PATHTAIL: PRESERVING CALIBRATION TARGETS FROM FINITE FORECAST SAMPLES

Anonymous authors
Paper under double-blind review

ABSTRACT

Calibration audits for probabilistic forecasters often access the predictive law only through a finite set of sampled trajectories. When the audited error is defined by a population projection, replacing that projection with an empirical fit can do more than add estimation noise: it can change the target itself. We give an exact counterexample in which the declared population component is null, yet the empirical residual has nonzero expectation. PATHTAIL constructs exact score–companion pairs whose expectations preserve both the population component and its tolerance under finite forecast sampling, enabling anytime-valid sequential auditing without redefining the claim. On finite domains, a target-specific primitive polynomial degree characterizes the minimum fixed forecast budget for bounded exact pairs; this budget can vary with the target even when the nuisance space is unchanged. Controlled experiments with native trajectory samplers from four forecasting models across eight data sources check the score constructions. Paired comparisons on four fixed laws show that exactness does not determine detection power: one exact score gains on rare-state laws but loses on uniform laws, with no average gain in native count replay. PATHTAIL thus separates target preservation, evidence accumulation and sampling efficiency.

1 INTRODUCTION

Forecasting systems sample future trajectories to assess events such as sustained demand or persistent adverse weather (Ansari et al., 2024; Woo et al., 2024; Liu et al., 2025b). For high demand with sharp fluctuations, an audit must distinguish errors in either feature from errors in their combination. A population projection removes the declared additive effects and defines the remaining joint-event error. This *residual component* is our audit target.

In a four-state example, ordinary least squares fitted to eight paths produces a residual of mean $-.01814$ although the population component is zero. This exact expectation averages over all forecast batches; outcome independence does not remove the bias. A sequential test can therefore accumulate evidence for an error introduced by the audit itself. Its input score must preserve the declared population quantity.

PATHTAIL constructs such scores when the predictive law is accessible only through samples. A score and companion share an unknown mean coefficient, preserving both the component and its tolerance. We characterize when exact pairs exist, then examine how their choice and acquisition affect sequential evidence. Our contributions are:

- **Finite-sample distortion of the audit target.** We formalize the population residual component and show that an empirical projection residual can have nonzero mean under the component null (Section 2).
- **Target-specific exact forecast budgets.** We identify the primitive polynomial degree that determines the minimum fixed budget for bounded exact pairs, with a matching construction. With one fixed nuisance space, three event targets require one, two and three paths (Section 3).
- **PATHTAIL: exact scores for sequential auditing.** We combine exact pairs with anytime-valid evidence updates, compare their detection behavior on paired inputs, and derive partition-specific acquisition (Sections 4–5).

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.

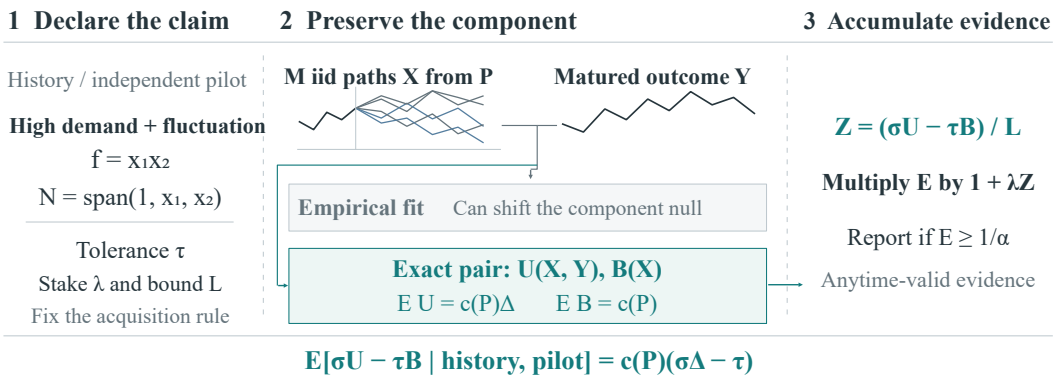


Figure 1: **An exact score–companion pair preserves the declared audit target.** The target, nuisance space, tolerance and update rule are fixed from permitted history or an independent pilot. Empirical projection can shift the component null. An exact pair instead shares a mean coefficient between the score and tolerance correction before evidence accumulates. Paths are schematic; origin indices are omitted. Appendix B.1 specifies the conditional sampling and phase protocol.

2 The Audit Target and Finite-Sample Bias

2.1 Defining the Error to Be Audited

At a forecast origin, let P be the predictive trajectory law and Q the outcome law. A bounded *target function* f maps a trajectory to a quantity of interest, such as an event indicator. Choose a linear *nuisance space* $\mathcal{N}$ of effects to remove from f, with dimension p and containing constants. The population projection under P defines the residual and its mean under Q:

$$r_P^f = (I - \Pi_{\mathcal{N}}^P) f, \qquad \Delta_f(P, Q) = \mathbb{E}_Q r_P^f(Y). \tag{1}$$

The *residual component* Δ_f is the audit target. The raw discrepancy $D = \mathbb{E}_Q f - \mathbb{E}_P f$ splits into the fitted component $D - \Delta_f$ and residual component Δ_f. For an event indicator, all three are in probability units.

For the demand example, let x_1 indicate a high trajectory mean and x_2 a large mean absolute first difference. Their thresholds are fixed before inference. Take $f = x_1 x_2$ and $\mathcal{N} = \mathrm{span}(1, x_1, x_2)$. The residual removes the best additive approximation under the predictive law. If the density ratio dQ/dP belongs to this space, projection orthogonality gives $\Delta_f = 0$, even when the raw event frequency differs. A nonzero component therefore identifies error beyond the declared additive effects (Appendix A.1). This decomposition describes forecast error without attributing a causal mechanism.

A saturated partition gives a useful special case. If g assigns trajectories to cells and $\mathcal{N}$ spans their indicators, then $r_P^f(y) = f(y) - \mathbb{E}_P[f(X) \mid g(X) = g(y)]$. Changing cell frequencies alone leaves Δ_f zero. Opposing within-cell errors can cancel, so separate cell claims require prespecified targets (Appendix A.9).

The audit receives M iid forecast paths $X_{1:M} \sim P$ and an independent matured outcome $Y \sim Q$, conditionally on the permitted history and any independent pilot. The target and nuisance space are fixed before the inference paths are inspected. The forecast budget M counts inference paths; an independent pilot is charged separately. The number T of outcome updates determines how long evidence can accumulate.

2.2 A Zero Component with a Biased Empirical Residual

For the binary interaction above, take state order $(00, 01, 10, 11)$, $P = (.4, .1, .2, .3)$ and $Q = (.22, .105, .21, .465)$. The density ratio $Q/P = (.55, 1.05, 1.05, 1.55)$ lies in $\mathcal{N}$, giving $\Delta_f = 0$. Ordinary least squares fitted to $M = 8$ forecasts has expected residual $-.01814$, with singular batches returning zero.

Table 1: **Notation.** Quantities may carry an origin index t.

Symbol	Meaning
$P, Q; X, Y$	Predictive and outcome laws; forecast path and matured outcome.
$f, \mathcal{N}, p$	Target function, nuisance space and its dimension.
r_P^f, Δ_f, D	Population residual, residual component and raw discrepancy.
$U, B, c(P)$	Exact score, companion and common mean coefficient.
q, M	Primitive residual degree and fixed forecast budget.
$C, C_j; N = C + e_y$	Forecast counts; counts pooled with outcome state y.
$g, K, r, M_{\max}$	Partition, cell count, occupancy quota and cap (OS only).
σ, τ, L, Z	Sign, original-scale tolerance, bound and normalized score.
λ, T, α	Predictable stake, outcome updates and error level.

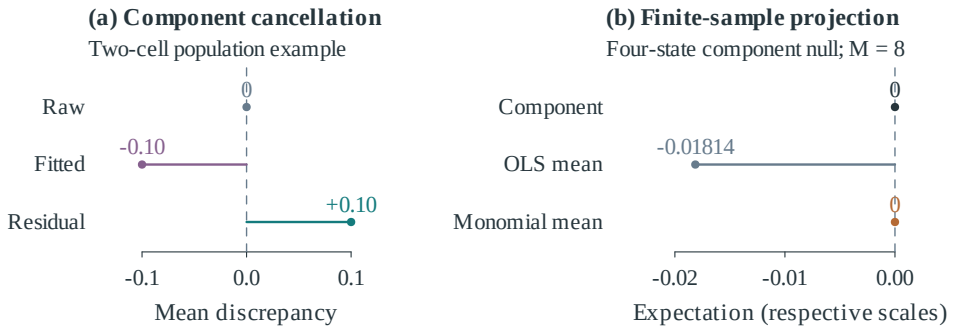


Figure 2: **Population cancellation and bias from empirical projection.** (a) Opposing fitted and residual components cancel in the raw discrepancy: event rates change from $(.2, .8)$ to $(.3, .9)$ and cell masses from $(1/2, 1/2)$ to $(2/3, 1/3)$. (b) In a separate four-state example, the population component is zero but an empirical projection has expected score $-.01814$ with eight paths. Monomial, the binary count form of Volume, remains centered. Panel (b) uses enumeration; the scores have different scales.

Exact enumeration covers all forecast count vectors (Appendix A.2); this discrepancy has no Monte Carlo error. The outcome is independent of the fitting sample throughout.

This failure concerns a fitted projection. In a saturated partition, unsmoothed within-cell sample means already preserve the component when an incomplete batch receives a neutral update. They provide an exact control. Smoothing the means or selecting batches by within-cell values requires a different argument.

2.3 Related Work

Calibration and forecast accuracy describe different properties (Adler et al., 2026). Proper scores, multivariate ranks and kernel tests assess complementary properties (Gneiting & Raftery, 2007; Ziegel & Gneiting, 2014; Allen et al., 2024; Widmann et al., 2021); multicalibration imposes subgroup constraints (Gopalan et al., 2022). Fair scores correct finite-ensemble bias (Ferro, 2014). Here the population projection defines the component; estimating that projection must preserve its null.

Random-design least-squares bias and volume-rescaled correction are established (Derezinski et al., 2019; Dereziński et al., 2022). We use their weighted regression identity for a score and companion. The connection between polynomial degree and minimum sample size is also established (McCormack, 2025). Our finite-domain result identifies the primitive degree of a specified projection residual and constructs its tolerance companion. This determines when common-factor cancellation permits fewer paths than the nuisance dimension. Appendix A.6 separates these steps from the inherited tools. Our sequential updates use established betting and confidence-sequence machinery (Arnold et al., 2023; Waudby-Smith & Ramdas, 2024; Howard et al., 2021; Ramdas et al., 2023), also used in feature-aware quantile audits (Antonov et al., 2026). The new step is constructing the target-specific input score.

