SAME ERROR RATES, DIFFERENT CERTIFICATES: PAIRED AUDITS OF PROCESS REWARD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Verification protocols with the same marginal error rates can provide different evidence for a certified comparison when their joint errors differ. We study this distinction in process reward models using seven fixed protocols, two related checkpoints, and design-based finite-population risk guarantees. On 235 previously unscored problems, a predeclared audit with a 192-problem budget finds that paired error bounds increase valid near-optimal certificate returns by 24.41 and 24.22 percentage points over marginal risk subtraction. Simultaneous conditional Monte Carlo lower bounds are 19.41 and 19.23 points. Permuting comparator error blocks while preserving each protocol's class- and domain-specific error counts nearly removes this advantage. Before new inference, forecasts retaining joint errors predict the relevant certificate probabilities with mean absolute error 0.0153, versus 0.1284 after marginal-preserving shuffling. Component-level analysis shows when absolute-risk requirements and candidate-selection costs remain the active certification bottlenecks. In the single prospective certificate draw, corrected multi-candidate testing certifies one checkpoint, while an unresolved absolute-risk requirement determines the other outcome. These results identify information omitted by marginal accuracy and separate protocol quality from evidence adequacy. The guarantees apply to the specified finite benchmark rosters and sampling designs.

1 INTRODUCTION

Shared errors are part of the evidence for comparing verification protocols. A process reward model (PRM) judges intermediate reasoning steps, while an evaluator must determine whether a verification rule meets explicit quality requirements. Process supervision and step-level evaluation make this question increasingly concrete (Lightman et al., 2024; Zheng et al., 2025; Zhang et al., 2025). Marginal accuracy describes each rule separately; the errors that rules share determine how informative their comparison can be.

Consider two fixed verification rules that disagree on only a few labeled steps. Estimating each error rate separately and subtracting confidence bounds ignores the cancellation of their common errors. Directly bounding their paired error difference can be more informative. Absolute quality remains a separate requirement: two rules can be close to each other while both have high risk. These are distinct questions about **absolute quality**, **relative quality**, and **the evidence required to certify them**.

We examine this distinction through a controlled audit of fixed PRM protocols. The unit is a mathematical problem with at most one correct and one first-error endpoint. Frozen rules combine a native PRM score with a history representation and a learned decision threshold. An audit samples complete problems without replacement and either certifies explicit risk targets or returns unresolved. Existing finite-sample risk-control and finite-population tools provide the validity layer (Angelopoulos et al., 2022; Waudby-Smith & Ramdas, 2020; 2024); our question is when the information supplied to those tools changes a real certification decision.

Three tests separate these questions. First, we compare paired and marginal bounds on *the same candidate, sample, target, and error budget*. Second, we permute comparator error blocks while preserving all marginal canonical error counts within class and source domain. This isolates statistical coupling at unchanged marginal risk. Third, we forecast certification probabilities from previously

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.

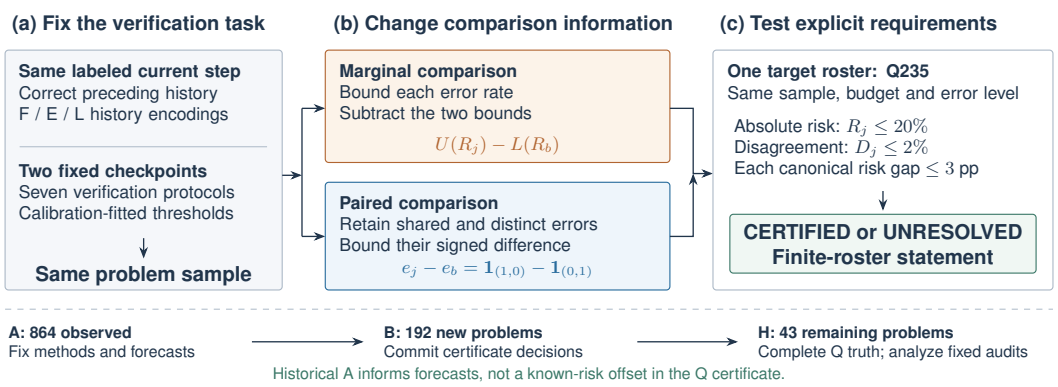


Figure 1: **Hold the verification rules fixed; change the information used to compare them.** Both paths use the same labeled endpoints, risk targets, and candidate-level error budget. The final prospective draw and its certificate decisions are sealed before the census remainder is observed. Historical data inform forecasts, but do not enter the new Q235 certificate as a known-risk offset. The coupling control in Figure 2 acts on error tables, not on model inputs.

observed joint errors, fix those forecasts before new model outputs, and evaluate them on a fully revealed holdout roster. Figure 1 separates these stages.

The final validation contains 235 previously unscored problems. Across 512 shared random samples at the predeclared 192-problem budget, paired atomic bounds improve valid near-optimal returns by about 24 percentage points for each checkpoint. A separate marginal-preserving control nearly eliminates the gain. Joint-error forecasts also substantially outperform their shuffled counterpart. Together, the comparisons establish *joint errors as decision-relevant information beyond marginal accuracy*.

The final study observes 192/235 problems (81.70%). Its prospective draw distinguishes unresolved fixed-protocol certificates from a certified alternative for M1 under corrected candidate selection. Absolute risk limits M2's complete certificates even when its relative-risk comparisons pass. Matched classical interval references further separate the value of paired information from the calibration of a particular statistical procedure.

2 Related Work

Process supervision and reward-guided reasoning. Outcome verifiers support answer selection on mathematical problems (Cobbe et al., 2021), while process supervision evaluates intermediate reasoning rather than only final outcomes (Uesato et al., 2022; Lightman et al., 2024). Math-Shepherd constructs process supervision automatically for verification and reinforcement learning (Wang et al., 2024). Under a specific reward parameterization, process rewards can also be derived from outcome-supervised training (Yuan et al., 2025). PRM-guided search is one mechanism for allocating test-time compute (Snell et al., 2025). These studies improve supervision or inference; we instead freeze the verifier and ask whether its protocol satisfies explicit audit requirements.

PRM evaluation. ProcessBench identifies the earliest erroneous step in mathematical solutions; PRMBench evaluates finer-grained process-reward behavior (Zheng et al., 2025; Song et al., 2025). Work on developing PRMs distinguishes step verification from completion-based value estimates and Best-of-N evaluation (Zhang et al., 2025). CoLD studies verbosity-related bias and counterfactual debiasing (Zheng et al., 2026). Our history encodings provide matched views of the same current step. We evaluate them through statistically qualified *audit decisions*, linking measured protocol quality to the evidence needed for certification.

Paired comparison and finite-population inference. Paired comparisons have a classical statistical foundation in correlated proportions and classifier evaluation (McNemar, 1947; Dietterich, 1998). Finite-population sampling theory and concentration inequalities establish the role of without-

replacement sampling (Cochran, 1977; Hoeffding, 1963). Confidence sequences, betting-based mean bounds, and variance-sensitive bounds extend the available tools (Waudby-Smith & Ramdas, 2020; 2024; Maurer & Pontil, 2009). We use exact hypergeometric event bounds and one frozen betting mixture. Newcombe's paired score intervals provide a classical risk-difference reference (Newcombe, 1998). We compare that nominal construction and a separately labeled finite-population width sensitivity on the same samples. Our contribution is the matched measurement of comparison information, interval calibration, and pre-output forecasting.

Risk control and multiple testing. Risk-controlling prediction sets and Learn then Test connect calibrated decisions to finite-sample guarantees (Bates et al., 2021; Angelopoulos et al., 2022). Multiple testing also offers Holm's step-down procedure and dependence-aware resampling procedures (Holm, 1979; Romano & Wolf, 2005). Our fixed-sequence test follows an order learned before the audit and stops at its first non-pass. Holm's procedure instead orders observed p-values. PEAK provides anytime composite-mean tests across bounded streams with stationary conditional means (Cho et al., 2024). Our without-replacement setting instead has a changing remaining mean, which defines the sampling condition used by the evaluated betting construction.

3 Verification protocols and certification targets

3.1 A fixed current-step task

We use ProcessBench's first-error annotations (Zheng et al., 2025). Its question sources include GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021), OlympiadBench (He et al., 2024), and Omni-MATH (Gao et al., 2024). The evaluated records and first-error labels come from ProcessBench. All supplied history precedes the first error. A selected current step is either correct ($y = 1$) or the first erroneous step ($y = 0$); all later steps are excluded. Exact normalized problem text defines a cluster. The frozen selection keeps at most one endpoint per class per cluster and requires at least three history steps. The resulting calibration set has 247 problems and 330 endpoints; the original audit roster P has 1,099 problems and 1,469 endpoints. Filtering, version pins, and class denominators are in Appendix A.

For h original history steps, F retains every original step boundary. E places a single internal history cut at $\lfloor h/3 \rfloor$, and L at $\lceil 2h/3 \rceil$, yielding two history blocks followed by the unchanged current step. E/L preserve non-boundary tokens, total length, three reward markers, and the current readout position. F preserves the current logical step with its original history boundaries and length. The native two-class score is read only at the current step's final reward marker.

The two checkpoints are Qwen2.5-Math-PRM-7B (M1) and Qwen2.5-Math-7B-PRM800K (M2). They share the Qwen2.5-Math backbone family (Yang et al., 2024); PRM development and checkpoint evaluation are described by Zhang et al. (2025); Zheng et al. (2025). Inference is frozen BF16/eager, batch one, with a 4,096-token limit and no truncation. The seven calibration-fitted protocols comprise three canonicalizations and four threshold rules (Table 7). All parameters and tie rules are fixed before the audit. $\textsc{Can}_L$ is its best canonical rule in both models and is the precommitted Fixed candidate. Canonicalization reconstructs the same selected encoding from either received view, so its E/L decisions agree and $D = 0$. Appendix A illustrates the boundaries for six history steps.

