best-of-N results. Risk is conditional on the 394 correct greedy incumbents; accuracy and oracle pass rate use all 512 tasks. The added $B = 17$ endpoint is a deterministic full-pool check.

B	PRESCALE risk	Schedule risk	PRM accuracy	Oracle pass@B
1	.0000	.0000	.7695	.7695
2	.0148	.0148	.8157	.8378
4	.0209	.0208	.8449	.8844
8	.0257	.0257	.8644	.9179
16	.0344	.0344	.8762	.9399
17	.0355	.0355	.8770	.9414

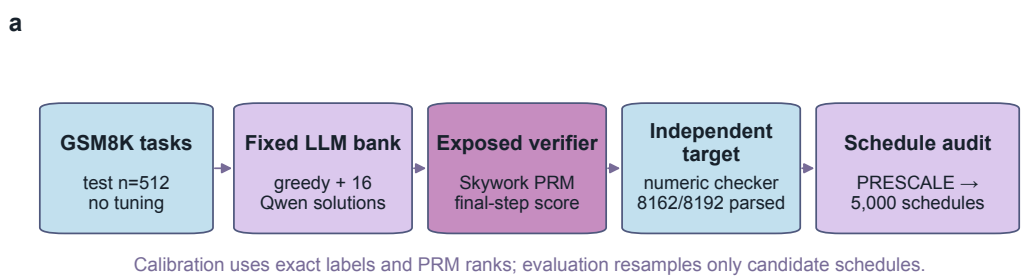

Figure 10: **Fixed-pool LLM operational audit.** Candidate generation, exposed PRM ranking, and independent target checking are separate operations. Resampled subsets audit the exact rank calculation on the frozen bank; they are not evidence of out-of-bank generation transfer.

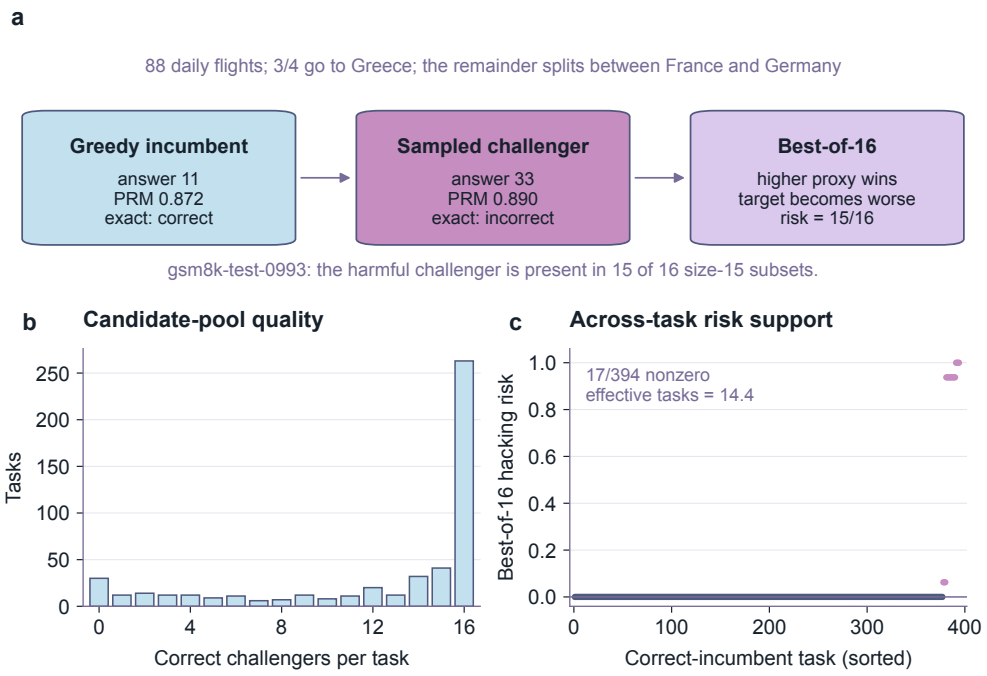

Figure 11: **Response-level mechanism and task-level support in the fixed-pool LLM audit.** (a) On GSM8K test item 993, a correct greedy answer is displaced by a higher-scoring wrong answer; at $B = 16$, the challenger occurs in 15 of 16 size-15 subsets. (b) Distribution of correct challengers in each 16-sample bank; 263 of 512 banks contain 16 correct answers. (c) Exact Best-of-16 risk over 394 correct-incumbent tasks. Seventeen tasks have nonzero risk, with effective risk-task count 14.37; the largest one and two tasks contribute 7.4% and 14.7% of aggregate risk.