3 EXACT SCORES AND FORECAST BUDGETS

3.1 PRESERVING THE COMPONENT AND ITS TOLERANCE

Suppose the audit allows $\sigma\Delta_f \le \tau$, where $\sigma \in \{-1, 1\}$ and $\tau \ge 0$. An exact score satisfies $\mathbb{E}_X U(X_{1:M}, y) = c(P) r_P^f(y)$ at every outcome state, with the same nonnegative coefficient $c(P)$. A *companion* B satisfies $\mathbb{E}_X B = c(P)$. Independence of the outcome yields

$$\mathbb{E}[\sigma U - \tau B] = c(P)\{\sigma\Delta_f(P, Q) - \tau\}, \qquad c(P) \ge 0. \tag{2}$$

Subtracting τB keeps the tolerance in the target's units. Subtracting τ alone would compare the scaled component $c(P)\Delta_f$ with the wrong threshold. A positive coefficient preserves the sign of the discrepancy from tolerance; a zero coefficient supplies no information. The identity concerns expectations, so the realized ratio U/B is not an unbiased component estimate.

For bounded features ϕ spanning $\mathcal{N}$, let $G = \mathbb{E}_P \phi\phi^\top \succ 0$. Determinant-weighted regression gives an exact pair from $M \ge p$ paths with $c(P) = \det G$; singular batches contribute zero. Appendix A derives this *Volume score* from volume-rescaled regression and supplies its deterministic bounds. The nuisance dimension is therefore a sufficient budget; the next result determines when fewer paths suffice.

3.2 HOW MANY FORECAST PATHS ARE NECESSARY?

The sample requirement depends on which part of the residual space the chosen function uses. Consider five states split into $A = \{a_0, a_1, a_2\}$ and $B = \{b_0, b_1\}$. Fix $\mathcal{N} = \mathrm{span}(\mathbf{1}_A, z\mathbf{1}_A, \mathbf{1}_B)$, where $z = 0, 1, 2$ on A. This rank-three space fits a line on A and a constant on B. An event inside B needs one path for exactness; an event inside A needs two. Their union needs three because one common mean coefficient must preserve both residual blocks (Table 2).

This difference can be computed symbolically. Multiply the population residual vector by $\det G(P)$, then cancel the greatest common homogeneous factor of its nonzero coordinates. The remaining *primitive residual vector* has polynomial degree $q(f, \mathcal{N})$. Only the target and feature matrix are specified; the state probabilities remain symbolic.

Theorem 1 (Minimum fixed budget for exact scores)**.** *Let the finite-domain nuisance matrix have full column rank, contain constants, and exclude f from its column span. For unknown interior P, consider bounded scores based on a fixed number of iid forecast paths that satisfy the outcome-wise identity, with a companion of mean $c(P) \ge 0$ not identically zero. A pair exists if and only if*

$$M \ge q(f, \mathcal{N}), \qquad q(f, \mathcal{N}) \le p. \tag{3}$$

A statistic of M categorical draws has an expectation polynomial of degree at most M. Exact preservation forces this expectation vector to be a polynomial multiple of the primitive residual vector, requiring $M \ge q$. Conversely, multinomial counts unbiasedly estimate each degree-q monomial and construct both score and companion. Appendix A.4 proves both directions. The threshold characterizes existence within this exact-score class; detection also depends on the law and the number of outcome updates.

In the five-state example, cancellation leaves degrees one and two for the separate blocks, but no shared factor for their union. Appendix A.5 gives the residuals. The same nuisance space therefore admits different exact budgets solely through target choice.

Table 2: **Changing only the target changes the minimum exact budget.** All rows use the same five states and the same rank-three nuisance space.

Event function f	Nuisance rank p	Minimum budget q
$\mathbf{1}\{b_1\}$	3	1
$\mathbf{1}\{a_2\}$	3	2
$\mathbf{1}\{a_2, b_1\}$	3	3

3.3 Choosing an Exact Score

Volume supplies a general exact pair. For the four-state demand interaction it reduces exactly to the following count form, called *Monomial*. Let C_j be forecast counts, $\chi = (1, -1, -1, 1)$ and $(M)_3 = M(M-1)(M-2)$:

$$U_{\mathrm{mon}}(C, y) = \frac{\chi(y) \prod_{j \neq y} C_j}{(M)_3}, \qquad B_{\mathrm{mon}}(C) = \frac{\sum_y \prod_{j \neq y} C_j}{(M)_3}.$$

Its common mean coefficient is $\sum_y \prod_{j \neq y} P_j$ and its minimum budget is three. It needs counts and products, without fitting a regression or estimating inverse probabilities at each update.

A second exact pair for the same component uses counts pooled with the outcome. Set $N = C + e_y$ and define the *Pooled* score by

$$U_{\mathrm{pool}}(C, y) = \chi(y) \frac{\min_j N_j}{N_y}, \qquad B_{\mathrm{pool}}(C) = \sum_y \frac{\min_j (C + e_y)_j}{C_y + 1}. \tag{4}$$

The outcome may complete a state missing from the forecasts. The score is zero when a state remains missing from the pooled counts. Appendix A.8 proves that this construction maximizes the common mean coefficient among unit-bounded symmetrized exact scores for the complete binary target. At $M = 3$, it equals Monomial after normalizing by Monomial's exact range. At larger budgets, the scores can differ in variance as well as mean. Section 5.3 compares their evidence accumulation on identical inputs.

4 PathTail: A Sequential Audit

4.1 From Forecast Paths to Statistical Evidence

Figure 1 shows the information available at each step. Before current inference, choose the target, nuisance space, tolerance, acquisition rule, stake and bound from the permitted history or independent pilot. The nuisance structure determines which score constructions apply. Draw the forecast paths, then evaluate U_t, B_t when the outcome matures. With a valid deterministic or predictable bound $L_t \geq |\sigma U_t - \tau B_t|$, form $Z_t = (\sigma U_t - \tau B_t)/L_t$ and multiply the evidence by $1 + \lambda_t Z_t$, where $\lambda_t \in [0, 1]$ was chosen before the current inference sample.

Under the conditional null $\sigma \Delta_t \leq \tau$ at every origin, Equation 2 makes the conditional mean of Z_t nonpositive. The evidence product is a nonnegative test supermartingale, so its probability of ever reaching $1/\alpha$ is at most α. Fixed mixtures over stakes retain the guarantee; mixing both signs tests $|\Delta_t| \leq \tau$ (Vovk & Wang, 2021; Waudby-Smith & Ramdas, 2024). Origins maturing together contribute a fixed average score. Separate phases use their own histories and a joint error allocation (Appendices A.11 and B.1).

4.2 Stopping Acquisition for Partition Targets

Occupancy stopping (OS) applies when the nuisance space spans all partition indicators. For K partition cells, sample until every cell has at least r paths or a fixed cap $M_{\max}$ is reached. Let A_r indicate quota success and $\widehat{\mu}_j$ the sampled target mean in cell j. For $f \in [0, 1]$, the OS pair is

$$U_{\mathrm{OS}}(y) = A_r \{f(y) - \widehat{\mu}_{g(y)}\}, \qquad B_{\mathrm{OS}} = A_r. \tag{5}$$

Stopping uses only cell labels. Conditional on the stopped labels, within-cell means remain unbiased, giving $c(P) = \Pr(A_r = 1)$ and bound $1 + \tau$. An incomplete acquisition yields a neutral update (Appendix A.10). At quota one and the same cap, a covered fixed batch has the same success event and more samples for estimating the cell means. OS saves paths while changing that averaging; its power and runtime require separate measurement.

4.3 Interpreting an Alert

For the demand target, a crossing rejects the claim that joint-event error beyond the additive effects stayed within tolerance at every origin. It does not identify the latest origin or the responsible feature. Prespecified target families refine the diagnosis. Historical confidence sequences for sums of U and B describe a component weighted by their mean coefficients. A raw score with the same weights recovers the fitted component by subtraction (Appendix D). Intervention benefits require separate evaluation.

5 Experiments

5.1 Questions, Data and Comparisons

We evaluate target preservation, score choice and acquisition cost (Table 3). Constructed outcome laws make false-alarm and detection rates measurable. The native controlled study uses Aurora, Sundial, Moirai and Moirai-MoE at two contexts from each of eight sources, giving 64 model–context pairs (Wu et al., 2026; Liu et al., 2025b; Woo et al., 2024; Liu et al., 2025a). Each forecast uses a 512-step context and 24-step horizon. Appendix B.4 lists the sources and their sampling grids.

Appendix C.3 reports the broader forecast-quality archive; its marginal quantiles do not specify joint trajectory laws.

Table 3: **Claims and the evidence used to assess them.**

Question	Evidence	What it establishes
Is the target preserved?	Enumeration and native controlled nulls	Exact identities and implementation checks.
Which budget permits exactness?	Theorem and fixed-space example	Target-specific thresholds and matching constructions.
Which score detects better?	Paired, effect-matched confirmation	Power and time differences on four fixed laws.
What transfers to model interfaces?	Native count replay and acquisition timing	Controlled-feedback behavior and measured cost.

5.2 Projection Bias and Null Control

Figure 2 establishes the projection bias by enumeration; 108 further rational configurations check the score–companion identity. A 271-law ablation includes nonzero tolerances and 136 null laws, with 4,000 confirmation streams per law and 512 updates. Across seven valid implementations, the largest null crossing rate is 2.23%; none of 1,904 excess checks is significant after Holm correction. A half-count plug-in reaches 100%, with 50 significant excesses among 272 budget–law checks. The unsmoothed partition score provides an exact control for this comparison.

To compare valid procedures, the projection envelope tests the same component conservatively. Across 24 alternatives at $M = 8$, Monomial detects 26.97% of streams on average and the envelope has no observed crossings; simultaneous paired intervals favor Monomial in 18 configurations. At a much larger budget in the five-state study, the envelope detects as reliably and crosses sooner (Appendix C.4).

5.3 Which Exact Score Accumulates More Evidence?

Exactness leaves a choice of score. Monomial and Pooled share paths, outcomes and an equal mixture over both signs and six stakes, at $M = 8, 32$. Four fixed predictive laws have minimum state masses from .25 to .01. Tilting to $\Delta_f = -.025$ gives the same feasible effect in all four. A 1,000-stream screen precedes 4,000-stream confirmation with disjoint seeds and a prespecified 2,048-update endpoint. Nulls, a weaker component and nonzero tolerances complete 20 configurations (Appendix B.2).

Table 4: **Exact score choice changes detection.** Four fixed laws, $\Delta = -.025$, 4,000 fresh streams per law, 2,048 updates and shared inputs/stakes. Paired differences (Pooled minus Monomial, percentage points) have simultaneous 95% coverage over eight rows. Bold marks resolved differences.