3.2 Three distinct risk requirements

For endpoint i and rule j, let $a_{jvi} \in \{0, 1\}$ denote acceptance under view $v \in \{E, L\}$, and define

$$e_{jvi} = \mathbf{1}\{a_{jvi} \neq y_i\}, \quad \ell_{ji} = \tfrac{1}{2}(e_{jEi} + e_{jLi}), \quad d_{ji} = \mathbf{1}\{a_{jEi} \neq a_{jLi}\}. \tag{1}$$

For a named finite roster $\mathcal{T}$, let $\mathcal{T}_y$ contain its selected class-y endpoints. The target risks are

$$R_j = \tfrac{1}{2}\sum_{y=0}^{1} \frac{\sum_{i\in\mathcal{T}_y} \ell_{ji}}{|\mathcal{T}_y|}, \qquad D_j = \tfrac{1}{2}\sum_{y=0}^{1} \frac{\sum_{i\in\mathcal{T}_y} d_{ji}}{|\mathcal{T}_y|}, \qquad G_j = \max_{b\in\mathcal{C}}(R_j - R_b), \tag{2}$$

where $\mathcal{C} = \{\textsc{Can}_F, \textsc{Can}_E, \textsc{Can}_L\}$. R measures balanced error across the two endpoint classes. Signed gaps may be negative.

An Abs claim requires $R_j \leq .20$ and $D_j \leq .02$. Near additionally requires $R_j - R_b \leq .03$ for *each* $b \in \mathcal{C}$. Thus "near-optimal" always means within three percentage points of the best of these three fixed canonical rules, within the declared comparator set. These targets remain unchanged throughout the final study. An unissued claim leaves the target unresolved.

3.3 VALID TOOLS, DIFFERENT INFORMATION

For a fixed binary event, n sampled class endpoints from a population of size N have a hypergeometric event count (Cochran, 1977). ATOM–MARG bounds candidate and comparator risks separately, then subtracts an upper and a lower bound. ATOM–PAIR instead uses

$$e_j - e_b = \mathbf{1}\{e_j = 1, e_b = 0\} - \mathbf{1}\{e_j = 0, e_b = 1\} \tag{3}$$

and bounds the harmful and beneficial events directly. Each scalar's error allocation is divided across its nonconstant one-sided event bounds; terms are removed only through symbolic identities. Within the atomic pair/marginal comparison, the absolute-risk and disagreement bounds are identical.

BET–MARG and BET–PAIR use the same fixed 19-component without-replacement betting mixture on bounded view-average losses. The paired version acts directly on a shifted signed gap; the marginal version subtracts two risk bounds. Their disagreement bounds remain hypergeometric. Appendix B gives all transformations and outward-certified lattice inversions.

Each fixed program allocates $\alpha = 1/40$ to each model. Fixed tests only the precommitted CAN_L; Seq tests the calibration-ordered candidates and stops at the first non-pass; Bonf tests all seven at $\alpha/7$; Split selects on one half of the sample and tests on the other with the correct remaining-population adjustment. An intersection–union argument controls the conjunction for a fixed candidate at its assigned level (Angelopoulos et al., 2022). The two-model false-certification probability is at most $1/20$ for each specified program, target, budget, and round. Empirical selection is reported as an uncertified reference.

4 FROZEN FORECASTS, PROSPECTIVE SAMPLING, AND CONTROLS

Earlier outcome-informed development leaves A, 864 already observed problems, and $Q = P \setminus A$, 235 problems whose model outputs remain unobserved. We freeze the final methods, all targets, the primary comparison, and forecasts using A before opening any Q output. The new certificate target is *Q itself*; A supplies historical forecasting information. One uniform random problem permutation allocates 192 problems to B and the remaining 43 to H. B decisions are fixed before H is evaluated. H then completes exact Q and P risks. Earlier certificates retain their original P targets and are reported separately.

Conditional validation after the census. The frozen audit is evaluated on 512 independently randomized problem permutations of Q, with shared prefixes at budgets 64, 128, and 192 across models and programs. Full-roster truth is available only to the scorer. A *valid return* is a returned protocol whose complete target conjunction is true. The primary comparison is Fixed–Near, ATOM–PAIR versus ATOM–MARG, at budget 192. The 512 randomizations estimate sampling power conditional on the fixed roster.

For each model, let $Z_r \in \{-1, 0, 1\}$ be paired minus marginal valid return in randomization r. We lower-bound $\Pr(Z = 1)$ and upper-bound $\Pr(Z = -1)$ using one-sided Clopper–Pearson limits (Clopper & Pearson, 1934), with failure level $1/80$ per tail. A union bound over four tails gives simultaneous 95% Monte Carlo lower bounds for the two conditional power differences. These bounds quantify uncertainty in the estimated power differences.

Forecast before observing the new errors. For each new problem's source domain and class-presence pattern, draw a whole-cluster donor from A, preserving both models, all present classes, and all three canonical error vectors. We generate 64 pseudo-rosters and use 32 sampling permutations per pseudo-roster. Joint retains these dependencies. Marginalized independently permutes the F and E comparator blocks within the same strata, holding L fixed. The two versions have identical marginal canonical error counts in every pseudo-roster. Forecasts of absolute-risk, relative-gap, and complete-Near pass probabilities are fixed before new inference. The primary mean absolute error

Table 1: **Paired errors turn comparison information into valid returns.** Fixed–Near at 192/235 problems, with the same CAN_L and $\alpha = 1/40$ per model. Counts are out of 512 shared random samples. The lower bounds concern conditional Monte Carlo power differences on Q, not future-task generalization.

Model	Marginal	Paired	Gain (pp)	Simultaneous MC lower
M1	0/512	125/512	+24.41	+19.41
M2	0/512	124/512	+24.22	+19.23

Table 2: **Paired interval construction changes return rates.** Complete returns at 192 problems, out of 512 shared samples per model. All rows use the atomic absolute-risk requirement. Newcombe and hFPC are post-hoc reference rules, not procedures covered by the finite-roster guarantee.

Relative-bound construction	M1	M2	Guarantee in this design
Atomic paired	125	124	Finite-roster error control
Newcombe method 10	0	1	Nominal reference
Newcombe + hFPC	191	131	Heuristic width adjustment

(MAE) equally weights 12 cells: two models, three budgets, and the relative-gap/complete-Near events for ATOM–PAIR. The forecasts are evaluated as predicted audit probabilities.

Vary coupling at fixed marginal risk. After revealing Q, create 32 additional block-permutation surrogates with the same marginal-preserving operation. All variants share 64 problem samples per budget. The control preserves each canonical protocol's finite class- and domain-specific error counts, hence its R and all canonical risk differences, while changing their joint errors. The 32 surrogate error tables share samples and provide a descriptive sensitivity to statistical dependence. Analogous controls on A and complete construction rules appear in Appendix C.

5 RESULTS

5.1 PAIRING CHANGES CONDITIONAL CERTIFICATION POWER

Retaining paired errors increases conditional certification power (Table 1). At budget 192, ATOM–PAIR returns valid Near claims in 125/512 M1 samples and 124/512 M2 samples; ATOM–MARG returns none. Neither model has a sample in which the marginal method alone returns a valid claim. The simultaneous conditional Monte Carlo lower bounds remain above 19 percentage points. Both point gains exceed the predeclared ten-point criterion.

The gain is budget-dependent. At 64 and 128 problems, both atomic methods have zero valid Fixed–Near returns in both models. The 192-problem budget was chosen before new outputs after a zero-event resolution calculation; the full budget comparison retains both smaller samples. At this budget, paired certification is obtained in approximately one quarter of the samples. Component-level analysis below identifies which requirement determines the remaining unresolved outcomes.

The atomic and betting implementations have different operating characteristics. At the same primary budget, BET–PAIR returns in 27/512 M1 samples and 4/512 M2 samples; both betting-marginal counts are zero. Atomic pairing provides the higher valid-return rate in this comparison. Full budget and selector tables are in Appendix F.

A matched classical paired reference. A post-hoc comparison applies Newcombe's paired-data method 10 to the same 512 samples, retaining the atomic absolute-risk bound and fixed candidate. At 192 problems, its nominal rule returns in 0/512 M1 and 1/512 M2 samples. Scaling its class-interval widths by a heuristic finite-population correction (hFPC) yields 191/512 and 131/512 returns (Table 2). The heuristic adjustment yields more returns than the atomic paired procedure. Table 2 distinguishes the three inferential conditions. Appendix H specifies the construction, all budgets, and fixed-roster coverage checks.

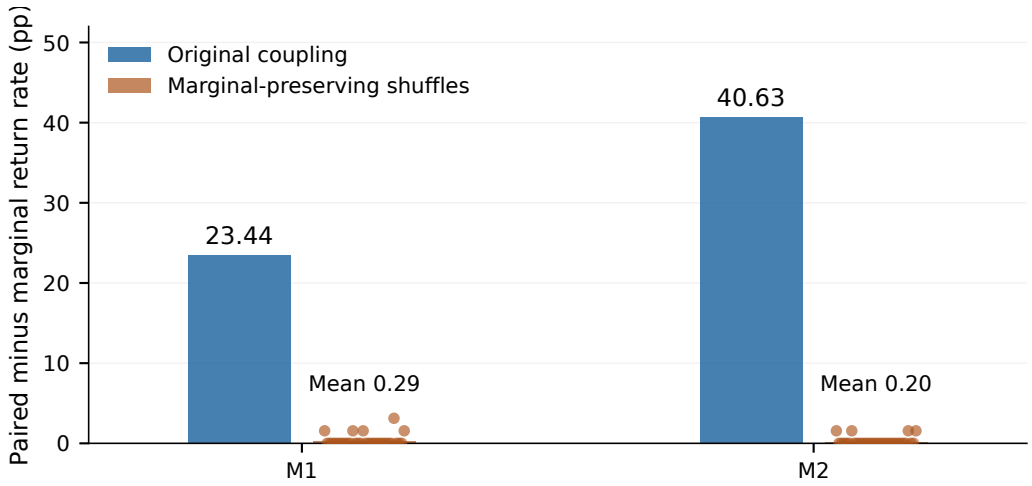


Figure 2: **Preserve marginal errors; nearly remove the paired advantage.** Q192 coupling control for Fixed–Near atomic bounds. Blue bars use the original joint error table; orange marks are all 32 marginal-preserving surrogates and the orange bar is their mean. All share 64 samples per variant, separate from Table 1's 512 samples. Error-count preservation applies to the finite population, not necessarily each sampled subset. The dependent surrogate results are summarized by their descriptive mean and range.