This design tests the paper's selected-winner and rank-integration interface in an actual language model stack. It identifies fixed-pool selection behavior; a changed generator, verifier, target channel, or within-run policy requires a new calibration object.

H Additional Implementation and Calibration Checks

H.1 Retrospective interpretation of the original PDE audit

The archived 300-task paired pilot increased the budget from 4 to 32 and applied separately frozen acceptance thresholds. At $B = 32$ it recorded one accepted structural error among 87 accepted responses, versus zero among 86 at $B = 4$ (`results/f1-g0c-test-only-pilot-v1/`

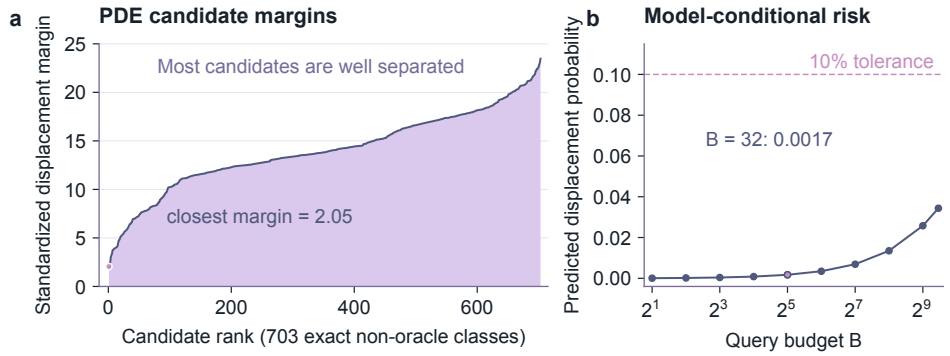

Figure 12: **Candidate margins explain the PDE operating point.** (a) Sorted margins for 703 exact non-oracle classes. (b) At the development-RMS reference scale, $B = 32$ has model-conditional risk 0.0017. This unthresholded risk is distinct from the historical pilot's accepted-error endpoint.

`analysis.json`). These selective-error summaries do not reconstruct the unthresholded harmful-winner count; the archived task-level records are unavailable. They are not a direct evaluation of $\mathcal{R}_{32}$.

The diagnostic supplies retrospective model-conditional context at the same development-RMS noise scale, rather than a prediction frozen before that pilot. It assigns $\mathcal{R}_{32} = 0.00172 \pm 0.00007$ Monte Carlo standard error; an independently seeded 50,000-event simulation gives 0.00178 ± 0.00019 (`results/geometry-diagnostic-v1/phase_diagram.csv`, empirical geometry, $B = 32$, noise multiplier 1). For 300 independent unthresholded selections under this model, the probability of at least one harmful winner is 0.403; this calculation does not evaluate the pilot's selective-error endpoint.

Figure 12 exposes the mechanism. The nearest candidate has standardized margin 2.05, but the median margin is 13.83; most randomly proposed candidates have negligible probability of crossing the incumbent. The complete-bank risk remains only 0.034 at $B = 704$ under the base noise model. The low model-implied risk follows from the mass of the margin distribution, rather than from query count alone.

H.2 Explicit subset-and-winner check

The original simulation samples a Bernoulli event from the same conditional hit formula used by the estimator. We therefore added a separately coded subset-and-winner oracle. On a four-challenger bank, all 81 gap vectors in $\{-1, 0, 1\}^4$, 15 nonempty harmful masks, and five budgets give 6,075 probability comparisons from 19,440 explicitly enumerated subset selections. The maximum discrepancy is 1.11×10^{-16}. Another 75 public-interface comparisons use direct quadratic scores with nonzero incumbents, mixed labels, duplicate scores, zero gaps, and deterministic errors. Ties use global candidate index, with the incumbent retained at zero gap.