Law	M	Monomial (%)	Pooled (%)	Difference [95% CI]
Uniform	8	**68.50**	66.40	-2.10 $[-3.72, -0.47]$
	32	**91.70**	90.70	-1.00 $[-1.91, -0.08]$
Asymmetric	8	65.15	66.12	$+0.98$ $[-0.83, +2.78]$
	32	93.67	92.88	-0.80 $[-1.68, +0.08]$
Correlated	8	67.77	67.60	-0.17 $[-2.14, +1.79]$
	32	99.62	99.60	-0.02 $[-0.47, +0.42]$
Rare states	8	12.53	**35.27**	$+22.75$ $[+20.58, +24.83]$
	32	52.90	**99.75**	$+46.85$ $[+44.35, +49.20]$

Across the eight primary comparisons, mean detection rises from 68.98% to 77.29%, a paired difference of 8.31 percentage points with a 95% interval $[6.16, 10.46]$.

The aggregate interval describes the fixed four-law matrix and accounts for shared budget prefixes. The gain is concentrated in the rare-state law: Pooled improves detection by 22.75 and 46.85 points at $M = 8$ and $M = 32$. Both uniform-law comparisons favor Monomial; four comparisons remain unresolved under simultaneous intervals. Thus maximizing the common mean coefficient does not order detection power, even at a matched component magnitude.

Exact count enumeration separates score choice from normalization. At tolerance zero, Monomial's bound is already tight, so tightening the bound leaves its updates unchanged. The 60-test null family has maximum crossing rate 3.70% and no Holm-significant excess. Nonzero-tolerance bounds and time-to-detection summaries, including all noncrossing streams, appear in Appendix C.1.

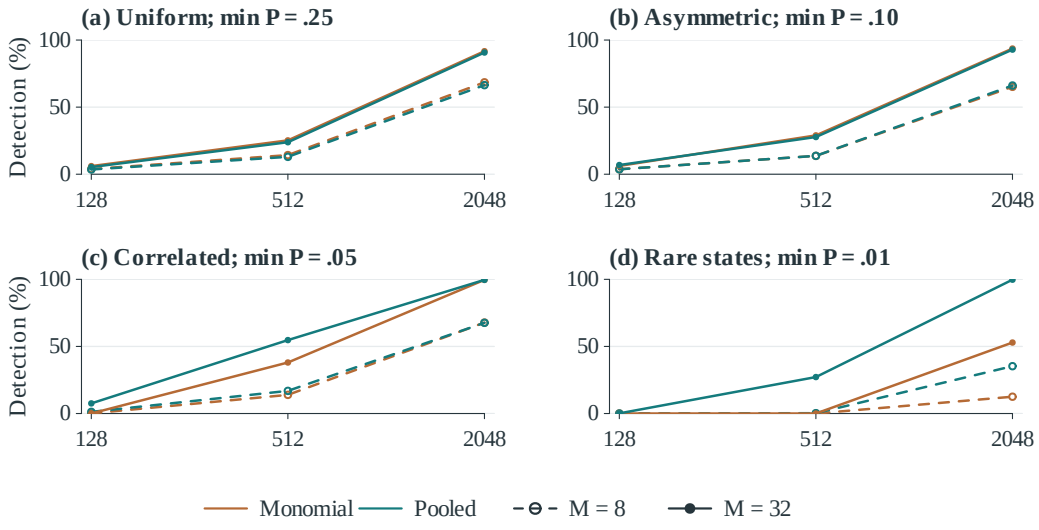


Figure 3: **State probabilities and outcome updates determine the benefit of score choice.** Detection at $128, 512, 2048$ outcome updates, using prefixes of the same 4,000 confirmation streams at $\Delta_f = -.025$. Color identifies score; line type identifies forecast budget. Rare-state gains coexist with uniform-law losses. Appendix C.1 gives the weaker component $\Delta_f = -.005$.

Evidence multiplies $1 + \lambda Z$, with log growth $\mathbb{E}\log(1 + \lambda Z)$. The mean signal $(c(P)/L)(\sigma\Delta_f - \tau)$ and variance alone do not order finite-horizon crossings. Rescaling U, B, L together leaves every factor unchanged. At $M = 8$ under the uniform law, Pooled raises $c(P)/L$ from 1.167 to 1.470 and variance from .152 to .248, yet detection falls. These moments describe the tradeoff without establishing its cause.

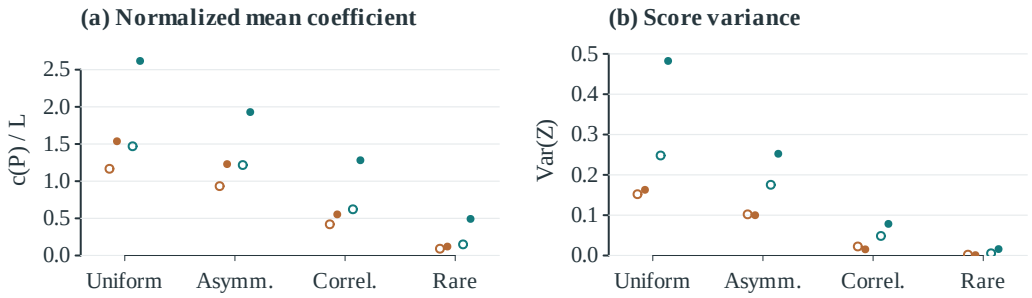


Figure 4: **A larger mean coefficient can accompany larger score variance.** Exact count enumeration at $\Delta_f = -.025$ gives (a) the common coefficient divided by the score bound and (b) normalized-score variance. Orange: Monomial; teal: Pooled. Open: $M = 8$; filled: $M = 32$. These moments alone do not order crossing probabilities.

The two scores are nonzero on the same count–outcome configurations. Under the rare-state law, 98.99% of updates are zero; Pooled changes the weight of the remaining updates (Figure 4).

In a separate study of 112 model-derived reference alternatives, Monomial averages only 3.48% detection at $M = 8$ and 8.03% at $M = 64$, with 512 updates. Replacing the score by its known-P conditional mean raises these rates to 6.71% and 8.60%. Forecast-sampling noise explains only part of the low power (Appendix B.7).

5.4 Native Forecast Controls and Count Replay

An independent 256-path pilot fixes median thresholds for trajectory mean and mean absolute first difference, giving the binary interaction target of Section 2.1. Fresh model draws supply both forecasts and synthetic feedback. Two known component nulls use $Q = P$ and $Q \propto (1 + .75x_1 + .5x_2)P$. The alternative uses $Q \propto (1 + 4f)P$, whose component magnitude varies across contexts. Each context has 256 streams, 256 updates and $M = 8$.

All 64 model–context pairs pass the fixed support gate. The maximum Monomial null crossing rate is 5.08% (13/256; pointwise 95% interval $[2.73, 8.53]\%$), with no significant excess over the nominal 5% level. Empirical OLS reaches 10.55% and the half-count diagnostic reaches 94.53%, with zero and 19 significant excesses, respectively, after the joint correction. The smallest adjusted OLS p-value is .087. Figure 5 shows score means and crossing rates for every null law; the tests assess crossing rates, so a displaced score mean alone does not establish a false-alarm excess. Appendix C.2 reports the model–source summaries and the earlier two-source study separately.

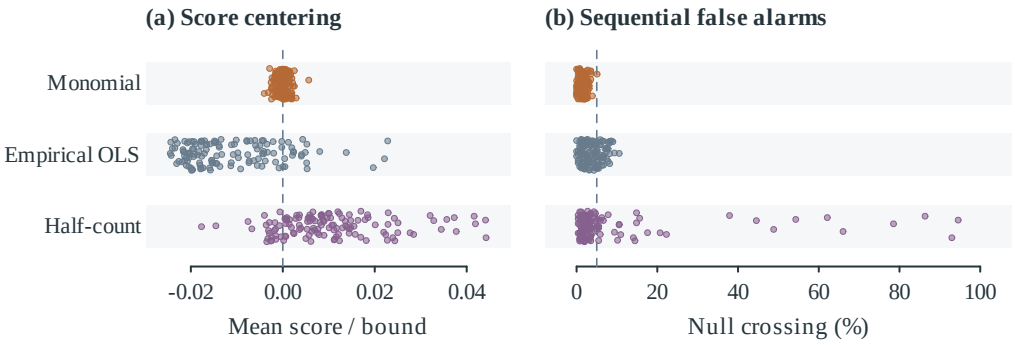


Figure 5: **Finite-sample centering and null crossings across native samplers.** All 128 model–context–null laws across eight sources. Left: mean score divided by its deterministic bound. Right: crossings over 256 streams; dashed line: 5%. Vertical spreading separates overlaps. Monomial preserves the component; OLS and half-count are the specified plug-in diagnostics.

We compare Pooled on the same 64 model–context pairs using recorded counts, outcomes, pilots and stakes. Count replay reproduces every original Monomial score and crossing. Mean alternative detection is 88.12% for Monomial and 87.71% for Pooled. Of 64 simultaneous paired comparisons, two favor Pooled and 62 are unresolved; no loss is resolved. Pooled's maximum null crossing rate is 3.91%, with no significant excess in the separate reanalysis family. Thus Pooled has no average detection gain on these archived inputs. This reanalysis uses no new model draws, and no score is selected for individual contexts.

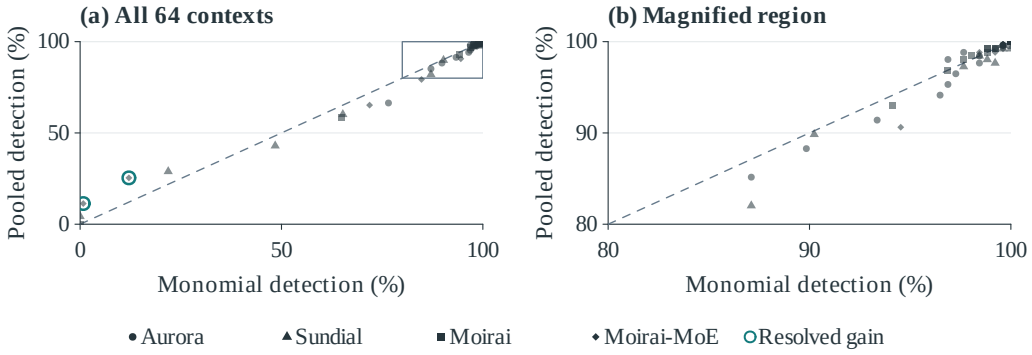


Figure 6: **Native count replay shows no average detection gain.** All 64 alternatives at $M = 8, T = 256$, with 256 streams and a context-specific effect per point. Left: all contexts, including three Monomial and two Pooled rates below 5%. Right: the 80–100% region. Circles mark the two gains resolved by simultaneous paired intervals; the other 62 cross zero.