Table 3: Original Q235 paired error counts for CAN_L against each nonself canonical comparator. Class 0 contains first-error endpoints; class 1 contains correct endpoints. L only and Other only are errors made by just one rule. These are census descriptions, not additional audit samples.

Model/class	Comparator	Both wrong	L only	Other only	Both correct	N
M1/0	CAN_F	8	4	1	91	104
M1/0	CAN_E	7	5	0	92	104
M1/1	CAN_F	24	0	7	173	204
M1/1	CAN_E	24	0	7	173	204
M2/0	CAN_F	13	2	2	87	104
M2/0	CAN_E	13	2	3	86	104
M2/1	CAN_F	38	1	3	162	204
M2/1	CAN_E	37	2	0	165	204

5.2 The gain depends on joint errors, not marginal risk

The coupling control changes certification power at fixed marginal risks (Figure 2). In the separate 64-sample control experiment at budget 192, the original paired-minus-marginal gains are 23.44 and 40.63 points. After marginal-preserving permutations, the mean gains across 32 surrogates are 0.29 and 0.20 points. The descriptive surrogate ranges are $[0, 3.13]$ and $[0, 1.56]$ points. This control uses 64 shared masks per variant; Table 1 uses a separate set of 512 masks.

For canonical binary errors, write $h = \Pr(e_j = 1, e_b = 0)$ and $b = \Pr(e_j = 0, e_b = 1)$. Then

$$\mathbb{E}[e_j - e_b] = h - b, \qquad \mathrm{Var}(e_j - e_b) = h + b - (h - b)^2. \tag{4}$$

Marginal risks determine $h - b$ but not $h + b$. The control preserves the first quantity while changing the co-occurrence of errors. Equation (4) gives the standard variance relationship; certificate issuance uses the bounds in Section 3. Together with the controlled results, it explains why a table of individual accuracies omits relevant information about certification difficulty.

The complete error tables make the dependence concrete (Table 3). For M1's erroneous-step class, CAN_L and CAN_E share seven errors, with five L-only and no E-only errors. The paired comparison uses these discordant events, while the marginal risks include the shared errors.

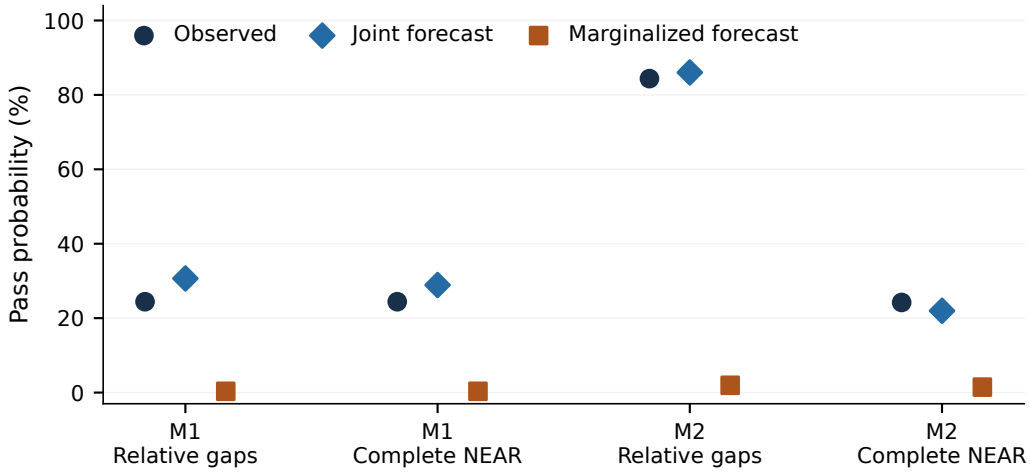


Figure 3: **Joint errors forecast which condition will block the certificate.** Predictions were sealed before any Q model output. Points show pass probabilities for ATOM–PAIR at budget 192; observed probabilities use 512 fixed-roster randomizations. M2's large relative-gap probability does not imply a large complete-Near probability. The primary MAE averages the 12 targets defined in Section 5.3 and listed in Table 8.

5.3 JOINT-ERROR INFORMATION PREDICTS THE NEW CERTIFICATION PATTERN

Joint-error information predicts certification bottlenecks before target evaluation. The fixed Joint forecasts have primary MAE 0.0153, compared with 0.1284 for Marginalized. Per-model MAEs are 0.0181 versus 0.0802 for M1 and 0.0125 versus 0.1766 for M2. The worst primary-cell absolute errors are 0.0625 and 0.8242, respectively. MAE averages the 12 prespecified model–budget–event cells; Table 8 reports each cell. On the four 192-problem cells alone, the corresponding MAEs are 0.0366 and 0.3833. This post hoc slice complements the unchanged primary score.

Figure 3 shows the four 192-problem cells; Appendix D retains all 12 primary cells, including the lower-budget near-zero probabilities. For M2, Joint forecasts a relative-gap pass probability of 0.8604 and complete-Near probability of 0.2197. The observed values are 0.8438 and 0.2422. Thus the forecast anticipates a specific bottleneck: relative closeness can be well supported while absolute quality still prevents a complete certificate. Marginalized forecasts 0.0195 and 0.0146 for these events despite preserving marginal error counts in its pseudo-rosters.

5.4 CERTIFICATE BOTTLENECKS AND CANDIDATE SELECTION

Absolute and relative requirements impose different bottlenecks (Table 4). For M1, atomic absolute-risk bounds pass in every sample, but relative gaps pass only 125 times. For M2, the gap condition passes 432 times while the absolute-risk condition passes 135 times; the complete intersection has 124 returns. The absolute-risk condition determines which precise comparisons receive a complete certificate. The complete component partition (Appendix I) also shows M2's disjoint bottlenecks: 308 samples pass only the paired gap condition, 11 only absolute risk, and 69 neither. The fixed betting implementation passes the two components in 106 and nine samples, with four in their intersection. These joint counts locate the two implementations' distinct operating points.

The one actual prospective sample B illustrates the distinction between power and issuance. It contains 78 erroneous and 170 correct endpoints. For M1's fixed CAN_L, ATOM–PAIR gives $U_R = 15.77\%$ but maximum gap upper bound 4.07 pp; Fixed–Near remains unresolved. Bonf–Near with atomic pairing instead certifies CAN_E, with $U_R = 15.81\%$ and maximum gap bound 2.70 pp, after testing all seven candidates at corrected levels. This is the only Near program to issue in that draw. M2 has no certificate: its fixed atomic CAN_L has a sufficiently small gap bound, 2.43 pp, but $U_R = 22.10\%$. The subsequent census gives true Q risk 16.77%, distinguishing an unresolved certificate from a violated requirement. All B bounds and returned candidates are reported in Appendix E.

Table 4: **A precise comparison does not certify absolute quality.** Counts out of 512 Q192 samples for the fixed CAN_L. Its disagreement is structurally zero. A complete return jointly passes the absolute-risk condition and every relative-gap condition.

Model	Engine	Absolute R	All relative gaps	Complete Near
M1	Atom–marginal	512/512	0/512	0/512
M1	Atom–paired	512/512	125/512	125/512
M1	Bet–marginal	509/512	0/512	0/512
M1	Bet–paired	509/512	27/512	27/512
M2	Atom–marginal	135/512	0/512	0/512
M2	Atom–paired	135/512	432/512	124/512
M2	Bet–marginal	9/512	0/512	0/512
M2	Bet–paired	9/512	106/512	4/512

Candidate selection trades access to an easier-to-certify rule against multiplicity cost. On Q192, atomic Bonf–Near returns valid claims in 333/512 M1 samples versus Fixed's 125/512, mainly choosing CAN_E; on M2 the corresponding counts are 17/512 versus 124/512. Selection can find an easier-to-certify alternative, but its multiplicity cost can also dominate. These secondary comparisons retain the fixed-candidate primary contrast. The existing target sensitivity also distinguishes the checkpoints: at a five-point gap tolerance, atomic pairing returns in 510/512 M1 samples versus nine for marginal subtraction, and 135/512 M2 samples versus zero. M2's unchanged absolute-risk requirement caps those returns (Appendix G).

The census distinguishes actual threshold violations from inconclusive certification. On Q, M1's I and $Q_{.02}$ have $D = 4.120\% > 2\%$. M2's Q_0 has $R = 20.433\%$ and $G = 3.672\,\text{pp}$, violating both thresholds. In contrast, both fixed CAN_L protocols satisfy the full Q conjunction. Every previously issued claim is true on its original completed target.

6 Implications and scope

Audit information and decision requirements. Paired errors and component-level requirements make certification outcomes interpretable. The matched comparisons, coupling control, and prospective forecasts identify information discarded by marginal accuracy. The classical paired reference further shows that the treatment of finite-population sampling affects the return rate even when pairing is retained. Information choice and interval calibration are therefore distinct parts of audit design.

Scope of statistical guarantees. The established error control applies to each specified program, target roster, budget, and round. It is not simultaneous coverage of every reported bound or authorization to choose the best program after observing outcomes. The Newcombe reference is nominal; its hFPC width adjustment has no finite-population coverage proof here. Both fixed candidates satisfy the finite-roster targets, so the observed zero false returns cannot evaluate error control under a violated conjunction. Component partitions locate operating bottlenecks but do not isolate a unique cause in the betting mixture.

Experimental scope. Q235 is the unscored remainder of one filtered ProcessBench population. The two checkpoints share a backbone family, and problem resampling describes conditional audit variation. The coupling and forecasting studies operate on joint error tables rather than interventions on neural computations. Their correlated forecast cells and surrogate tables are not independent task replications. The guarantee concerns the fixed roster, not a new deployment population.

Historical observations also have a cost. Retrospective P1099 and known-fraction analyses keep known, incremental, and total problem counts separate (Appendix G). Newly scoring an endpoint requires three F/E/L forwards per checkpoint in this instrument, compared with one for a deployed canonical or received-view rule. Benchmark labels already exist. The problem budget measures observed model evaluations; annotation and deployment costs require separate accounts.

7 CONCLUSION

Joint errors carry information that changes a verification audit's conclusions. At fixed models, samples, and targets, pairing increases valid certificate returns; preserving marginal errors while changing their coupling nearly removes the advantage. Prospective joint-error forecasts anticipate which requirement will constrain certification. These results connect protocol quality, comparison information, and the adequacy of statistical evidence, providing a concrete basis for auditing PRM alternatives.