For the existing empirical 703-candidate bank, we fixed budgets $\{2, 32, 704\}$ and noise multipliers $\{1, 2.8, 5.6\}$ before the check. Each regime has 5,000 explicit subset-and-winner selections. The same 5,000 evaluation error vectors are reused across regimes, so the 45,000 selections are not 45,000 independent error vectors. A second independent set of 5,000 errors supplies integrated predictions. The maximum same-error paired difference is 0.002395; full-pool $B = 704$ agrees on every draw. The largest independent-error difference is 0.0166. For nine contrasts, a two-sided Hoeffding union bound gives radii 0.02427 (conditional subset differences) and 0.03432 (two independent means), each rounded upward to five decimal places and each at familywise level 0.05. All differences lie within these conservative bounds; this is an implementation check, not a statistical-equivalence or real-error-model validation claim. The ten geometry unit tests pass in the accompanying execution record. The Gaussian software entry point requires positive noise scale; zero-error checks use its empirical-error interface. An all-benign bank has mathematical risk zero and may be returned directly rather than passed to the current nonempty-harmful-mask API.

H.3 Exact independence baseline and archived approximations

Baselines. For $k = B - 1$, let $p_c = \overline{\Phi}(\|r_c\|/(2\sigma))$ and $\bar{p} = M^{-1}\sum_c p_c$. The exact independent-projection baseline is

$$R_B^{\mathrm{ind}} = 1 - \frac{1}{\binom{M}{k}} \sum_{S:\,|S|=k} \prod_{c\in S}(1-p_c). \tag{17}$$

It preserves all candidate marginals and uniform sampling without replacement while removing shared-error dependence. A normalized elementary-symmetric recurrence computes it exactly. The union-bound baseline is $\min\{1, k\bar{p}\}$. The nearest-margin comparator is $1-(1-p_{\max})^k$, a hypothetical homogeneous bank of distinct challengers each having the most vulnerable margin, not repeated draws of the same candidate. Both receive the current noise scale. Budget-only calibration instead reuses the empirical bank's σ_0 risk curve at every geometry and noise level; it removes scale as well as geometry information. Appendix H.3 records the recurrence and the historical Poisson approximations separately.

For $q_c = 1 - p_c$, define $a_{j,k} = \binom{j}{k}^{-1} \sum_{S\subseteq\{1,\dots,j\},|S|=k} \prod_{c\in S} q_c$ for $0 \le k \le j$. Set $a_{0,0} = 1$, $a_{j,0} = 1$, and $a_{j,k} = 0$ outside $0 \le k \le j$. For $j \ge 1$ and $1 \le k \le j$, the stable recurrence is

$$a_{j,k} = \frac{j-k}{j} a_{j-1,k} + \frac{k}{j} q_j a_{j-1,k-1}.$$

Then $R_B^{\mathrm{ind}} = 1 - a_{M,B-1}$. This preserves the finite-bank marginals and without-replacement design. The earlier code instead used $1-\exp(-k\bar{p})$ and $1-\exp(-kp_{\max})$. Those are Poisson approximations, not exact independence controls; the revised Table 4 and Figure 5 use exact expressions. The original results remain archived, with independent-approximation RMSE 0.12184 and nearest-approximation RMSE 0.55228. No generation or verifier scoring was rerun for this correction.