5.5 Acquisition Costs and Historical Diagnosis

In 18 controlled alternatives, chunked OS improves mean detection by 4.06 points over an ideal fixed budget matched to its expected path use, but loses 2.21 points against fixed-256. Across four native models and 14 sources, the historical partition saves 36–63% of paths while taking 2.18–5.31 times the batched fixed-256 time. For a pilot-defined partition, every acquisition stops after 32 paths and agrees exactly with fixed-32. Appendix C.5 reports the complete cost comparisons, including pilot and failed acquisition costs.

The historical pedestrian and wind study covers 12 traces and 597,990 system–origin pairs. Four residual intervals exclude zero, all among seven raw intervals excluding zero. In the Chronos pedestrian trace with 16 cells, the fitted interval is positive and the residual interval negative: the two components offset each other in the raw discrepancy. The intervals describe weighted historical components; their nominal coverage excludes the retrospective choice of this analysis (Appendix D.4).

6 Discussion and Conclusion

The population projection defines the audit target; PathTail preserves it and its tolerance from finite forecast samples. The primitive degree determines the minimum fixed budget for this exact-score class, including reductions below the nuisance dimension. Exactness permits different detection behavior: Pooled gains on the controlled rare-state law, loses on uniform laws and gives no average native gain.

Current native controls use synthetic feedback, two archived contexts per source and one binary target. Prospective outcomes and larger target families are needed to establish broader diagnostic value. Acquisition efficiency depends on the model interface; pretraining overlap and benefits from subsequent recalibration remain unmeasured.

AI USE STATEMENT

The authors led the research conception and experimental design. Generative AI tools assisted with manuscript revision, literature retrieval, mathematical exposition and proof drafting, code implementation, result analysis, and figure preparation. The authors independently reviewed and verified the mathematical claims, experimental results, citations, code, and AI-assisted artifacts, and take responsibility for the scientific claims and final manuscript.

ETHICS STATEMENT

The studies use public time-series data and synthetic feedback. They do not recruit human participants. A component signal refers to a declared target and nuisance space; it does not establish a causal failure or the value of a subsequent intervention. The Berlin and Tokyo weather archives include reanalysis/model-derived values. Pretraining overlap has not been ruled out.

REPRODUCIBILITY STATEMENT

Anonymous code repository: https://anonymous.4open.science/r/path-7901. Appendix A gives the score formulas and proofs. Appendix B specifies experimental protocols and model interfaces; Appendix C reports additional results. Appendix D derives and applies the historical effect intervals. The native replay uses recorded forecast counts, outcome states and fixed pilot definitions for all 64 contexts to check score and evidence calculations without new model calls (Appendix B.2). Reproducing native trajectory generation additionally requires the original model checkpoints and data contexts; count replay does not regenerate those paths.

REFERENCES

Coen Adler, Yuxin Chang, Felix Draxler, Samar Abdi, and Padhraic Smyth. Beyond accuracy: Are time series foundation models well-calibrated? In *International Conference on Learning Representations*, 2026.

Sam Allen, Johanna Ziegel, and David Ginsbourger. Assessing the calibration of multivariate probabilistic forecasts. *Quarterly Journal of the Royal Meteorological Society*, 150(760A):1315–1335, 2024. doi: 10.1002/qj.4647.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=gerNCVqqtR.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, Mononito Goswami, Shubham Kapoor, Danielle C. Maddix, Pablo Guerron, Tony Hu, Junming Yin, Nick Erickson, Prateek Mutalik Desai, Hao Wang, Huzefa Rangwala, George Karypis, Yuyang Wang, and Michael Bohlke-Schneider. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025. URL https://arxiv.org/abs/2510.15821.

Ivane Antonov, Sohom Mukherjee, Richard Pibernik, and Yo Joong Choe. Bet on features: Anytime-valid and feature-aware auditing of conditional quantile forecasters. *arXiv preprint arXiv:2607.11653*, 2026.

Sebastian Arnold, Alexander Henzi, and Johanna F. Ziegel. Sequentially valid tests for forecast calibration. *The Annals of Applied Statistics*, 17(3):1909–1935, 2023. doi: 10.1214/22-AOAS1697.

Youngseog Chung, Ian Char, and Jeff Schneider. Sampling-based multi-dimensional recalibration. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 8919–8940, 2024.

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.

Michal Derezinski, Manfred K. Warmuth, and Daniel Hsu. Correcting the bias in least squares regression with volume-rescaled sampling. In *International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 944–953. PMLR, 2019. URL https://proceedings.mlr.press/v89/derezinski19a.html.

Michał Dereziński, Manfred K. Warmuth, and Daniel Hsu. Unbiased estimators for random design regression. *Journal of Machine Learning Research*, 23(167):1–46, 2022. URL https://jmlr.org/papers/v23/19-571.html.

Yu Ding. Wind spatio-temporal dataset1. https://doi.org/10.5281/zenodo.5516543, 2021.

Tristan Farran. When your model stops working: Anytime-valid calibration monitoring. *arXiv preprint arXiv:2603.13156*, 2026.

Christopher A. T. Ferro. Fair scores for ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 140(683):1917–1923, 2014. doi: 10.1002/qj.2270.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Rakshitha Godahewa, Christoph Bergmeir, Geoffrey I. Webb, Rob J. Hyndman, and Pablo Montero-Manso. Monash time series forecasting archive. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

Google Research. TimesFM 3.0 (PyTorch): Official model card, 2026. URL https://huggingface.co/google/timesfm-3.0-pytorch. Accessed September 24, 2026.

Parikshit Gopalan, Michael P. Kim, Mihir A. Singhal, and Shengjia Zhao. Low-degree multicalibration. In *Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 3193–3234. PMLR, 2022. URL https://proceedings.mlr.press/v178/gopalan22a.html.

Nicholas J. Higham. *Accuracy and Stability of Numerical Algorithms*. Society for Industrial and Applied Mathematics, 2 edition, 2002.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021. doi: 10.1214/20-AOS1991.

Mert Kayaalp, Caner Türkmen, Oleksandr Shchur, Pedro Mercado, Abdul Fatir Ansari, Michael Bohlke-Schneider, and Bernie Wang. Test-time efficient pretrained model portfolios for time series forecasting. In *International Conference on Learning Representations*, 2026.

Emaad Khwaja, Chris Lettieri, Gerald Woo, Eden Belouadah, Marc Cenac, Guillaume Jarry, Enguerrand Paquin, Xunyi Zhao, Viktoriya Zhukov, Othmane Abou-Amal, Chenghao Liu, Ameet Talwalkar, and David Asker. Toto 2.0: Time series forecasting enters the scaling era. *arXiv preprint arXiv:2605.20119*, 2026. URL https://arxiv.org/abs/2605.20119.

Qi Kuang and Yin Xia. Anytime-valid distribution shift detection via predictive rank martingales. *arXiv preprint arXiv:2609.00536*, 2026.

Xu Liu, Juncheng Liu, Gerald Woo, Taha Aksu, Yuxuan Liang, Roger Zimmermann, Chenghao Liu, Junnan Li, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. Moirai-MoE: Empowering time series foundation models with sparse mixture of experts. In *International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 38940–38962, 2025a.

Yong Liu, Guo Qin, Zhiyuan Shi, Zhi Chen, Caiyin Yang, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Sundial: A family of highly capable time series foundation models. In *International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 39295–39317, 2025b.

Max Planck Institute for Biogeochemistry. Weather station data, beutenberg campus, jena. https://weather.bgc-jena.mpg.de/weather_data.html, n.d. Accessed 2026-08-16.

Andrew McCormack. On the existence of unbiased hypothesis tests: An algebraic approach. *arXiv preprint arXiv:2506.08259*, 2025. URL https://arxiv.org/abs/2506.08259.

Open-Meteo. Historical weather API documentation, 2026. URL https://open-meteo.com/en/docs/historical-weather-api. Accessed September 24, 2026.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023. doi: 10.1214/23-STS894.

Raphael Rossellini, Jake A. Soloff, Rina Foygel Barber, Zhimei Ren, and Rebecca Willett. Can a calibration metric be both testable and actionable? In *Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 4937–4972, 2025.

Salesforce AI Research. Moirai 2.0-R-Small: Official model card, 2025. URL https://huggingface.co/Salesforce/moirai-2.0-R-small. Checkpoint revision 30f43ff; accessed September 25, 2026.

Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021.

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.

David Widmann, Fredrik Lindsten, and Dave Zachariah. Calibration tests beyond classification. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=-bxf89v3Nx.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. doi: 10.1080/01621459.1927.10502953.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53140–53164. PMLR, 2024. URL https://proceedings.mlr.press/v235/woo24a.html.

Xingjian Wu, Jianxin Jin, Wanghui Qiu, Peng Chen, Yang Shu, Bin Yang, and Chenjuan Guo. Aurora: Towards universal generative multimodal time series forecasting. *arXiv preprint arXiv:2509.22295*, 2026. URL https://arxiv.org/abs/2509.22295.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11106–11115, 2021. doi: 10.1609/aaai.v35i12.17325.

Johanna F. Ziegel and Tilmann Gneiting. Copula calibration. *Electronic Journal of Statistics*, 8(2): 2619–2638, 2014. doi: 10.1214/14-EJS964.

A Proofs and Score Constructions

We first establish the population interpretation and projection-bias counterexample, then derive exact score–companion pairs, their minimum fixed budgets and the conditions for sequential validity.

A.1 Population Interpretation

Assume $Q \ll P$ and $h = dQ/dP \in L^2(P)$. The target function f belongs to $L^2(P)$, and $\mathcal{N}$ is a finite-dimensional linear subspace containing constants.

Proposition 2 (Representability and target alignment)**.** *Equation 1 implies: (i) $h \in \mathcal{N}$ gives $\Delta_f = 0$ for every f; (ii) one nonzero Δ_f proves $h \notin \mathcal{N}$; and (iii) $\Delta_f = 0$ for all $f \in L^2(P)$ iff $h \in \mathcal{N}$. If $\dim(\mathcal{N}^\perp) = 1$, every nonzero population projection residual spans the full complement. On a finite interior domain, no fixed target is complete when $\dim(\mathcal{N}^\perp) > 1$: an outside-space density ratio can be orthogonal to it.*

Proof. Write $h = \Pi^P_{\mathcal{N}} h + (I - \Pi^P_{\mathcal{N}})h$. Since $r^f_P \perp \mathcal{N}$,

$$\mathbb{E}_Q r^f_P = \langle h, r^f_P \rangle_P = \langle (I - \Pi^P_{\mathcal{N}})h, r^f_P \rangle_P.$$

This proves (i) and (ii). If the component vanishes for every f, choose $f = h$. Then $r^h_P = (I - \Pi^P_{\mathcal{N}})h$ and the component is $\|r^h_P\|^2_P$, so $h \in \mathcal{N}$. The converse has already been shown.