REPRODUCIBILITY STATEMENT

Appendices A–C specify the data, frozen rules, certificate mathematics, and information chronology. The numerical package supplies P/Q decisions and losses, all 512 original sample permutations, saved fixed-candidate bounds, and CPU reconstruction code. It rebuilds the reported truths, component and sensitivity counts, and classical paired comparisons; independent high-precision calculations check interval endpoints and decisions. Saved forecast and coupling summaries support independent arithmetic reconstruction of their reported statistics. Appendix J specifies this layer and the configurations needed for fresh model inference.

AI USE STATEMENT

AI tools assisted scientific planning, implementation, statistical analysis, interpretation, figure preparation, and manuscript drafting. Codex assisted the local experimental workflow, and ChatGPT assisted evidence review and reconstruction checks. Numerical figures use recorded experimental values; workflow diagrams are editable vector drawings checked against the study design. The authors take responsibility for the claims, references, and submitted materials.

ETHICS STATEMENT

This study uses public mathematical reasoning data and involves no newly recruited participants or collection of personal records. The certificate describes benchmark performance; deployment safety requires its own evaluation. Models and data retain their original licenses. Model weights are obtained from the original providers.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control, 2022. URL https://arxiv.org/abs/2110.01052v5.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):1–34, 2021. doi: 10.1145/3478535. URL https://doi.org/10.1145/3478535.

Brian M. Cho, Kyra Gan, and Nathan Kallus. Peeking with PEAK: Sequential, nonparametric composite hypothesis tests for means of multiple data streams. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 8487–8509. PMLR, 2024. URL https://proceedings.mlr.press/v235/cho24a.html.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404. URL https://doi.org/10.1093/biomet/26.4.404.

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, 2021. URL https://arxiv.org/abs/2110.14168.

William G. Cochran. *Sampling Techniques*. John Wiley & Sons, New York, third edition, 1977. ISBN 9780471162407. URL https://www.wiley-vch.de/en/areas-interest/mathematics-statistics/sampling-techniques-978-0-471-16240-7.

Thomas G. Dietterich. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation*, 10(7):1895–1923, 1998. doi: 10.1162/089976698300017197. URL https://doi.org/10.1162/089976698300017197.

Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Runxin Xu, Zhengyang Tang, Benyou Wang, Daoguang Zan, Shanghaoran Quan, Ge Zhang, Lei Sha, Yichang Zhang, Xuancheng Ren, Tianyu Liu, and Baobao Chang. Omni-MATH: A Universal Olympiad Level Mathematic Benchmark For Large Language Models, 2024. URL https://arxiv.org/abs/2410.07985.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. OlympiadBench: A Challenging Benchmark for Promoting AGI with Olympiad-Level Bilingual Multimodal Scientific Problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3828–3850. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.211. URL https://aclanthology.org/2024.acl-long.211/.

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, 2021. URL https://arxiv.org/abs/2103.03874.

Wassily Hoeffding. Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830. URL https://doi.org/10.1080/01621459.1963.10500830.

Sture Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

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. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/aca97732e30bcf1303bc22ac3924fd16-Abstract-Conference.html.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization, 2009. URL https://arxiv.org/abs/0907.3740.

Quinn McNemar. Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages. *Psychometrika*, 12(2):153–157, 1947. doi: 10.1007/BF02295996. URL https://doi.org/10.1007/BF02295996.

Robert G. Newcombe. Improved confidence intervals for the difference between binomial proportions based on paired data. *Statistics in Medicine*, 17(22):2635–2650, 1998. doi: 10.1002/(SICI)1097-0258(19981130)17:22<2635::AID-SIM954>3.0.CO;2-C.

Joseph P. Romano and Michael Wolf. Stepwise Multiple Testing as Formalized Data Snooping. *Econometrica*, 73(4):1237–1282, 2005. doi: 10.1111/j.1468-0262.2005.00615.x. URL https://doi.org/10.1111/j.1468-0262.2005.00615.x.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM Test-Time Compute Optimally Can be More Effective than Scaling Parameters for Reasoning. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1b623663fd9b874366f3ce019fdfdd44-Abstract-Conference.html.

Mingyang Song, Zhaochen Su, Xiaoye Qu, Jiawei Zhou, and Yu Cheng. PRMBench: A fine-grained and challenging benchmark for process-level reward models, 2025. URL https://arxiv.org/abs/2501.03124v5.

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, 2022. URL https://arxiv.org/abs/2211.14275.

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. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.510. URL https://aclanthology.org/2024.acl-long.510/.

Ian Waudby-Smith and Aaditya Ramdas. Confidence sequences for sampling without replacement. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/e96c7de8f6390b1e6c71556e4e0a4959-Abstract.html.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, 2024. doi: 10.1093/jrsssb/qkad009. URL https://doi.org/10.1093/jrsssb/qkad009.

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, 2024. URL https://arxiv.org/abs/2409.12122.

Lifan Yuan, Wendi Li, Huayu Chen, Ganqu Cui, Ning Ding, Kaiyan Zhang, Bowen Zhou, Zhiyuan Liu, and Hao Peng. Free Process Rewards without Process Labels. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 73511–73525. PMLR, 2025. URL https://proceedings.mlr.press/v267/yuan25c.html.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. The lessons of developing process reward models in mathematical reasoning, 2025. URL https://arxiv.org/abs/2501.07301v2.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. ProcessBench: Identifying process errors in mathematical reasoning, 2025. URL https://arxiv.org/abs/2412.06559v4.

Congmin Zheng, Jiachen Zhu, Jianghao Lin, Xinyi Dai, Weiwen Liu, Haoxuan Li, Yong Yu, Weinan Zhang, and Mengyue Yang. CoLD: Counterfactually-guided length debiasing for process reward models in mathematical reasoning, 2026. URL https://arxiv.org/abs/2507.15698v2.

A Data, native measurement, and frozen protocols

A.1 Source versions and endpoint selection

The dataset is `Qwen/ProcessBench` at revision

`3bdcd5371ed567559a78f559c01c13a6deee7604`.

Its four files draw on GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021), OlympiadBench (He et al., 2024), and Omni-MATH (Gao et al., 2024). ProcessBench supplies the solution-level first-error annotations used here (Zheng et al., 2025). The original benchmark contains solution records that can share a problem. We cluster by exact problem text after Unicode NFC normalization, CRLF-to-LF conversion, and removal of outer whitespace only. Internal whitespace and case are retained. Cross-file duplicates are merged; a fixed source ordering assigns a canonical domain while retaining provenance. A conflict in the earliest-error label for identical problem and step sequences excludes the whole cluster.

The initial construction considers current-step indices $i \geq 3$ in zero-based indexing. If the earliest error is $e \geq 0$, only $i < e$ is labeled correct and $i = e$ erroneous; no endpoint after e is eligible. When the source marks an entirely correct solution, its eligible endpoints are correct. All content after the selected endpoint is omitted. The construction excludes malformed labels or steps, empty steps, embedded special-token literals, conservative step-index references, and any input exceeding the 4,096-token native limit in either checkpoint. Selected inputs are evaluated without truncation or outcome-dependent replacement.

Index-reference exclusions cover explicit step/stage/paragraph numbers, including Roman numerals, and finite positional terms such as previous, next, first, second, and last followed by those nouns. This syntactic filter determines eligibility before model evaluation. Eligible endpoints are chosen by a deterministic hash of the source-row identity and current-step index, retaining at most one per class per problem. Eligibility and selection precede the original model outcomes. The historical split uses fixed hashes within source-domain and class-presence strata: calibration, discovery, and the original audit roster are disjoint. Calibration has 247 clusters; discovery has 331 (120 early-screen and 211 later-discovery clusters); the original audit roster has 1,099. The final Q235 is the still-unscored remainder of that roster.

Table 5: Problem and endpoint accounting. A problem can supply both classes, so class counts need not sum to the problem count. A864 is development information in the final study, not another new sample.

Role	Problems	Error endpoints	Correct endpoints
Calibration	247	118	212
Original roster P	1,099	531	938
Previously observed A	864	427	734
Final roster Q	235	104	204
Prospective B within Q	192	78	170
Census remainder H within Q	43	26	34

Table 6: Source-domain support in the complete rosters. These are descriptive strata, not independent model-family replications or per-domain certificate guarantees.

	Q235			P1099		
Domain	Problems	Error	Correct	Problems	Error	Correct
GSM8K	39	6	38	146	38	128
MATH	74	21	66	349	131	294
OlympiadBench	56	38	48	269	164	238
OmniMath	66	39	52	335	198	278

A.2 Models and inference configuration

M1 is `Qwen/Qwen2.5-Math-PRM-7B`, revision

`0610740060112df12585d00a1c5f4624d2f59051`.

M2 is `Qwen/Qwen2.5-Math-7B-PRM800K`, revision

`9d6e292f6ccfd474fa44461ce6d5b80d08d8f3c7`.

The two checkpoints share the Qwen2.5-Math backbone family (Yang et al., 2024) but differ in supervision recipe; their PRM provenance is documented in Zhang et al. (2025); Zheng et al. (2025). The comparison therefore spans two supervision recipes within a common backbone family.

The instrument uses PyTorch 2.7.1+cu128, Transformers 4.40.2, BF16 weights, eager attention, batch size one, evaluation/inference mode, and KV caching disabled on an RTX 5090. Each evaluation applies the fixed checkpoint as a classifier. The system string is:

```
Please reason step by step, and put your final answer within
\boxed{}.
```

The user field is the unmodified problem. The assistant field contains the selected prefix, with two LF characters appended to each history step, and the frozen step marker <extra_0>. The native chat template is applied to the complete string before tokenization. The classifier produces two logits at each position; we use the positive-class probability at the final inserted step marker, as the current-step score. The audit retains the two logits and computes probabilities in float64 with higher-precision decision checks where needed. These checks verify decisions computed from the recorded BF16 logits; inference precision remains part of the specified measurement configuration.