H.4 Calibration inputs

Obtaining the calibration object. Calibration requires independent complete score-gap vectors on the same frozen labeled bank, or independent errors in a specified residual measurement model. Direct gap vectors suffice: recovering a latent error vector is unnecessary. The finite-sample bound covers sampling uncertainty under their common law. Monte Carlo draws from a fitted Gaussian model certify that fitted-model risk only; uncertainty in the fitted law requires an additional bound. Our controlled study stipulates the error law, and PDE development residuals set its scale. The LLM study conditions on each observed score vector and separately evaluates a development-to-confirmation task-average forecast. Different questions are not repeated error draws on one fixed bank, and a scalar PRM score does not identify a latent covariance. Operationally, repetition i must supply aligned gaps $G_c^{(i)} = V(c; e_i) - V(c_0; e_i)$ for every candidate under one shared verifier realization. Writing A for the matrix with rows $\Delta_c^\top$, the quadratic Gaussian model gives $G = \gamma - 2Ae$ and $\mathrm{Cov}(G) = 4A\Sigma A^\top$. The gap law is the risk-relevant object; its latent factorization need not be unique. Repeating a deterministic PRM on unchanged text supplies no new error realization, while input perturbations define a different, explicitly specified observation law.

Worked structural-label example. Our conservative one-dimensional equation is $u_t = -\frac{1}{2}\partial_x(u^2) + \frac{3}{32}u_{xx} + q(u; m)$, with metadata $m = (m_1, m_2)$ and reference correction

$$q_*(u; m) = -\tfrac{1}{8}(1 + m_1/4)\partial_x(u^2) + \tfrac{1}{32}(1 + m_2/4)\partial_x(uu_x). \tag{18}$$

A challenger replaces $1 + m_2/4$ by $1 + m_1 m_2/4$. Its exact discrepancy is $m_2(m_1 - 1)(u_x^2 + uu_{xx})/128$: it is structurally wrong but vanishes at the two intervention corners with $m_1 = 1$. Four corners and 16 normalized weak moments per corner form its 64-dimensional residual, with norm 0.209066 and 32 zero coordinates. Rational canonicalization yields 1,088 classes; principal-part and positivity checks exclude 384, leaving the reference and 703 harmful challengers.

H.5 Development scale and measurement coordinates

Each of three development trajectories contains 33 snapshots on 512 spatial points. Eight four-interval windows and 16 sine/cosine probes per window yield 384 weak coordinates. For coordinate a, let y_a be the weak endpoint increment and q_a the composite-Simpson integral of the reference right-hand side. With window duration $\Delta = 1/16$,

$$s_a = \max\{\Delta/8, |y_a|, |q_a|\}, \qquad e_a = (y_a - q_a)/s_a, \qquad \sigma_0 = \sqrt{384^{-1}\sum_{a=1}^{384} e_a^2} = 0.0443171141.$$

This is an uncentered RMS, not an IID covariance estimate. Separately, the ideal candidate vectors contain 64 normalized weak differences under the four metadata interventions. For the challenger in Equation 18, the first four coordinates at $(-1,-1)$ are $(-0.001772, -0.000129, -0.021802, 0.014914)$; the full norm is 0.209066.

Ideal candidate coordinates. For intervention j and probe a, let $Q_{c,ja}$ and $Q_{*,ja}$ be the weak moments of the candidate and known corrections, and $F_{*,ja}$ the weak moment of the full known right-hand side. The frozen bank uses

$$s^*_{ja} = \max\{s_{\mathrm{floor}}, |F_{*,ja}|\}, \qquad r_{c,ja} = (Q_{c,ja} - Q_{*,ja})/s^*_{ja}.$$

Four interventions and sixteen probes give $r_c \in \mathbb{R}^{64}$ and $r_{c_0} = 0$. The frozen builder uses $s_{\mathrm{floor}} = 1/8$. If Y_{ja} is a noisy observation of the full weak moment, then $e_{ja} = (Y_{ja} - F_{*,ja})/s^*_{ja}$ yields residual $r_{c,ja} - e_{ja}$. Repeated independent observations would supply complete error vectors. The controlled experiments instead impose the stated law; the 384 development coordinates above determine its scale, not an IID sample of 64-dimensional errors. The supplied bank construction, estimated scale, and imposed error law are thus separate inputs.

H.6 Full-pool LLM endpoint

At $B = 17$ all 16 cached challengers are available. Directly taking the highest PRM score gives 449 correct winners out of 512 and oracle availability 482/512. Among 394 correct incumbents, 14 become incorrect; among 118 incorrect incumbents, 69 become correct. Thus harmful replacement is $14/512 = 0.02734$ over all tasks and $14/394 = 0.03553$ conditional on incumbent correctness. These are deterministic cached-bank endpoints, not new independently generated evaluation data. The $B \le 16$ schedule RMSEs remain the four-nontrivial-budget checks ($B = 2, 4, 8, 16$); neither $B = 1$ nor $B = 17$ enters them.