If $\dim(\mathcal{N}^\perp) = 1$, every nonzero projection residual spans the complement. Conversely, on a finite interior domain with complement dimension greater than one, choose nonzero $v \in \mathcal{N}^\perp$ with $\langle v, r^f_P \rangle_P = 0$. Because v is bounded and $1 \in \mathcal{N}$, $h_\varepsilon = 1 + \varepsilon v$ is a positive normalized density ratio for sufficiently small $|\varepsilon|$. It lies outside $\mathcal{N}$ but has zero component for f. □

A.2 Empirical Projection under the Component Null

The empirical OLS counterexample uses states $(00, 01, 10, 11)$, target function $f(x) = x_1 x_2$, nuisance basis $(1, x_1, x_2)$, and

$$P = (.4, .1, .2, .3), \qquad Q = (.22, .105, .21, .465).$$

The ratio $Q/P = (.55, 1.05, 1.05, 1.55)$ belongs to the nuisance space, so $\Delta_f = 0$. For $M = 8$, let the empirical OLS score be $f(y) - \phi(y)^\top (\Phi^\top \Phi)^{-1} \Phi^\top F_X$ on full-rank batches, and zero otherwise. Exact multinomial enumeration over all forecast count vectors gives full-rank probability $2870217/3125000$ and mean score $-124726893/6875000000$. The Monomial score has mean zero and mean companion $1/20$. Both calculations include singular batches under the stated rule.

A.3 Volume Score and Independent Pilot

With forecast feature matrix Φ, let $A = \Phi^\top \Phi$, $F_X = (f(X_1), \ldots, f(X_M))^\top$, $b_X = \Phi^\top F_X$, and $(M)_p = M(M-1)\cdots(M-p+1)$. The Volume score and companion are

$$B_M = \frac{\det A}{(M)_p}, \qquad U_M(y) = \frac{\det(A) f(y) - \phi(y)^\top \operatorname{adj}(A) b_X}{(M)_p}. \tag{A.1}$$

Theorem 3 (Volume score identity)**.** *Let f and the p-dimensional feature map ϕ be bounded, with constants in the span of ϕ and $G = \mathbb{E}_P \phi\phi^\top \succ 0$. For $M \geq p$ iid forecasts and an independent outcome, the pair in Equation A.1 satisfies $\mathbb{E}U_M(y) = \det(G) r^f_P(y)$ at every outcome state and $\mathbb{E}B_M = \det G$. Thus Equation 2 holds, including on singular forecast batches.*

Proof of Theorem 3. Cauchy–Binet gives

$$\det(\Phi^\top \Phi) = \sum_{I \subset [M],\, |I| = p} \det(\Phi_I)^2, \qquad \mathbb{E} \det(\Phi_I)^2 = p! \det G.$$

Since $(M)_p = p!\binom{M}{p}$, this proves $\mathbb{E}B_M = \det G$. The volume-rescaled least-squares identity (Derezinski et al., 2019) gives, for each p-row subset,

$$\mathbb{E}\Big[\det(\Phi_I)^2\{f(y) - \phi(y)^\top \widehat{\beta}_I\}\Big] = p!\det(G)r(y).$$

Summing over subsets gives $\mathbb{E}U_M(y) = \det(G)r(y)$; averaging over $Y \sim Q$ and using the identity for B_M proves Equation 2. The sum is exactly the adjugate expression in Equation A.1. If A is singular, $b_X \in \mathrm{range}(A)$; hence $\mathrm{adj}(A)b_X = 0$ when $\mathrm{rank}(A) = p - 1$, while the adjugate itself is zero at smaller rank. Thus the polynomial expression gives $B_M = U_M = 0$ without inverting A. □

Exhaustive enumeration on 240 randomly generated finite-domain designs with $p \le 3$ and $M \in \{p, p+1\}$ gives maximum expectation error 2.66×10^{-15}. Another 500 checks of the one-hot specialization give maximum discrepancy 2.61×10^{-18}.

If every p-row feature determinant is bounded by D_ϕ and every $(p+1)$-row determinant after appending f is bounded by D_{aug}, the same subset expansion gives

$$0 \le B_M \le D_\phi^2/p!, \qquad |U_M(y)| \le D_\phi D_{\mathrm{aug}}/p!. \tag{A.2}$$

Consequently $(D_\phi D_{\mathrm{aug}} + \tau D_\phi^2)/p!$ is a deterministic bound for $|\sigma U_M - \tau B_M|$. Hadamard's inequality supplies finite prespecified choices from bounds on $\|\phi\|_2$ and $|f|$.

Corollary 4 (Independent pilot)**.** *Let an independent forecast pilot H define bounded f_H, ϕ_H, and an availability gate $a(H) \in \{0, 1\}$. Conditional on H, multiply both scores in Equation A.1 by $a(H)$. Then*

$$\mathbb{E}(B_M \mid H) = a(H)\det G_H, \qquad \mathbb{E}\{U_M(Y) \mid H\} = a(H)\det(G_H)\Delta_{f_H}.$$

A failed pilot gate therefore gives a neutral exact update. Replacing the declared map after seeing the outcome, or silently reducing its dimension, would define a different claim.

For $\phi = (1, x_1, x_2)^\top$ on $(00, 01, 10, 11)$, every three-state feature minor has squared determinant one. Cauchy–Binet therefore gives $\det A = \sum_y \prod_{j \ne y} C_j$. The determinant-weighted residual at y is $\chi(y)\prod_{j \ne y} C_j$: only the three-state subset omitting y contributes. Dividing both quantities by $(M)_3$ gives exactly $U_{\mathrm{mon}}, B_{\mathrm{mon}}$ in Section 3.3. Monomial is therefore the four-state count specialization of Volume.

A.4 Target-Specific Exact-Score Degree

On a finite n-state domain, let $\Psi \in \mathbb{R}^{n \times p}$ have full column rank and contain the constant vector, and assume $f \notin \mathrm{col}(\Psi)$. Let $D_P = \mathrm{diag}(P)$, $G(P) = \Psi^\top D_P \Psi$, $d(P) = \det G(P)$, and

$$\mathcal{R}(P) = d(P)f - \Psi\,\mathrm{adj}\{G(P)\}\Psi^\top D_P f.$$

Then $r_P^f = \mathcal{R}(P)/d(P)$. Let κ be the greatest common homogeneous polynomial divisor of the nonzero coordinates of $\mathcal{R}$, oriented so that $a = d/\kappa > 0$ on the interior simplex; put $V = \mathcal{R}/\kappa$ and $q = \deg V$.

Theorem 5 (Exact-score degree)**.** *A bounded unknown-P score pair based on M iid forecast draws, with*

$$\mathbb{E}U_M(X_{1:M}, y) = c(P)r_P^f(y), \qquad \mathbb{E}B_M(X_{1:M}) = c(P) \ge 0$$

for every interior P and state y, and mean coefficient c not identically zero, exists if and only if $M \ge q$. Consequently the minimal exact degree is specific to the target and nuisance space and is at most p; it need not equal the nuisance rank.

Proof. Choose a fixed vector v satisfying $v^\top\Psi = 0$ and $v^\top f \ne 0$. Since $v^\top\mathcal{R}(P) = d(P)v^\top f$, κ divides d. Thus a is a homogeneous degree-q polynomial. The coordinates of $\mathcal{R}$ never vanish simultaneously on the interior, so κ has constant sign there; orient it so that $a(P) > 0$.

For necessity, average any randomized score over its internal randomness and forecast permutations. For each outcome state y, its forecast expectation $H_y(P)$ has a homogeneous degree-M multinomial representation (a lower-degree expression can be homogenized using $\sum_x P_x = 1$). The common-component identity implies $H_y V_z = H_z V_y$ on the positive simplex and hence as a polynomial

identity. Over the fraction field, $H = tV$. If $t = A/B$ is reduced, polynomiality of every H_y makes B divide every nonzero V_y. Their greatest common divisor is one, so B is constant. Therefore t is a polynomial of degree $M - q$; a nonzero score pair requires $M \geq q$.

For sufficiency, any homogeneous degree-q polynomial $F(P) = \sum_{|c|=q} F_c P^c$ has the unbiased multinomial estimator assigning count vector c the value $F_c / \binom{q}{c_1,\ldots,c_n}$. Apply this to V_y and a, ignoring extra draws. A deterministic scale S_0 bounds all finite count/state values. Then $\mathbb{E}U(X, y) = a(P) r_P^f(y)/S_0$ and $\mathbb{E}B = a(P)/S_0$; independence of $Y \sim Q$ gives $\mathbb{E}U(X, Y) = a(P)\Delta_f(P, Q)/S_0$. A signed tolerance-adjusted score has mean $a(P)\{\sigma\Delta_f(P, Q) - \tau\}/S_0$. □

Table 5: Exact symbolic checks of the algebraic degree. The degree depends on the target–space geometry, not only on nuisance rank.

Finite-domain design	States n	Nuisance rank p	Exact degree q
Complete binary, $s = 2$	4	3	3
Five-state cubic target	5	3	3
Partially saturated indicator target	4	3	1

For the partially saturated design in Table 5, take nuisance columns $(1, \mathbf{1}\{x = 0\}, \mathbf{1}\{x = 1\})$ and target $\mathbf{1}\{x = 2\}$. Then

$$U = \mathbf{1}\{X = 3, Y = 2\} - \mathbf{1}\{X = 2, Y = 3\}, \qquad B = \mathbf{1}\{X \in \{2, 3\}\}$$

already gives $\mathbb{E}U = (P_2 + P_3)\Delta$ and $\mathbb{E}B = P_2 + P_3$ from one forecast. This example rules out replacing q by a universal function of rank alone.

Relation to algebraic unbiased testing. McCormack (2025) relates separating-polynomial degree to minimum sample size and discusses product multinomial sampling. Theorem 5 uses this established principle to identify a computable primitive degree for the declared residual vector. Different sample roles alone do not imply a strict threshold separation. Indeed, consider a randomized level-α test based on M forecast draws and one outcome that has power greater than α for every Q with $\Delta_f(P, Q) > 0$, and power at most α throughout $\Delta_f(P, Q) \leq 0$. For fixed interior P, put $a_P(y) = \mathbb{E}_{X_{1:M}}[\varphi(X_{1:M}, y)] - \alpha$. Power minus α is $Q^\top a_P$, linear in Q. Continuity from the two sides forces it to vanish on the interior hyperplane $Q^\top r_P^f = 0$. Because $P^\top r_P^f = 0$ and P is interior, this implies $a_P = c(P) r_P^f$ with $c(P) > 0$. Thus an outcome-wise common coefficient can already be forced by unbiased testing under the matched scalar sign null. Theorem 5 requires the general algebraic companion to have nonnegative *expectation*. Its realized value may be negative; the Volume companion is nonnegative on every batch.