A.3 THREE ENCODINGS AND SEVEN RULES

With h history steps, F marks each original step separately. E has history blocks $[0, k)$ and $[k, h)$ for $k = \lfloor h/3 \rfloor$; L uses $k = \lceil 2h/3 \rceil$. The current step remains a separate final block. E and L each have three markers and must match in non-marker tokens, current-step bytes/tokens, total token count, marker count, and current-marker absolute index. The only intended change is one history marker's position. F preserves the current logical step with its original marker count and input length.

A threshold is fitted exclusively on the original calibration set. For one representation, sort its distinct probabilities $v_0 < \cdots < v_{K-1}$ and enumerate all $K + 1$ monotone acceptance sets, including accepting and rejecting everything. Interior thresholds are adjacent probability midpoints; decisions use $p \geq \tau$. Minimize calibration balanced error, resolving exact ties by the largest cut index (the more restrictive acceptance set).

I uses the independently fitted E and L thresholds. S fits one threshold on the pooled E/L probability grid, minimizing the mean of the two calibration errors with the same cut-index tie rule. For each $\epsilon \in \{0, .02\}$, Q_ϵ enumerates the Cartesian product of E and L cut sets, minimizes mean calibration error subject to $D_C \leq \epsilon$, then breaks ties by smaller D_C, larger E cut, and larger L cut. The constraint is fitted on calibration data and independently evaluated on each audit roster.

Each CAN_b reconstructs fixed encoding $b \in \{F, E, L\}$ from the original steps and uses that representation's independent threshold. Its two input-view decisions therefore agree by construction. The full family is $\{\text{CAN}_F, \text{CAN}_E, \text{CAN}_L, \text{I}, \text{S}, Q_0, Q_{.02}\}$. Training error, training disagreement, and this ID order determine the frozen Seq order. Both checkpoints' best calibration CAN is CAN_L, the Fixed candidate. All thresholds, candidates, and ordering are retained when evaluating A, B, H, and the complete Q/P truths. Multiplicity correction uses the full seven-candidate family.

An illustrative encoding example. For six history steps, F retains the blocks $[0, 1), \ldots, [5, 6)$; E uses $[0, 2), [2, 6)$; and L uses $[0, 4), [4, 6)$. The current step remains a separate final block in all three cases. F therefore has a different marker count and length from E and L. CAN_L reconstructs L from the original steps for either received view, then applies the same L score and frozen threshold. Its two decisions agree by construction, giving $D = 0$. The example specifies input boundaries; canonical agreement follows from using the same reconstructed input and threshold.

Table 7: Frozen verification protocols. Each threshold is fitted on the separate calibration set. Canonical rules reconstruct the same encoding under either received view.

Rule	Encoding and threshold	Calibration constraint
CAN_F	Canonical F; F threshold	None
CAN_E	Canonical E; E threshold	None
CAN_L	Canonical L; L threshold	None; fixed candidate
I	Received E/L; separate thresholds	None
S	Received E/L; shared threshold	None
Q_0	Received E/L; paired thresholds	Disagreement ≤ 0
$Q_{.02}$	Received E/L; paired thresholds	Disagreement $\leq .02$

B CERTIFICATE MATHEMATICS AND NUMERICAL IMPLEMENTATION

B.1 WHY SAMPLING PROBLEMS PERMITS CLASS-SPECIFIC EVENT BOUNDS

Fix a finite roster and all candidate decisions on it. Draw a uniform subset of complete problem clusters, without replacement. Each class contributes at most one endpoint per cluster. Conditional on the number of selected endpoints of class y, their identities form a simple random sample without replacement from that class's endpoints. For sequential values, the induced class order is uniform as well. The two class samples can be dependent, and both views and checkpoints can be arbitrarily dependent. The scalar combination rules accommodate this dependence.

For a class-specific binary event with population size N, observed count k in n sampled endpoints, the feasible population count is $K \in [k, N-n+k]$. Write $X \sim \text{Hypergeom}(N, K, n)$. The one-sided count bounds at error level γ are

$$K_U = \max\{K : \Pr_K(X \le k) \ge \gamma\}, \tag{5}$$

$$K_L = \min\{K : \Pr_K(X \ge k) \ge \gamma\}. \tag{6}$$

Tail equality is retained. Empty samples give support bounds; a census gives the exact count. Probabilities are K_U/N and K_L/N. Integer-tail inversion is a standard exact finite-population operation within the sampling framework described by Cochran (1977). Independent exact tail calculations verify each returned integer and its adjacent boundary.

For an atomic scalar expression, combine positive coefficients with upper event bounds and negative coefficients with lower bounds. Divide that scalar's assigned level q equally over its nonconstant, nonduplicate one-sided events. For a canonical candidate, R has two class-error atoms; a noncanonical view-averaged candidate can have four. D has two class-disagreement atoms unless it is structurally zero. A paired difference uses harmful and beneficial events separately in each class/view. Structural equalities determine cancellation and merging before allocation, with the same terms retained for every sample.

B.2 KNOWN OBSERVATIONS AND SAMPLE SPLITTING

When a target $\mathcal{T} = A_0 \dot\cup U$ contains an already known set, let $S_{A_0,y}$ be its exact class loss sum. The whole-target class mean is

$$\mu_{\mathcal{T},y} = \frac{S_{A_0,y} + N_{U,y}\mu_{U,y}}{N_{\mathcal{T},y}}. \tag{7}$$

A bound on $\mu_{U,y}$ passes through this increasing affine transformation. The known term has no new sampling uncertainty. In the final Q certificate, A_0 is empty except for Split's choose half. Conditional on that half and its chosen candidate, the test half samples from the remainder; both the remaining population count and the known sum are adjusted before recovering the same original Q target. This adjustment preserves the original Q target while conditioning on selection.

Earlier P certificates used their historical known sets and remain statements about P. The final Q certificate uses a separate target, with the 864 historical problems reserved for development and forecasting. Known-fraction replays deliberately vary A_0 under Equation (7) and retain the associated observation cost.

B.3 THE FIXED BETTING ALTERNATIVE

The betting component follows the bounded-mean, without-replacement framework of Waudby-Smith & Ramdas (2024; 2020). For a population of N values in $[0,1]$, a null mean m, and observations $X_1, \dots, X_t$, let

$$r_t(m) = \frac{Nm - \sum_{i<t} X_i}{N-t+1}, \qquad f_{t,c}(m) = 1 + c\frac{r_t(m) - X_t}{1 - r_t(m)}. \tag{8}$$

We fix $c \in \{1/20, \dots, 19/20\}$ and form

$$K_{t,c}(m) = \prod_{i=1}^{t} f_{i,c}(m), \qquad E_t(m) = \frac{1}{19}\sum_c K_{t,c}(m). \tag{9}$$

For $r_t < 1$, factors are nonnegative because their minimum over $X_t \le 1$ is $1-c$. Under the composite null $\mu \ge m$, the true conditional remaining mean is at least $r_t(m)$, so $\mathbb{E}[f_{t,c} \mid \mathcal{F}_{t-1}] \le 1$. Products and their fixed mixture are therefore nonnegative test supermartingales. When $r_t = 1$ and $X_t = 1$ the factor is one; $r_t > 1$, or $r_t = 1$ with $X_t < 1$, makes the null deterministically incompatible with the observed support. Rejection uses the fixed mixture criterion $E_n(m) > 1/\gamma$. The construction applies the cited test-supermartingale results.

Values lie on a known lattice: view-average losses have step $1/2$, canonical errors step one, and transformed signed gaps step $1/4$. If the step is $1/d$ and observed unit sum is S, feasible population units are $T \in [S, S + d(N-n)]$, with $m = T/(dN)$. Invert at the largest nonrejected T. The wealth is nondecreasing in m, permitting bisection. A numerical upper bound must either reach the deterministic feasible maximum or certify rejection at its next larger lattice point. Interval arithmetic, followed when necessary by exact rational evaluation, checks strict comparisons. Unresolved numerical comparisons widen the bound.

BET–MARG applies this procedure directly to each class's ℓ_j, with lower bounds obtained from $1-\ell_j$, and subtracts marginal risk bounds. BET–PAIR instead transforms $g_{jb} = \ell_j - \ell_b$ to $z = (g_{jb}+1)/2$, bounds each class's z, and maps back by $2U_z - 1$. Error allocations are split across classes with unobserved outcomes. Both betting engines retain hypergeometric D bounds. The complete programs use fixed-time evaluation because the conjunction includes fixed-time hypergeometric components.

B.4 SELECTION AND CONJUNCTION VALIDITY

Proposition 1 (Guarantee for a declared program)**.** *Fix its target roster, candidate family, calibration ordering, and sample budget. Suppose each scalar test of a candidate's risk threshold has false-passage probability at most q under that scalar's violation. Then the declared Fixed, Seq, Bonf, and Split programs control false certification for one model by $\alpha = 1/40$, and for either of the two models by $1/20$.*

Proof. If a candidate violates Abs or Near, at least one of its fixed scalar constraints is false. Passing every constraint is a subset of falsely passing that scalar. Thus every scalar can use the same q for this intersection–union test; no independence is assumed. Fixed tests one precommitted candidate at $q = \alpha$. Bonf uses $q = \alpha/7$, so a union bound protects the passed set. Seq uses $q = \alpha$ but stops at the first non-pass. A false certification requires passage of the first truly invalid candidate in the frozen order, an event of probability at most α. Split fixes one candidate conditional on choose, then applies its valid remaining-population test at α. Finally, a union bound over the two models gives $2\alpha = 1/20$. □

These are standard risk-control arguments (Angelopoulos et al., 2022). They provide the declared 5% two-model false-certification guarantee separately for each program, target, and budget. Seq evaluates and may return candidates only through its first non-pass; any later displayed bounds are descriptive.

B.5 INFORMATION IDENTITIES AND DETERMINISTIC EXCLUSIONS

For a finite one-class population with variance defined using denominator N, simple random sampling without replacement gives

$$\mathrm{Var}(\overline{X}_n) = \frac{N-n}{n(N-1)}\,\mathrm{Var}_{\mathrm{pop}}(X), \qquad N > 1. \tag{10}$$

This exposes the role of the sampling fraction. Combining classes additionally requires their covariance induced by shared problems. Certificate decisions use the risk bounds, while this identity explains the role of the sampling fraction.