I Applicability Contract and Extensions

Scope and classical ingredients. This dependence-sensitive setting does not contradict ROC sufficiency under IID sampling. Gaussian extrema, conditional expectation, union bounds, and Hoeffding concentration are classical tools, not new inequalities claimed here. The construction concerns candidate-level score-gap marginals under shared uncertainty, not a general failure of ROC characterization.

The geometry layer requires a specified error law. Under Gaussian measurement error, standardized margins and their correlation matrix determine the all-harmful crossing probability. Mixed-bank winner risk also requires gap scales and target labels; non-Gaussian averaging uses the full specified joint score-gap law. A scalar verifier score does not identify a latent covariance (Appendix H.4); obtaining aligned complete score-gap vectors requires repeated verifier realizations on one frozen bank. Our controlled study stipulates the law and takes its scale from PDE development residuals; the LLM study exercises only the rank layer, which needs observed score vectors and no Gaussian assumption. In a mixed bank, gap scales and the harmful mask are also required to resolve winner competition.

Every capacity statement is conditional. Capacity is indexed by the reference, the labeled bank, the sampling policy, the independent target channel, and the joint error law. A generator update

changes the bank, a verifier update changes the error law, and a target-channel update changes the harmful mask; each calls for recalibration. The checkpoint study quantifies this cost but does not remove it, and the finite-bank formula does not cover within-run adaptive proposal generation.

The identifying bank is degenerate in one respect. All 703 PDE challengers are harmful, so every replacement counts as a failure and pointwise risk is monotone in B. Mixed banks need not be monotone, because a benign winner can shield the incumbent; this is why the running envelope of Equation 1 is required. Mixed-bank evidence consists of the explicit implementation checks and the LLM ranking evaluation, not a matched-margin LLM geometry intervention.

The certificate is conservative at small calibration sizes. At $\delta = 0.05$ and $n = 256$, the additive Hoeffding bound withholds all expansion in the controlled finite-calibration study; on MATH-500 its radius exceeds the tolerance, so certification returns the zero-additional-search fallback (Appendix F). Appendix D.1 shows that a fixed-sequence rule retains power on the same controlled data. This separates the conservatism of the additive bound from the availability of an admissible budget; it does not establish a useful LLM expansion certificate.

Empirical scale is measured at specific operating points. Harmful replacement is 0.36–1.09% over all 276 MATH-500 confirmation tasks with a 7B policy and a 7B process reward model, and 3.44% conditional on 394 correct GSM8K incumbents with a 1.5B policy and a 1.5B process reward model (Appendix G). These settings differ in model, benchmark, verifier, reference, sampling, and denominator; the contrast is an observation, not a controlled comparison. They are not estimates of a universal rate, and they do not establish that reward hacking is negligible in general. On MATH-500, development forecasts exceed recorded confirmation-prefix harm at all evaluated expansion budgets. This direction of error is descriptive, not a coverage guarantee for live deployment; forecast subset sampling and recorded-prefix evaluation also have distinct targets.

The primary utility comparison does not reject equality. On MATH-500 the prespecified $N = 8$ minus $N = 64$ accuracy difference is -1.09 percentage points with a 95% paired bootstrap interval $[-2.90, 0.72]$ and exact McNemar $p = 0.453$. No noninferiority margin was prespecified, so this interval does not establish noninferiority either. Harm at $N = 8$ relative to its own first-eight reference is zero by construction, not an empirical safety advantage of the diagnostic.

Every capacity statement is indexed by four calibrated objects: the reference and labeled candidate bank, the candidate-sampling policy, the independent target channel, and the joint verifier-error law. A generator update changes the bank, a verifier update changes the error law, and a target-channel update changes the harmful mask; each calls for recalibration. The guarantee then applies to the frozen search policy, with explicit shift allowance when calibration and live distributions differ.