A.5 Changing the Target with a Fixed Nuisance Space

Order the five states as $(a_0, a_1, a_2, b_0, b_1)$, with positive probabilities $(p_0, p_1, p_2, p_3, p_4)$. Use the same feature matrix for all three targets:

$$\Psi = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \quad s_B = p_3 + p_4, \quad D_A = p_0 p_1 + 4 p_0 p_2 + p_1 p_2.$$

The columns span $\mathbf{1}_A, z\mathbf{1}_A, \mathbf{1}_B$; thus the constant vector belongs to the space and its rank is three. The Gram determinant is $d = s_B D_A$.

For $f_B = \mathbf{1}\{b_1\}$, the residual is

$$r_P^{f_B} = (0, 0, 0, -p_4, p_3)/s_B.$$

The nonzero coordinates of $d r_P^{f_B}$ have greatest common factor D_A, leaving a primitive vector of degree one. For $f_A = \mathbf{1}\{a_2\}$,

$$r_P^{f_A} = (p_1 p_2, -2 p_0 p_2, p_0 p_1, 0, 0)/D_A.$$

The common factor is s_B, and the primitive degree is two. Finally, the disjoint union target $f_A + f_B$ has numerator

$$(s_B p_1 p_2, -2 s_B p_0 p_2, s_B p_0 p_1, -D_A p_4, D_A p_3).$$

Its coordinates have no nonconstant common factor: the first three share exactly s_B, the last two share exactly D_A, and s_B and D_A involve disjoint variable blocks. Its primitive degree is three. Theorem 5 gives the respective minimum budgets $1, 2, 3$ and matching count constructions. Exact symbolic calculations verify all three residuals and $\Psi^\top D_P r = 0$.

This example isolates target choice: the nuisance matrix alone cannot determine whether a given fixed forecast budget admits an exact pair. In the four-state, rank-three interaction design, the residual complement is one-dimensional; every nonzero residual target is a constant multiple of the same direction and has the same primitive degree.

A.6 Relation to Prior Theory

Table 6: **Theoretical sources and the residual-audit derivation.**

Source	Established tool	Use in this work
Derezinski et al. (2019); Dereziński et al. (2022)	Unbiased least-squares correction under volume rescaling.	Weight an iid-batch residual and pair it with the same determinant mean to preserve the component and tolerance.
McCormack (2025)	Polynomial degree characterizes sample requirements for unbiased testing, including multinomial products.	Compute the primitive degree of the specified projection residual and construct a matching companion.
Howard et al. (2021); Ramdas et al. (2023)	Nonnegative evidence processes, mixtures and anytime error control.	Supply an exact conditional score before applying the established evidence update.

Volume supplies an exact pair at budget p. The primitive-vector argument in Appendix A.4 identifies common-factor cancellations that permit fewer paths and proves that no smaller fixed budget can preserve the same residual.

A.7 Residuals for Complete Binary Targets

For the binary specification of Section 3.3, let $x \in \{0,1\}^s$, $P_x > 0$, and $f(x) = \prod_{j=1}^s x_j$. The nuisance space is the span of all proper-subset monomials. Write $\chi(x) = (-1)^{s - \sum_j x_j}$. The residual has the form

$$H(P) = \sum_x P_x^{-1}, \quad r_P(x) = \frac{\chi(x)}{P_x H(P)}. \tag{A.3}$$

Its one-dimensional complement gives

$$\Delta(P, Q) = 0 \quad \Longleftrightarrow \quad dQ/dP \in \mathcal{N}. \tag{A.4}$$

Let $m_A(x) = \prod_{j \in A} x_j$. For every proper subset $A \subsetneq [s]$,

$$\left\langle \frac{\chi}{P}, m_A \right\rangle_P = \sum_{x \in \{0,1\}^s} \chi(x) m_A(x) = 0,$$

because summing over any coordinate outside A cancels parity signs. Thus χ/P spans $\mathcal{N}^\perp$. Projecting the all-high indicator f onto this direction gives coefficient $\langle f, \chi/P \rangle_P / \|\chi/P\|_P^2 = 1/H(P)$, proving Equation (A.3). Since the complement is one-dimensional, equivalence (A.4) follows from Proposition 2.

A.8 Pooled Score for Complete Binary Targets

We derive the Pooled pair and its maximum common mean coefficient among unit-bounded exact scores for the complete binary target. Section 5.3 compares its detection power with Monomial.

For pooled counts $N = C + e_y$, define

$$a(C, y) = \frac{\min_x N_x}{N_y}, \qquad U_{\text{pool}}(C, y) = \chi(y)a(C, y), \qquad B_{\text{pool}}(C) = \sum_y a(C, y), \tag{A.5}$$

with zero score if a state is missing from the pooled counts. Then

$$\mathbb{E}_{P^M} U_{\text{pool}}(C, y) = c_M(P) r_P(y), \quad \mathbb{E}\{\sigma U_{\text{pool}}(C, Y) - \tau B_{\text{pool}}(C)\} = c_M(P)\{\sigma\Delta(P, Q) - \tau\}, \tag{A.6}$$

where $c_M(P) = \mathbb{E}_{P^M} B_{\text{pool}}(C)$.

Theorem 6 (Complete binary target)**.** *For an interior law on $K = 2^s$ binary states, the all-high target adjusted for all proper-subset monomials has minimum exact budget $K-1$. The count construction in Equation A.5 maximizes the common nonnegative mean coefficient among unit-bounded symmetrized scores in this exact class.*

Proof of Theorem 6. For a full-support pooled count vector n with total $M + 1$, the contribution of forecast count $C = n - e_y$ to $P_y \mathbb{E}a(C, y)$ is

$$P_y \frac{M! \prod_x P_x^{n_x - \mathbf{1}\{x=y\}}}{(n_y - 1)! \prod_{x \neq y} n_x!} \frac{\min_x n_x}{n_y} = \frac{M! \min_x n_x}{\prod_x n_x!} \prod_x P_x^{n_x},$$

which is independent of y. Call the sum of these terms $A_M(P)$. Then $\mathbb{E}a(C, y) = A_M(P)/P_y$ and $\mathbb{E}B_{\text{pool}}(C) = A_M(P)H(P) = c_M(P)$. Multiplication by $\chi(y)$ gives the first identity in (A.6); averaging over $Y \sim Q$ and subtracting the companion gives the second. For interior P, $A_M(P) > 0$ exactly when a pooled count can contain all K states, namely $M + 1 \geq K$.

For the amplitude statement, symmetrize any unit-bounded exact score pair and call its count value $t(C, y)$. The polynomials $P_y \mathbb{E}t(C, y)/\chi(y)$ must agree across y. Their common homogeneous polynomial is divisible by $\prod_x P_x$. Hence it is zero for $M + 1 < K$, and all count coefficients vanish. When a full-support pooled count is possible, coefficient matching gives

$$t(C, y) = w(C + e_y)\chi(y) \frac{\min_x (C + e_y)_x}{C_y + 1}, \qquad |w| \leq 1.$$

The common coefficient equals $H(P)\mathbb{E}_{N \sim \text{Mult}(M+1, P)}\{w(N) \min_x N_x\}/(M+1)$ and is maximized simultaneously for every P by $w \equiv 1$, giving companion (A.5). □

The coefficient bound applies to bounded exact scores using a fixed budget under unknown P. It does not cover known-P scores, conservative or approximate identities, or additional pilot samples.

A.9 Cell-Specific Diagnostic Targets

The choice of target determines what can be localized. A single partition residual averages errors across cells. To examine a particular cell j, one can instead declare the target $f_j(x) = \mathbf{1}\{g(x) = j\}f(x)$ with the same cell-indicator nuisance space. Its component is the outcome probability of that cell times its within-cell target error. Prespecified cell targets can distinguish opposing errors that cancel in the aggregate, provided error is allocated across targets and signs. An alert identifies the target and its tolerance. Our experiments evaluate individual components; extending them to larger families requires measuring the cost of this simultaneous error control.

A.10 Occupancy Stopping for Partition Targets

Condition throughout on the pre-inference history and any independent pilot. Let g be a fixed partition into K cells, with probabilities $p_j = P(g(X) = j) > 0$, and let $f \in [0, 1]$ be fixed. Set $\mu_j = \mathbb{E}_P[f(X) \mid g(X) = j]$ and $r_P^f(y) = f(y) - \mu_{g(y)}$. The nuisance space is the full span of the partition indicators. The forecasts $X_1, X_2, \ldots$ are iid P and independent of $Y \sim Q$. Put $L_i = g(X_i)$.

Fix an integer quota $r \geq 1$ and finite cap $M_{\max} \geq Kr$. Define

$$C_j(m) = \sum_{i=1}^{m} \mathbf{1}\{L_i = j\}, \qquad H_r = \inf\{m : \min_j C_j(m) \geq r\},$$
$$N_{\mathrm{OS}} = \min(H_r, M_{\max}), \quad A_r = \mathbf{1}\{H_r \leq M_{\max}\}.$$

On $A_r = 1$, set $\widehat{\mu}_j = C_j(N_{\mathrm{OS}})^{-1} \sum_{i \leq N_{\mathrm{OS}}: L_i = j} f(X_i)$. The scores $U_{\mathrm{OS}} = A_r(f(Y) - \widehat{\mu}_{g(Y)})$ and $B_{\mathrm{OS}} = A_r$ are set to zero on $A_r = 0$.

Proposition 7 (Stopped partition identity)**.** *For every fixed y,*

$$\mathbb{E}_X U_{\mathrm{OS}}(X, y) = \Pr(H_r \leq M_{\max}) r_P^f(y), \qquad \mathbb{E}_X B_{\mathrm{OS}} = \Pr(H_r \leq M_{\max}).$$

Hence Equation 2 holds for any independent outcome law Q, sign σ and tolerance $\tau \geq 0$.

Proof. Condition on N_{OS} and the stopped label sequence $L_{1:N_{\mathrm{OS}}}$. The event that this sequence is a stopped sequence depends only on the labels. Conditional on it, the values $f(X_i)$ retain their conditional laws given L_i. In particular, on success, $\mathbb{E}[\widehat{\mu}_j \mid N_{\mathrm{OS}}, L_{1:N_{\mathrm{OS}}}] = \mu_j$ for every cell. Since A_r is label-measurable,

$$\mathbb{E}[U_{\mathrm{OS}}(X, y) \mid N_{\mathrm{OS}}, L_{1:N_{\mathrm{OS}}}] = A_r\{f(y) - \mu_{g(y)}\}.$$

Integrating over the label sequence gives the outcome-wise identity. Integrating over independent Y gives the population component. The bounds $|U_{\mathrm{OS}}| \leq A_r \leq 1$ give the normalization $L = 1 + \tau$. □

Permitted stopping decisions. The proof also allows independently chosen positive integer quotas for individual cells; the experiments use equal quotas one and two. During inference, the target function, partition and quotas stay fixed. Stopping may use counts, but neither within-cell target values nor the outcome. The argument relies on the partition structure: it does not justify dividing by a realized determinant for continuous nuisance features or estimate μ_j in a P-null cell visited under Q.