The conjunction passage probability is at most the passage probability of each component. Separately, known nonnegative R or D contributions divided by their *full-target* denominators are deterministic lower bounds. A lower bound above the target establishes deterministic infeasibility. In the earlier P study, already observed losses established $D \ge 32531/996156 = 3.266\%$ for M1 I and $Q_{.02}$. This arithmetic exclusion is different from a statistical upper bound that remains above 2%, which only leaves the claim unresolved.

C FROZEN ANALYSIS DESIGN AND INFORMATION BOUNDARIES

The final study retained the 20% absolute-risk, 2% disagreement, and 3-point comparison targets. It chose the 192-problem primary budget before Q outcomes after a program-resolution calculation. Using illustrative rounded class counts $(57, 111)$ for 128 problems, a zero-harm/zero-benefit canonical paired comparison had upper gap $77/1768 = 4.355\,\text{pp}$. Counts $(85, 167)$ for 192 gave $77/5304 = 1.452\,\text{pp}$. These zero-event calculations characterize the resolution of the specified inference program. The actual B class counts were $(78, 170)$.

The realized permutation of Q was generated once and fixed before evaluation, without redrawing. B inference preceded its certificate seal; H inference followed it. The final analysis program and forecasts were sealed before any Q inference. Q validation uses 512 independent uniform problem permutations and shared prefixes at $n \in \{64, 128, 192\}$. Each permutation defines a conditional design replicate using the sampled losses alone. Complete roster truth is supplied only to a separate scoring step. Each reported budget is evaluated at its fixed sample size.

For the primary contrast, define Z_r as the difference of valid-return indicators for ATOM–PAIR and ATOM–MARG, so $Z_r \in \{-1, 0, 1\}$. Let $p_+ = \Pr(Z = 1)$ and $p_- = \Pr(Z = -1)$ under the fixed-roster sampling design. A one-sided Clopper–Pearson lower bound L_+ and upper bound U_-, each at error level $1/80$, give $L_+ - U_-$ as a lower bound for $p_+ - p_-$. Four tails across two models yield joint error at most $4/80 = 0.05$. The union bound allows dependence between the plus/minus events and models. Continuous binomial roots must be rounded outward, unlike a discrete hypergeometric integer inversion. These Monte Carlo bounds quantify conditional audit-power uncertainty.

C.1 FORECAST CONSTRUCTION AND COUPLING CONTROLS

Static Q metadata specify source domain and whether a cluster contains error, correct, or both endpoint classes. Each forecast pseudo-roster draws 235 donor clusters from A with replacement in those strata. A donor is a whole block: both models, every present class, and all three CAN error vectors. When a domain/presence stratum is unsupported, the rule allows same-presence donors pooled over domains; the actual forecast required zero such fallbacks. The frozen design used 64 pseudo-rosters and 32 audit permutations per pseudo-roster, with three fixed random seeds recorded in the companion configuration. The resulting 2,048 simulated audits per cell are nested within the 64 pseudo-rosters.

Joint retains the donor blocks. Marginalized shuffles each F or E comparator's whole-cluster block independently within the same source-domain/class-presence strata, leaving L unchanged. This preserves each model's class- and domain-specific marginal error counts. It preserves pseudo-roster canonical risk differences while changing covariance with L. Singleton strata are unchanged. Pseudo-roster construction, sampling, and variants share frozen randomization where prescribed.

The observed-roster coupling control uses the same kind of permutation on A and Q, with 32 surrogate tables and 64 shared audit samples for each budget. Marginal counts are preserved at the roster level, while individual sampled subsets can have different counts after shuffling. The surrogate tables test the information carried by the original statistical dependence. The companion data include the cell-level summaries and complete error tables used for the reported comparisons.

D ALL PRIMARY FORECAST CELLS

Table 8 gives every cell entering the primary MAE. The score includes every lower-budget near-zero value. Absolute-risk forecasts were also fixed prospectively and are retained separately from this 12-cell score. The MAE summarizes these 12 prespecified probability targets; other component forecasts are reported separately.

For the four displayed 192-problem cells, the forecast identifies an asymmetry: M1 is predominantly comparison-limited, whereas M2 has much higher relative-gap passage than complete-Near passage. The primary MAE advantage is $0.128418 - 0.015299 = 0.113118$, exceeding the frozen 0.05 descriptive criterion. That threshold and the 32-surrogate effect criterion define the descriptive analysis.

Table 8: All 12 primary forecast cells for atomic pairing. Percentages are simulated/observed pass probabilities, not risk estimates. Joint and shuffled forecasts are sealed predictions; observed probabilities use the completed Q roster and 512 random samples.

Model	Budget	Event	Joint (%)	Shuffled (%)	Observed (%)
M1	64	Relative gaps	0.00	0.00	0.00
M1	64	Complete Near	0.00	0.00	0.00
M1	128	Relative gaps	0.05	0.00	0.00
M1	128	Complete Near	0.05	0.00	0.00
M1	192	Relative gaps	30.66	0.34	24.41
M1	192	Complete Near	28.91	0.34	24.41
M2	64	Relative gaps	0.00	0.00	0.00
M2	64	Complete Near	0.00	0.00	0.00
M2	128	Relative gaps	4.00	0.00	0.78
M2	128	Complete Near	0.39	0.00	0.00
M2	192	Relative gaps	86.04	1.95	84.38
M2	192	Complete Near	21.97	1.46	24.22

E PROSPECTIVE B DECISIONS AND COMPLETED-ROSTER TRUTH

Table 9: The single prospective B draw, fixed CAN_L, Near. Every row is unresolved. All D bounds are structurally zero. These are marginal scalar bounds used for threshold testing, not a joint confidence box. G_F, G_E denote the corresponding canonical comparisons.

Model	Engine	U_R (%)	$\max_b U_{G_b}$ (pp)	Blocking condition
M1	Atom–marginal	15.77	7.23	G_F,G_E
M1	Atom–paired	15.77	4.07	G_E
M1	Bet–marginal	17.47	9.17	G_F,G_E
M1	Bet–paired	17.47	4.80	G_E
M2	Atom–marginal	22.10	7.52	R,G_F,G_E
M2	Atom–paired	22.10	2.43	R
M2	Bet–marginal	24.04	9.71	R,G_F,G_E
M2	Bet–paired	24.04	2.65	R

M1's only issued Near claim comes from Bonf–ATOM–PAIR, returning CAN_E, with $U_R = 15.81\%$ and largest gap bound $2.70\,\text{pp}$. It uses candidate level $1/280$ rather than the fixed-candidate $1/40$. All 16 M1 Abs programs issue a claim, giving 17 issued program-level claims in total on B. M2 issues none. All 17 results are correlated decisions on the same sample. Empirical choices remain explicitly labeled as having no guarantee.

The completed truths are given in Tables 10–11. The fully observed finite-roster means provide exact evaluation targets after certification. Both fixed CAN_L rules satisfy the Q conjunction, even though neither obtains Fixed–Near certification from B. Conversely, some noncanonical candidates actually violate the targets. The census thus separates true but uncertified candidates from actual target violations.

Table 10: Complete Q235 truth. R and D are percentages; G is a percentage-point difference from the best of the three CAN rules. Values are rounded only for display.

	M1			M2		
Protocol	R (%)	D (%)	G (pp)	R (%)	D (%)	G (pp)
CAN_F	11.925	0.000	0.962	17.261	0.000	0.500
CAN_E	10.963	0.000	0.000	16.761	0.000	0.000
CAN_L	11.652	0.000	0.688	16.770	0.000	0.009
I	11.308	4.120	0.344	16.766	2.894	0.005
S	11.897	0.490	0.933	17.119	1.207	0.358
Q_0	11.312	1.207	0.349	20.433	0.726	3.672
$Q_{.02}$	11.308	4.120	0.344	16.879	1.687	0.118

Table 11: Complete original P1099 truth. This is a different target from Q235. Earlier P claims are checked against this table, not against Q or an H-only estimate.

	M1			M2		
Protocol	R (%)	D (%)	G (pp)	R (%)	D (%)	G (pp)
$\textsc{Can}_F$	14.038	0.000	0.227	18.672	0.000	0.618
$\textsc{Can}_E$	13.811	0.000	0.000	18.638	0.000	0.584
$\textsc{Can}_L$	13.873	0.000	0.062	18.053	0.000	0.000
I	13.842	4.646	0.031	18.346	2.347	0.292
S	13.851	0.709	0.040	18.660	1.192	0.607
Q_0	13.926	1.233	0.115	20.387	1.381	2.333
$Q_{.02}$	13.842	4.646	0.031	18.599	1.045	0.545

The earlier first certification study issued no valid-program certificate. The second issued 21 M1 program-level claims, all on P; each is true on the completed P roster. The final B claims are true on Q. This comparison checks the truth of the realized correlated claims against their original completed rosters.

F Budget and selector results on Q

Table 12: All four engines and four selectors at the Q192 primary budget. Entries are valid returns out of 512. Abs and Near are separately defined target conjunctions. No cross-program confidence guarantee is asserted.

		M1		M2	
Selector	Engine	Abs	Near	Abs	Near
Fixed	Atom–marginal	512/512	0/512	135/512	0/512
Fixed	Atom–paired	512/512	125/512	135/512	124/512
Fixed	Bet–marginal	509/512	0/512	9/512	0/512
Fixed	Bet–paired	509/512	27/512	9/512	4/512
Seq	Atom–marginal	512/512	0/512	135/512	0/512
Seq	Atom–paired	512/512	125/512	135/512	124/512
Seq	Bet–marginal	509/512	0/512	9/512	0/512
Seq	Bet–paired	509/512	27/512	9/512	4/512
Bonf	Atom–marginal	512/512	0/512	32/512	0/512
Bonf	Atom–paired	512/512	333/512	32/512	17/512
Bonf	Bet–marginal	510/512	0/512	2/512	0/512
Bonf	Bet–paired	510/512	65/512	2/512	0/512
Split	Atom–marginal	442/512	0/512	70/512	0/512
Split	Atom–paired	442/512	162/512	70/512	38/512
Split	Bet–marginal	434/512	0/512	4/512	0/512
Split	Bet–paired	434/512	40/512	4/512	1/512

Selection has different effects across the checkpoints. Among the 333 M1 atomic Bonf–Near returns at Q192, 306 select $\textsc{Can}_E$, 18 select $\textsc{Can}_L$, and nine select $\textsc{Can}_F$. M2's Bonf count is 17, versus 124 for Fixed. These comparisons separate an opportunity to choose an easier-to-certify rule from the penalty for certifying a larger candidate family. Their relative effects determine whether candidate search improves the audit.