Mappings beyond PDE discovery. The reduction operates on pairwise score perturbations rather than PDE semantics. For proof search, for example, V can be a learned proof-score model and T kernel checking plus statement equivalence; for a quadratic latent scorer, Δ_c is the difference between whitened candidate and incumbent proof embeddings. Across domains, small margins call for generator constraints or an independent filter, large projected uncertainty calls for verifier improvement, and low capacity despite safe marginals calls for modeling joint correlation.

Table 17: Reference-relative mappings of the geometric-capacity interface.

Domain	Current reference c_0	Exposed verifier V	Independent target channel T
Proof search	Best accepted derivation	Learned process or proof score	Kernel checking plus statement equivalence
Program synthesis	Current implementation	Public tests or neural critic	Hidden tests or formal specification
Combinatorial design	Current feasible design	Surrogate objective and constraint score	Exact constraints plus high-fidelity evaluation

Mixed harmful and benign banks. The all-harmful bank makes the correlation mechanism identifiable, but the implementation supports a boolean harmful mask and arbitrary incumbent. In a mixed bank, benign challengers can prevent a harmful negative-gap candidate from being selected, so pointwise risk need not be monotone in B. For example, with one harmful challenger above the incumbent and one still higher-ranked benign challenger, the harmful-selection risks at $B = 1, 2, 3$ are $0, 1/2, 0$. Equation 6 is therefore the estimand, and the running envelope in Equation 1 is necessary.

Non-quadratic scores. For a smooth verifier, let $g_c(e) = V(c; e) - V(c_0; e)$. Around the calibration state,

$$g_c(e) = g_c(0) + a_c^\top e + O(\|e\|_2^2), \qquad a_c = \nabla_e g_c(0). \tag{19}$$

For the linearized gaps, crossings use $a_c^\top e < -g_c(0)$ with joint covariance $a_c^\top \Sigma a_{c'}$. This is a local approximation, not the exact quadratic reduction. Exact observed score gaps can instead be ranked directly in Equation 10. Transferring a certificate from the linearization to the original scorer requires an explicit remainder bound.

Distribution shift. Let P_{cal} and P_{live} denote calibration and live distributions over the labeled selection event. Since its indicator is bounded in $[0, 1]$, any fixed search policy obeys

$$\mathcal{R}_B(P_{\mathrm{live}}) \le \mathcal{R}_B(P_{\mathrm{cal}}) + \mathrm{TV}(P_{\mathrm{live}}, P_{\mathrm{cal}}). \tag{20}$$

A certificate at $\tau - \epsilon$ therefore transfers at tolerance τ when the shift is at most ϵ; covariate-weighted calibration is another route when density ratios are estimable (Tibshirani et al., 2019).

Adaptive proposal generation. If later proposals depend on earlier verifier outputs, a single static bank is insufficient. One can condition on the search history and evaluate the induced distribution over terminal banks, but the present finite-bank formula does not by itself provide that sequential model. Periodic recalibration, as in the checkpoint study, handles exogenous generator updates but not arbitrary within-run adaptation.

Hidden-channel integrity. The calibration target must remain independent of the exposed selection score. If candidate generation is optimized directly against the hidden channel, that channel becomes another exposed proxy and the target-worse labels can cease to be reliable. This is an evaluation-design condition, not a numerical property of PRESCALE.

J REPRODUCIBILITY DETAILS

Start with saved-evidence verification. From the repository root, run `python scripts/run_final_cpu_checks.py`. It checks hashes, recomputes the quantities supported by Table 18, and runs 54 tests without generating observations, calling models, or accessing the network. `ARTIFACT_REGISTRY.json` gives full paths and command scopes.

Current accounting entry point. For already observed complete score vectors, use `scripts/cached_rank_20260920.py`: `integrate(gaps, harmful, draws)` accepts matching row-by-challenger arrays and a tuple of challenger draw counts, returning row-by-budget risks. Gaps are lower-is-better; zero retains the reference, challenger ties follow fixed column order, and positive infinity is unselectable but remains in the sampling denominator. This computes conditional risk, not a certificate. The README demonstrates the separate `decisions` calibration API on saved D.2 rows.