At fixed budget M, write $A(C) = \mathbf{1}\{\min_j C_j > 0\}$. Every fixed count-only weight $a(C) \in [0, 1]$ that vanishes when a cell is unobserved gives $\mathbb{E}[a(C)(f(Y) - \widehat{\mu}_{g(Y)})] = (\mathbb{E}a(C))\Delta_f$. Both $A(C)$ and the normalized product-count weight B of Appendix D.1 qualify, with $\mathbb{E}A \geq \mathbb{E}B$. A larger mean coefficient does not order detection power: the variance and the historical weighting also change.

OS has a random acquisition cost bounded by its cap, so the fixed-budget characterization in Theorem 5 does not apply to its expected path use.

Outcome matching and its access requirements. A matched-cell estimator draws until the cell $g(Y)$ has enough matches. To stop generation this way, it must know $g(Y)$ while querying the original forecast law. That access is distinct from online acquisition before feedback. If all paths are generated and cached before Y, later matching must be charged the entire generated batch. The 225-row comparison includes both access modes, with rejected paths charged as calls.

Cap handling also changes the guarantee. For the positive tail, an incomplete matched quota is completed with the adverse endpoint one for each missing slot. The resulting conditional reference mean is conservative. Returning a neutral update only on unsuccessful outcome-cell matches would select the outcome cell through its hit probability and need not preserve the component. Thus outcome matching has distinct access requirements and uses a conservative update when its quota is incomplete.

A.11 From Per-Origin Scores to Sequential Evidence

Condition on the pre-inference history and independent pilot at each origin. Let L_t be a prespecified bound on $|\sigma U_t - \tau B_t|$; for OS, take $L_t = 1 + \tau$. Put $Z_t = (\sigma U_t - \tau B_t)/L_t$. Under the joint null $\sigma\Delta_t \leq \tau$ for every origin, Equation 2 gives $\mathbb{E}[Z_t \mid \mathcal{F}_{t-1}] \leq 0$. For past-predictable $\lambda_t \in [0, 1]$,

$$E_T = \prod_{t \leq T} (1 + \lambda_t Z_t)$$

is nonnegative and has conditional expected multiplier at most one. Ville's inequality bounds its probability of ever reaching $1/\alpha$ by α. Fixed mixtures and fixed allocations across signs, targets and phases retain this guarantee (Ramdas et al., 2023; Howard et al., 2021).

If several origins mature together, their fixed average score also has nonpositive conditional mean. Cross-origin independence within that group is unnecessary; each origin's conditional forecast/outcome identity remains necessary. A phase updates after all its target coordinates mature and before its next group of forecasts. Other-phase feedback does not enter its stakes. A union bound covers phases without multiplying processes adapted to different histories. Timestamps make the schedule admissible under these assumptions but do not prove the conditional sampling model.

In numerical work, QR or Cholesky and a log determinant avoid explicit matrix inversion (Higham, 2002). The polynomial definition specifies zero scores on singular batches. Rank truncation or regularization changes the nuisance space unless its effect is explicitly controlled.

B Experimental Protocols

The controlled studies use either native model draws with synthetic feedback or specified mixtures of empirical reference banks. Their component nulls are known by construction. Historical outcomes have unknown components and are analyzed separately in Appendix D. This appendix specifies targets, sampling designs and error allocations; Appendix C reports supplementary results.

B.1 Information and Sampling Order

Let $\mathcal{H}_t$ contain the permitted phase history, the current forecast context and an independent pilot, when one is used. Conditional on $\mathcal{H}_t$, fix $f_t, \mathcal{N}_t, \tau_t$, the acquisition rule, the normalizer L_t and stake λ_t. For a fixed budget, require $X_{t,1:M} \sim P_t^M$ and $Y_t \sim Q_t$ independently given this information. A pilot may select the target; the score identity is then applied conditionally to that selected target. The pilot is not reused as an iid inference batch. The normalizer must bound the score over every permitted forecast batch and outcome.

When Y_t matures, compute U_t, B_t and update the phase's evidence. This is when the outcome first enters the update. Acquisition may have stopped earlier under the label-only rule of Appendix A.10. The next stake can depend on this completed update. For overlapping horizons, a phase completes its fixed group of outcomes before issuing its next inference group; feedback from other phases is excluded from its stakes. Fixed error allocations cover the phases jointly. These rules specify the filtration. Conditional iid forecast sampling and forecast–outcome independence remain assumptions; nonoverlapping timestamps alone do not establish them.

B.2 Paired Comparison of Exact Binary Scores

Fixed laws and effects. The predictive laws, in state order $(00, 01, 10, 11)$, are

$$P^{\text{uniform}} = (.25, .25, .25, .25), \quad P^{\text{asymmetric}} = (.4, .1, .2, .3),$$
$$P^{\text{correlated}} = (.45, .05, .05, .45), \quad P^{\text{rare}} = (.49, .01, .01, .49).$$

All use $f = x_1 x_2$ and $\mathcal{N} = \text{span}(1, x_1, x_2)$. Besides $Q = P$, the nuisance null is $Q_0 = P(1 + .5h)$, where h is the centered sum $x_1 + x_2$ divided by its maximum absolute centered value. Alternatives exponentially tilt Q_0 along the population residual, solving for components $-.005$ and $-.025$. The negative sign makes the same strong effect feasible in every law. The audit mixes both signs equally. The asymmetric and rare-state laws add tolerance $\tau = .005$, at boundary component $-.005$ and alternative $-.015$. This gives 20 predictive/outcome-law configurations for confirmation.

Pairing and confirmation. The prespecified seed is 2026092601. A 1,000-stream, 512-update screen uses the four nuisance nulls and four strong alternatives. Confirmation uses a disjoint seed namespace, 4,000 streams per configuration and 2,048 updates. The primary family comprises the eight strong-alternative comparisons. The screen does not select laws, scores or stakes. Budgets $M = 8, 32$ share forecast prefixes through independent multinomial increments, and scores share each forecast count and outcome. Horizons $128, 512, 2048$ are prefixes of the same streams. The equal mixture over two signs and stakes $\{.05, .1, .2, .4, .8, .95\}$ crosses at 20.

Separating score choice from normalization. We compare Monomial under two bounds and Pooled under its tight bound. The first Monomial bound adds separate bounds for U and τB; the tight bound enumerates $\max_{C,y,\sigma} |\sigma U(C, y) - \tau B(C)|$ exactly for each M, τ. At zero tolerance the Monomial bounds coincide. At $M = 3$, the normalized Monomial and Pooled pairs agree for every count vector and outcome.

Inference and reporting. Each primary paired interval bounds the probabilities of the two discordant crossing events using exact binomial intervals. Bonferroni allocation covers both tails and all eight differences simultaneously at 95%. The single aggregate contrast averages budgets within each replicate before averaging the four fixed laws. A bounded-variable Hoeffding interval accounts for this shared-prefix dependence. Its scope is the fixed four-law comparison. Other intervals are descriptive unless a family is explicitly identified. Null excess tests compare crossing counts with .05 using exact one-sided binomial tests. The confirmation's 60-test Holm family includes all three implementations, both budgets, the eight zero-tolerance nulls and two nonzero-tolerance boundaries. We report restricted mean crossing time $\mathbb{E}[\min(T_{\text{cross}}, T)]$ over all streams, together with the fraction that never crosses.

Archived native replay. We evaluate Pooled on all 64 archived model–context pairs, keeping pilot thresholds, forecast counts, outcomes, stakes and horizons fixed. The 64 alternative differences use simultaneous paired intervals; 384 null checks form a separate Holm family that includes both Monomial normalizations. A separate replay verifies the original Monomial scores and crossings and reproduces crossing counts and restricted mean crossing times for all 384 method–law results. These checks use recorded synthetic feedback and require no new model draws; they provide no prospective outcome evidence.

B.3 Native Sampling and Target Construction

Initial interface study. Each model uses one Weather and one Electricity context: the first evaluation origin of the first metadata-selected series in the fourteen-source archive. This gives eight model–context pairs, with 512 independent streams per outcome law.

Domain and context extension. The extension fixes the eight sources in Table 7 and two contexts per source. For context index $c \in \{0, 1\}$, use metadata-selected series c when it exists, and series zero otherwise. Within that series, select evaluation row $\lfloor (c + 1)n_{\text{eval}}/3 \rfloor$ using zero-based indexing. The selection uses source metadata and row indices, not forecasts, future values or audit results. Four models give 64 model–context pairs, with 256 independent streams per outcome law. The lower stream count is fixed across the extension to bound computation; the budget, horizon and evidence rule are unchanged. For each context, we record source hashes, series identifiers, origin indices and context hashes.

The nulls hold conditional on each fixed context and pilot. The extension uses fresh simulation draws at archived contexts.

Table 7: **Sources for native controlled experiments.** Every source has two contexts per model; single-series sources use two origins. Horizons follow the archived sampling grid. Source citations and processing details appear in Appendix B.4.

Source	Domain	Step	Horizon	Series
Weather	Meteorology	10 min	4 h	2
Electricity	Electric load	1 h	24 h	2
Traffic	Road occupancy	1 h	24 h	2
Solar	Solar output	10 min	4 h	2
Exchange	Exchange rates	1 day	24 days	2
ETTh1	Transformer sensors	1 h	24 h	2
US Births	Demography	1 day	24 days	1
Sunspot	Solar activity	1 day	24 days	1

Targets and outcome laws. For each model–context pair, a separate 256-path pilot defines median thresholds for the trajectory mean and the mean absolute first difference. Let x_1 and x_2 indicate that these two features exceed their respective thresholds. Fix $f = x_1x_2$ and $\mathcal{N} = \mathrm{span}(1, x_1, x_2)$. All eight initial pilots and all 64 extension pilots observe all four states; the smallest extension pilot cell has four paths. A context with a missing pilot state would remain unavailable and would not be replaced. The independent pilot fixes the target for subsequent draws.

The null outcome laws are $Q = P$ and

$$Q_h(dy) = \frac{h(y)P(dy)}{\mathbb{E}_P h}, \qquad h(y) = 1 + .75x_1(y) + .5x_2(y).$$

Draw a new native path and accept it with probability $h(y)/2.25$. This rejection sampler requires neither a density evaluation nor the normalizing constant. Since $h \in \mathcal{N}$, projection orthogonality gives $\mathbb{E}_{Q_h} r_P^f = \mathbb{E}_P[hr_P^f]/\mathbb{E}_P h = 0$. Proposition 2 therefore gives the component null.