For each simulation the scoring record retains whether an issued claim is actually valid, as well as separate R, D, and G violations. The nominal error level controls unconditional false certification; errors conditional on return are a separate quantity. The companion data retain the validity and violation summaries, together with the paired valid-return indicators used in the primary contrast.

G Retrospective P budgets, known information, and sensitivity

After P is fully observed, 256 frozen random permutations evaluate budgets $64, 128, 256, 512, 768, 1024$. This supplies exact scoring truth for retrospective finite-population

Table 13: Fixed-candidate Near results at every predeclared Q budget. Zero valid returns at 64 and 128 are retained. A zero in 512 simulations is not an exact zero conditional power.

		M1		M2	
Budget	Engine	Abs	Near	Abs	Near
64	Atom–marginal	39/512	0/512	0/512	0/512
64	Atom–paired	39/512	0/512	0/512	0/512
64	Bet–marginal	1/512	0/512	0/512	0/512
64	Bet–paired	1/512	0/512	0/512	0/512
128	Atom–marginal	386/512	0/512	16/512	0/512
128	Atom–paired	386/512	0/512	16/512	0/512
128	Bet–marginal	95/512	0/512	0/512	0/512
128	Bet–paired	95/512	0/512	0/512	0/512
192	Atom–marginal	512/512	0/512	135/512	0/512
192	Atom–paired	512/512	125/512	135/512	124/512
192	Bet–marginal	509/512	0/512	9/512	0/512
192	Bet–paired	509/512	27/512	9/512	4/512

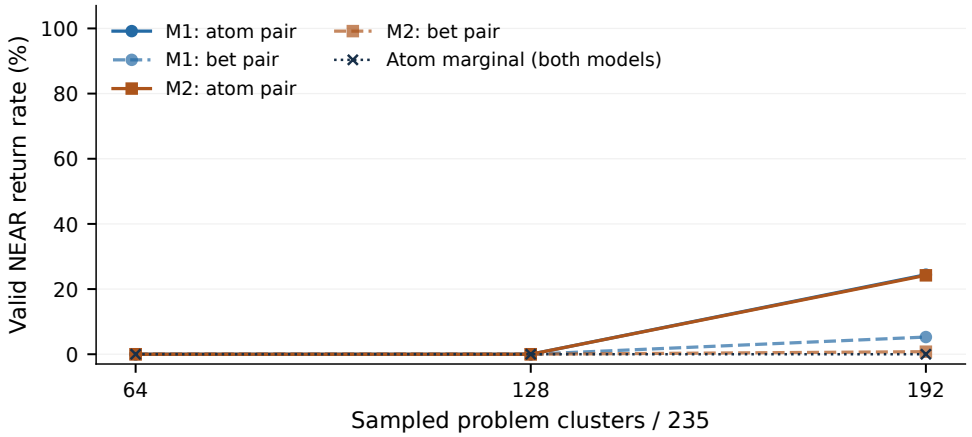


Figure 4: Conditional Fixed–Near valid-return rates on Q. Lines connect the three predeclared fixed-budget evaluations.

power curves. The replay conditions on the completed original roster. The replay includes all atomic engines/selectors/goals, fixed-candidate betting controls, and empirical selectors; we report the same fixed canonical comparison.

Table 14: Retrospective Fixed–Near valid returns on P1099, out of 256 per cell. Larger finite-population coverage changes information as well as the absolute number of problems.

	M1			M2		
Budget	Atom pair	Atom marg.	Bet pair	Atom pair	Atom marg.	Bet pair
64	0/256	0/256	0/256	0/256	0/256	0/256
128	0/256	0/256	0/256	0/256	0/256	0/256
256	1/256	0/256	0/256	0/256	0/256	0/256
512	159/256	0/256	89/256	11/256	0/256	0/256
768	256/256	0/256	256/256	74/256	0/256	2/256
1024	256/256	256/256	256/256	256/256	256/256	165/256

For M1, atomic pairing first exceeds 80% valid return at the tested budget 768, while marginal subtraction does so at 1,024. For M2, both first reach that level at 1,024. The required total budget therefore depends on the checkpoint and its active certification conditions.

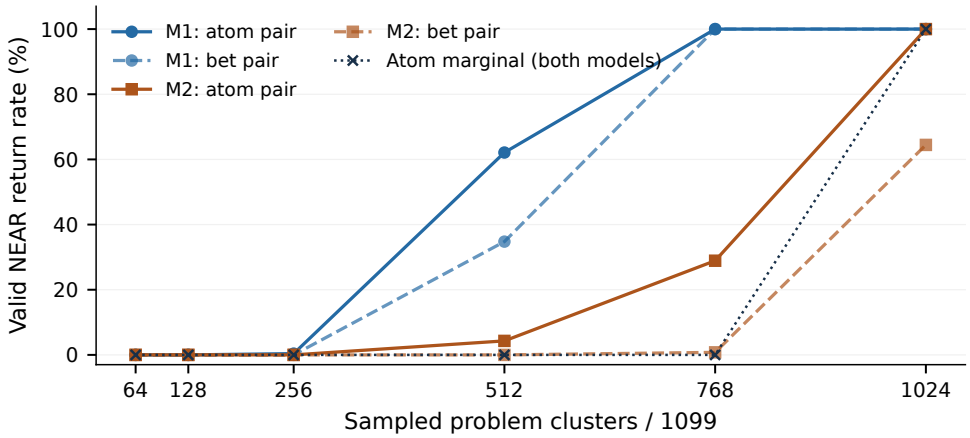


Figure 5: P1099 retrospective power by total newly sampled problems when no outcomes are assumed known. The earliest tabulated budget reaching 80% is a grid summary, not an optimal or continuous sample-complexity estimate.

A separate known-fraction experiment uses 128 permutations, known problem counts $0, 274, 549, 824$, and incremental sample counts $128, 256$. The known part is conditioned on and restored through Equation (7). Known and newly observed outcomes contribute separately to the total observation count. Table 15 and Figure 6 keep the two counts explicit.

Table 15: Known-information ablation on P1099. Entries are Fixed–Near valid returns out of 128. Total observed problems equal the known and new columns added together.

Known	New	Total	M1		M2	
			Paired	Marginal	Paired	Marginal
0	128	128	0/128	0/128	0/128	0/128
0	256	256	1/128	0/128	0/128	0/128
274	128	402	0/128	0/128	0/128	0/128
274	256	530	22/128	0/128	3/128	0/128
549	128	677	7/128	0/128	2/128	0/128
549	256	805	120/128	0/128	28/128	0/128
824	128	952	128/128	5/128	53/128	8/128
824	256	1080	128/128	128/128	128/128	128/128

The predeclared sensitivity analysis retains disagreement target 2%, varies absolute-risk targets over 15%, 20%, and 25%, and relative-gap tolerances over one, three, and five points. It evaluates the saved fixed-CAN_L bounds under alternative target thresholds; the original target remains primary. All four engines and three Q budgets are retained in the numeric evidence. Figure 7 displays the atomic paired 192-problem slice; the complete grid remains a descriptive sensitivity.

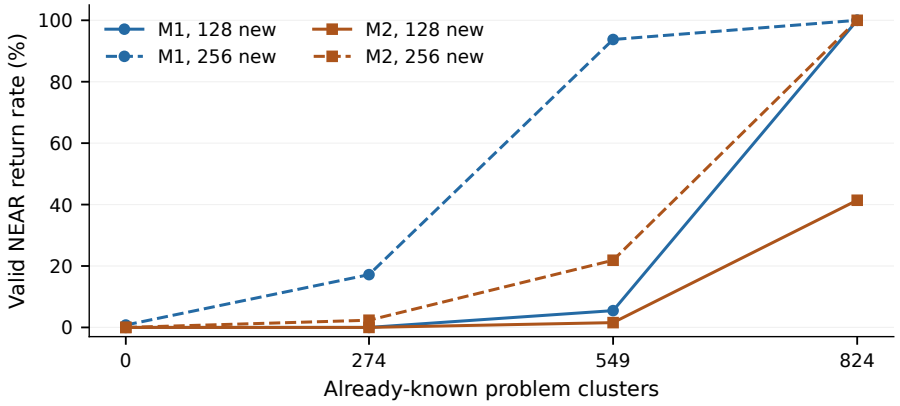


Figure 6: The same incremental budget can have very different value when much of the finite target is already known. This is a conditional-information comparison, not free additional training or an external-population guarantee.

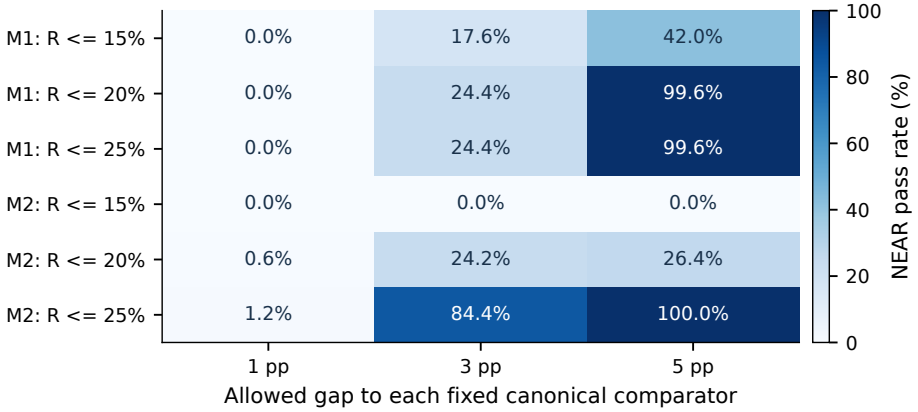


Figure 7: Auxiliary risk-target sensitivity for Fixed–Near ATOM–PAIR at Q192. Cell values are conditional valid-return rates. The original 20%/three-point target remains the primary comparison regardless of this grid.