Historical provenance and corrections. Historical code, the 16-test log, and original/updated independence baselines remain separate (Appendix H.3). MATH receipts are local, not external timestamps. D.1 preserves 32 paired decisions and 704 traces, not independent UCB/test replications. D.2 preserves the auxiliary correction and unchanged predictions; D.3 preserves both protocols and the timing reversal.

Table 18: Artifact registry. Scripts are under `scripts/`; full input/output paths are in `ARTIFACT_REGISTRY.json`. Historical generation entry points document provenance, not default verification commands.

Study	Script and existing input	Saved output and paper location
Controlled geometry	`run_geometry_diagnostic.py`; historical bank construction; packaged verification uses per-regime tables	`geometry_variants.csv`, `phase_diagram.csv`; corrected `baseline_geometry_variants_exact.csv`. Figure 2(a); Appendix H.3, Figure 5.
Heavy tails	`run_nongaussian_extension.py`; bank, seeds and sampling laws in that source	`robustness.csv` and `analysis.json`. Appendix A, Figure 3.
Finite calibration	`run_finite_calibration_study.py`; packaged verification uses replicate tables	`replicates.csv` and `analysis.json`. Appendix D.
Checkpoint drift	Historical `run_pythia_checkpoint_landscape.py`; verification input: `candidate_scores.csv`	`risk_curves.csv` and `analysis.json`. Appendix E.
LLM Best-of-N	`run_llm_best_of_n.py`; `scored_candidates.jsonl`	`task_metrics.csv` and `risk_curves.csv`. Appendix G, Table 16.
MATH confirmation	`verify_llm_archives.py`; `confirmation-candidates-evaluated.jsonl`	`task-by-budget.csv`, `summary.json` and `qa-verification.json`. Section 5.4; Appendix F.
D.1: calibration rules	`run_cpu_calibration_comparison.py`; `calibration-means.json`; matched tests: `run_matched_marginal_sequence_20260920.py`	`paired-decisions.csv`, `test-trace.csv`, original `replicates.csv` and `reference.csv`. Appendix D.1, Table 2.
D.2: measurement bridge	`measurement-bridge-20260920/run.py`; `fixed-inputs.npz` and `protocol.json`	`calibration.npz`, `evaluation.npz`, `frozen-predictions.json`, `evaluation.json` and `audit.json`. Figure 2(b); Appendix D.2, Table 9.
D.3: cached rank	`cached_rank_20260920.py`; the MATH confirmation scores above; `protocol.json` and `batched-protocol.json`	`exact-task-risks.csv`, `repeats.csv`, `batched-repeats.csv` and both MC estimate arrays. Appendix D.3.

Delivered package scope. Fresh extraction of the anonymous package passed all 54 tests. Registry levels are **L1**: table aggregation (geometry, heavy tails, finite calibration, checkpoints); **L2**: score/statistic recomputation (GSM8K, MATH, D.1–D.3; timings remain L1); and **L3**: input-level regeneration (not run). Missing original 703-by-64 residuals block controlled regeneration; specified labels and transformations alone are insufficient. D.1 preserves means, not raw draws. D.2 includes minimal trajectory inputs and protocol, but verification reuses scores. README details these boundaries and the exclusion of identities, private material, weights, sealed data and GSM8K text.

Computational complexity. For n vectors, M candidates, p coordinates and J budgets, projections cost $O(nMp)$; given weights, integration costs $O(n(M+J))$ for all-harmful banks or $O(nM\log M + nJM)$ with mixed-bank sorting. Caching removes historical $O(LJM^2)$ validation over L calls, building weights in $O(JM)$ per bank-size/grid pair; scoring cost is unchanged. Batch-q working storage is $O(qp + qM + Mp + JM)$, excluding covariance, sampling and generation. Gaussian sampling adds $O(np)$ isotropically or $O(np^2)$ time and $O(p^2)$ storage with a supplied dense factor; factorization is separate. Candidate/verifier updates require affected projections and ranks; grid-only updates reuse ranks and labels.