H MATCHED CLASSICAL PAIRED-INTERVAL REFERENCES

This post-hoc comparison fixes Q235, CAN_L, the original 512 problem permutations, and budgets 64, 128, and 192. Only the relative-risk bounds change. The atomic absolute-risk bound, structurally zero disagreement, three-point comparison tolerance, and self-comparison remain fixed. Nonself comparisons are against CAN_F and CAN_E; sample coincidences retain the declared comparison.

H.1 PAIRED SCORE CONSTRUCTION

For a class and comparator, let (a, b, c, d) count both-error, candidate-only error, comparator-only error, and neither-error pairs. Their sum is the sampled class count n_y. Write $p = (a + b)/n_y$, $q = (a + c)/n_y$, and $\widehat{\delta}_y = p - q$. Newcombe's paired-data method 10 (Newcombe, 1998) combines uncorrected Wilson score limits $[l_p, u_p]$, $[l_q, u_q]$ with the paired association estimate. We use nominal upper-tail level $\eta = 1/80$ per class, equivalently a two-sided confidence level $1 - 2\eta = 0.975$. If

$v = ad - bc$ and $D = \sqrt{(a+b)(c+d)(a+c)(b+d)}$, the corrected association is

$$r = \begin{cases} \max(0, v - n_y/2)/D, & v > 0,\ D > 0, \\ v/D, & v \le 0,\ D > 0, \\ 0, & D = 0. \end{cases}$$

The lower and upper limits are $\widehat{\delta}_y - d_-$ and $\widehat{\delta}_y + d_+$, restricted to $[-1, 1]$, where

$$d_-^2 = (p - l_p)^2 + (u_q - q)^2 - 2r(p - l_p)(u_q - q),$$
$$d_+^2 = (u_p - p)^2 + (q - l_q)^2 - 2r(u_p - p)(q - l_q).$$

Balanced lower and upper endpoints are the arithmetic means of the two class endpoints. Each comparison passes when its balanced upper endpoint is at most 0.03; a complete return additionally requires the unchanged atomic $U_R \le 0.20$. Empty classes receive support bounds, and a fully observed class contributes its exact difference. The paired-binomial construction provides the nominal reference defined in Section 6.

H.2 A SEPARATE FINITE-POPULATION WIDTH SENSITIVITY

For population class size N_y and sample size n_y, the heuristic sensitivity rescales each distance from $\widehat{\delta}_y$ by

$$f_y = \sqrt{\frac{N_y - n_y}{N_y - 1}}, \qquad [L_y^{\mathrm{hFPC}}, U_y^{\mathrm{hFPC}}] = \widehat{\delta}_y + f_y\big([L_y, U_y] - \widehat{\delta}_y\big).$$

Here $N_0 = 104$ and $N_1 = 204$; observed class counts follow each original sample. This separate heuristic rescales Newcombe's interval width; its inferential scope is given in Section 6. The point estimate, sampling identities, and all target thresholds are unchanged.

Table 16: All classical-reference results on the original shared samples. Counts are out of 512. Both fixed Q235 conjunctions are true. Every returned conjunction is true on this roster; this does not test error control under a false conjunction.

Reference	Model	Problem budget	Gap passes	Complete	No return
Newcombe	M1	64	0	0	512
Newcombe	M1	128	0	0	512
Newcombe	M1	192	0	0	512
Newcombe	M2	64	0	0	512
Newcombe	M2	128	0	0	512
Newcombe	M2	192	1	1	511
Newcombe + hFPC	M1	64	0	0	512
Newcombe + hFPC	M1	128	6	6	506
Newcombe + hFPC	M1	192	191	191	321
Newcombe + hFPC	M2	64	0	0	512
Newcombe + hFPC	M2	128	49	1	511
Newcombe + hFPC	M2	192	494	131	381

At 192 problems, the nominal and adjusted rules return in 0 and 191 M1 samples, and one and 131 M2 samples, respectively. The adjusted rule yields more returns than the original atomic paired procedure under this heuristic calibration. Table 17 reports widths and empirical upper coverage. Every balanced upper bound covers its corresponding fixed-roster gap in all 512 samples at each tested budget. Both CAN_L targets are true on Q235. These empirical coverage measurements therefore describe a fixed roster with satisfied targets.

The computation uses the paired-data construction. Every numeric interval calculation returned a resolved endpoint. The implementation reproduces four published examples and satisfies support and swap-symmetry checks. Independent 60-decimal evaluation gives the same threshold and coverage decisions for all class intervals. The numerical release includes contingency counts, unrounded endpoints, and the precise endpoint-agreement tolerance.

Table 17: Empirical one-sided coverage of each balanced upper difference on the fixed Q roster, and mean full interval width in percentage points. Each row retains all 512 samples. These descriptive values do not establish uniform or simultaneous finite-population coverage.

Reference	Model	n	Against	Coverage	Mean width (pp)
Newcombe	M1	64	CAN_F	512/512	25.406
Newcombe	M1	64	CAN_E	512/512	25.479
Newcombe	M1	128	CAN_F	512/512	15.043
Newcombe	M1	128	CAN_E	512/512	15.019
Newcombe	M1	192	CAN_F	512/512	11.218
Newcombe	M1	192	CAN_E	512/512	11.196
Newcombe	M2	64	CAN_F	512/512	22.703
Newcombe	M2	64	CAN_E	512/512	22.694
Newcombe	M2	128	CAN_F	512/512	13.262
Newcombe	M2	128	CAN_E	512/512	13.216
Newcombe	M2	192	CAN_F	512/512	9.859
Newcombe	M2	192	CAN_E	512/512	9.800
Newcombe + hFPC	M1	64	CAN_F	512/512	21.815
Newcombe + hFPC	M1	64	CAN_E	512/512	21.878
Newcombe + hFPC	M1	128	CAN_F	512/512	10.204
Newcombe + hFPC	M1	128	CAN_E	512/512	10.189
Newcombe + hFPC	M1	192	CAN_F	512/512	4.809
Newcombe + hFPC	M1	192	CAN_E	512/512	4.800
Newcombe + hFPC	M2	64	CAN_F	512/512	19.496
Newcombe + hFPC	M2	64	CAN_E	512/512	19.490
Newcombe + hFPC	M2	128	CAN_F	512/512	8.997
Newcombe + hFPC	M2	128	CAN_E	512/512	8.966
Newcombe + hFPC	M2	192	CAN_F	512/512	4.227
Newcombe + hFPC	M2	192	CAN_E	512/512	4.201

I COMPONENT BOTTLENECKS AND TARGET SENSITIVITY

The saved fixed-candidate bounds determine both the marginal passage counts and the joint component states. Table 18 reports the complete partition at every original budget. "Neither" fails both the absolute-risk and gap tests, "R only" passes only the former, "G only" only the latter, and "Both" passes the complete conjunction, since CAN_L has structural $D = 0$. Each row sums to 512. Joint counting retains the dependence between the conditions.

For M2 at 192 problems, atomic pairing has 69 neither, 11 R-only, 308 G-only, and 124 complete outcomes. Its betting counterpart has 401, five, 102, and four. The component partitions locate the operating differences of the fixed implementations.

The complete sensitivity retains all 216 model/engine/budget/target cells described in Appendix G. Table 19 extracts the 20% absolute-risk, 2% disagreement slice at 192 problems. Relaxing the gap requirement changes the comparison threshold at a fixed absolute-risk bound. For M2, the 135 atomic-paired returns at five points therefore coincide with its absolute-risk passage count. The original three-point target remains primary, and the grid provides a descriptive sensitivity.

Table 18: Complete joint component partition for the fixed candidate at all original budgets. Each row sums to 512; Both is the Near return count. The disagreement constraint is structurally satisfied.

Model	Bound construction	Budget	Neither	R only	G only	Both
M1	Atomic marginal	64	473	39	0	0
M1	Atomic marginal	128	126	386	0	0
M1	Atomic marginal	192	0	512	0	0
M1	Atomic paired	64	473	39	0	0
M1	Atomic paired	128	126	386	0	0
M1	Atomic paired	192	0	387	0	125
M1	Betting marginal	64	511	1	0	0
M1	Betting marginal	128	417	95	0	0
M1	Betting marginal	192	3	509	0	0
M1	Betting paired	64	511	1	0	0
M1	Betting paired	128	417	95	0	0
M1	Betting paired	192	3	482	0	27
M2	Atomic marginal	64	512	0	0	0
M2	Atomic marginal	128	496	16	0	0
M2	Atomic marginal	192	377	135	0	0
M2	Atomic paired	64	512	0	0	0
M2	Atomic paired	128	492	16	4	0
M2	Atomic paired	192	69	11	308	124
M2	Betting marginal	64	512	0	0	0
M2	Betting marginal	128	512	0	0	0
M2	Betting marginal	192	503	9	0	0
M2	Betting paired	64	512	0	0	0
M2	Betting paired	128	512	0	0	0
M2	Betting paired	192	401	5	102	4

Table 19: Existing sensitivity slice at absolute-risk target 20%, disagreement target 2%, and 192 problems; each count is out of 512. The primary gap remains three percentage points. The complete 216-cell sensitivity is provided in the numerical data.

Model	Gap target (pp)	Atom marginal	Atom paired
M1	1	0	0
M1	3	0	125
M1	5	9	510
M2	1	0	3
M2	3	0	124
M2	5	0	135

J Computational setting and reproducibility

All evaluation inference used one NVIDIA RTX 5090 GPU, with peak allocated memory of 14.1 GiB. Models and decision rules remained fixed. Evaluating all three F/E/L encodings requires three classifier forwards per endpoint per checkpoint; applying a received-view or canonical rule requires one. CPU statistical reconstruction reuses these fixed observations.

The numerical package supplies P/Q decisions and integer losses, fixed candidate definitions, all 512 original sample permutations, and recorded fixed-candidate bounds. The quick command reconstructs finite-roster truths, component and target-sensitivity counts, and original paired-return vectors. The full command computes the classical paired references and verifies the atomic absolute-risk boundaries with exact integer hypergeometric tails; an independent high-precision implementation checks interval endpoints and passage decisions.

Original forecast-count and coupling-return summaries have a separate arithmetic checker for their means, ranges, and primary forecast error. This reconstructs saved summaries; the original donor-construction and block-shuffle drivers are not distributed. Model inference uses the licensed checkpoints and Appendix A settings. The release contains statistical inputs and model decisions; weights remain available from their original providers.