The alternative is

$$Q_\varepsilon(dy) = \frac{(1 + \varepsilon f(y))P(dy)}{1 + \varepsilon \mathbb{E}_P f}, \qquad \varepsilon = 4.$$

Accept an independent native path with probability $(1 + 4f(y))/5$. Orthogonality of r_P^f to $\Pi_{\mathcal{N}}^P f$ gives

$$\Delta_f(P, Q_\varepsilon) = \frac{\varepsilon \mathbb{E}_P[(r_P^f)^2]}{1 + \varepsilon \mathbb{E}_P f} > 0.$$

Strict positivity holds whenever the population residual energy is positive; the four-state interior design has $f \notin \mathcal{N}$. Pilot coverage does not give a positive lower bound on population cell probabilities. A common ε also permits different component magnitudes across models. We therefore use zero tolerance; a nonzero tolerance would require a certified component magnitude. All rejected proposals count toward physical path use. A rejection cap of eight times the required feedback count is fixed in advance; none of the contexts exhausts it.

Scores and confirmation. Monomial (Section 3.3) uses $M = 8$ and deterministic bound $3/56$. Empirical OLS fits the same target and nuisance features on the forecast batch and returns zero on rank-deficient batches. Its outcome-uniform bound is obtained by enumerating every count vector. The half-count diagnostic uses $\widehat{P}_j = (C_j + .5)/(M + 2)$ and residual $\chi(y)/\{\widehat{P}_y \sum_j \widehat{P}_j^{-1}\}$, bounded by one. Each method is divided by its own deterministic bound. No sample-estimated range or knowledge of the native probabilities is used for normalization.

Each stream has 256 updates, with 512 streams per law in the initial study and 256 in the extension. Forecasts are shared across methods and outcome laws; feedback is independent of the forecasts. The fixed mixture assigns equal initial wealth to both signs and stakes $\{.05, .1, .2, .4, .8, .95\}$, and crosses at 20. The initial null family has $4 \times 2 \times 2 \times 3 = 48$ method–context–law checks. The extension has $4 \times 8 \times 2 \times 2 \times 3 = 384$ planned checks in its own family, including both null laws. An unavailable context would remain in the planned family with excess-test p-value one. We use one-sided exact binomial excess tests against .05 and Holm correction (Holm, 1979). Pointwise exact binomial intervals (Clopper & Pearson, 1934) are used for descriptive rate uncertainty.

Protocols and score implementations were frozen before hardware, interface and support checks; hashes were verified before confirmation. All planned contexts were retained. Model identifiers, context hashes, seeds and path counts were recorded for both studies. The initial study generated 13,279,232 physical paths; the extension generated 53,097,472. Counts include pilots, probes, rejected feedback proposals and unused final batch tails.

B.4 Model Interfaces and Data Processing

The sources are ETTh1, ETTh2, ETTm1, ETTm2, Weather, Electricity, Traffic, Solar, Exchange, London Smart Meters, US Births, Sunspot, Berlin and Tokyo. Metadata hashes select at most eight series per source, giving 80 series in total. Shared contexts have 512 steps and forecast horizons 24. The stride is 24, with at most 272 origins per series. The first 16 are development origins; subsequent windows are used for evaluation. Standard datasets use the last 30% of their histories; the three additional archives and two cities reuse the specified historical windows. There are 19,142 shared

origin windows. Standard benchmark and archive sources follow Zhou et al. (2021); Godahewa et al. (2021); Weather uses the Jena station archive (Max Planck Institute for Biogeochemistry, n.d.).

Four models issue the marginal quantiles $.1, .5, .9$: Chronos-2, TimesFM 3.0, Toto 2.0-2.5B and Moirai 2.0-R-small (Ansari et al., 2025; Google Research, 2026; Khwaja et al., 2026; Salesforce AI Research, 2025). Four supply native joint trajectories: Aurora, Sundial, Moirai 1.0-R-small and Moirai-MoE 1.0-R-small. For the latter, a fixed 256-path batch defines empirical quantile forecasts for the common forecast-quality comparison. These empirical quantiles are finite-sample summaries, distinct from the population quantities in the exact-score theory.

The fixed checkpoint revisions are Chronos-2 `29ec3766`, TimesFM `43046b85`, Toto `51a2812b`, Moirai 2.0 `30f43ff0`, Aurora `c495b02c`, Sundial `3212e425`, and Moirai 1.0 `f5bd5d01`. Moirai-MoE is identified by checkpoint-file hashes. Official package versions and inference settings are captured with each model. No quantile-to-joint-path reconstruction is used.

Related methods recalibrate a joint predictive distribution against a highest-density-region criterion (Chung et al., 2024), monitor PIT values (Farran, 2026), or construct predictive-rank processes (Kuang & Xia, 2026). Their guarantees concern different properties. Comparisons with the present audit require matching the component null and forecast access.

B.5 Computing the Primitive Residual Degree

The rational solver takes a finite nuisance matrix and target vector, checks full column rank, inclusion of constants and exclusion of the target, and expands the determinant residual over nonsingular row subsets. It removes the greatest common polynomial factor and exports the primitive coefficients. Each homogeneous degree-q coefficient can be converted into a multinomial score as in Theorem 5. The solver completed all 16 specified designs, including an eight-state three-bit interaction with $p = q = 7$ and a partially saturated 32-state design with $p = 31, q = 1$. The latter leaves a contrast between two states. These examples demonstrate the construction at finite scale. Subset enumeration, polynomial factorization, coefficient bit length and output size can still limit larger designs.

B.6 Four-State Projection and Budget Comparisons

Use the four predictive laws, state order and binary target from Appendix B.2. The outcome laws here use additive perturbations rather than that study's exponential tilts. Let $r = r_P^f$, $v = r/\|r\|_\infty$, and let h be the centered value of $x_1 + x_2$, divided by its maximum absolute value. Outcome probabilities are $Q_x = P_x(1 + \theta h_x + \eta v_x)$, with $\theta \in \{-.5, 0, .5\}$ and $\eta \in \{-.25, 0, .25\}$. The ratio is at least .25, and the component is $\eta \mathbb{E}_P r^2/\|r\|_\infty$. Thus η is a perturbation parameter, rather than the actual component magnitude. Exact enumeration at $M = 3, 4, 8$ gives 108 rational configurations. The separately reported asymmetric example with $\theta = .5$ has empirical OLS mean $-.01824087$; its Q differs from the original example in Appendix A.2.

The power study uses 4,000 streams, 512 updates and budgets $\{3, 4, 8, 16, 32, 64\}$. The primary comparison is $M = 8$ across the 24 nonzero-η alternatives. Monomial and envelope scores use the same counts, outcomes and fixed-stake mixture. Simultaneous paired intervals use Clopper–Pearson bounds for the two discordant outcomes, with the allocated family error divided across comparisons (Clopper & Pearson, 1934). A partial four-state design separately compares its degree-one primitive score with the degree-three Volume construction; its own residual defines its alternatives.

For the interaction envelope, form simultaneous binomial intervals for the four state probabilities with total failure probability .005 and intersect them with the simplex. The residual is $r_P(y) = \chi(y)/\{P_y \sum_x P_x^{-1}\}$, where $\chi = (1, -1, -1, 1)$. Its sign is fixed at each outcome and its magnitude is in $[0, 1]$. Analytic active-set calculations bound that magnitude over the confidence region; a probability-box bound is the recorded fallback if certification fails. The lower signed residual minus .005 has mean at most the signed component. Division by $1 + \tau + .005$ bounds the tolerance-adjusted score. This is a conservative comparator for the same null, without an exact score–companion identity.

B.7 Reference Laws and Sources of Power Loss

Four native samplers across 14 datasets provide 56 model–dataset cells. Four independent reference banks per cell give empirical four-state probabilities. Mixing with the uniform law at weights .02 and .10 defines positive known predictive laws. Outcome laws use the fixed nuisance perturbation and $\eta \in \{0, .25\}$, yielding 224 configurations, of which 112 are alternatives. Each has 4,000 streams and 512 updates, with $M = 8$ primary and $M = 64$ for budget sensitivity. The four reference contexts cycle through each stream under a prespecified schedule.

The actual average components range from .00125 to .03668; the average companions range from .000074 to .05286. The paired $M = 8$ family compares all 112 alternatives simultaneously. These values refer to constructed empirical mixtures. A state missing from a finite reference bank need not have zero native probability, and historical alerts do not reveal the true component.

To separate forecast noise from normalization, we compare the finite score, its known-P conditional mean given the outcome under the same bound, and the population residual under its own bound. Their average detection rates are $3.48/6.71/15.20\%$ at $M = 8$ and $8.03/8.60/15.20\%$ at $M = 64$, respectively. The conditional mean removes forecast-sampling noise; the population residual also changes the bound. An additional 23 feasible synthetic laws use the same comparison. The two known-P quantities serve as diagnostics, not attainable unknown-P competitors or universal power upper bounds.

B.8 Score and Betting Ablations

The ablation has 271 laws: 247 inherited reference/synthetic laws and 24 additional tolerance configurations. The latter use $\tau = .003$, components $-.003, .003, .008$, four predictive laws and two target families (interaction and partition). There are 136 null laws and 135 alternatives. Budgets are 8 and 64; development uses 1,000 streams and confirmation uses 4,000, with 512 updates. Fixed-stake, predictable-stake, conditional-average and mixture variants share all draws. The development criterion selected the hybrid predictable mixture.

The selected candidate improves detection over the original fixed mixture by .75 percentage points, with simultaneous interval $[-.07, 1.56]$; against the original score with predictable stakes, the difference is .06 points with interval $[-.76, .87]$. A supplementary comparison adds stake .95 to the original fixed grid $\{.05, .1, .2, .4, .8\}$. Its mean gain is .88 points with simultaneous interval $[.10, 1.66]$, averaging the historical, new synthetic and nonzero-tolerance groups equally. This wider grid was selected as a working implementation after the supplementary comparison; it was not the development-selected candidate. This supplementary comparison has a separate error allocation and maximum null crossing rate of 2.60% among valid methods.

The smoothing diagnostic uses $\widehat{P}_x = (C_x + .5)/(M + 2)$ for the interaction and $(S_j + .5)/(C_j + 1)$ for a partition mean. The interaction diagnostic substitutes $\widehat{P}$ into the residual; the partition diagnostic assigns mean $1/2$ to an empty cell. This diagnostic is excluded from valid power rankings and model selection. The comparison changes the projection and smoothing, so it is not an isolated removal of a single algebraic factor. Fixed-batch unsmoothed partition means with an all-cell gate remain an exact positive control by Appendix A.10.

B.9 Betting Baselines

The small comparison uses interaction and partition targets under the asymmetric and correlated predictive laws, with components 0 and .025: eight laws in total. The screen has 300 streams per law; confirmation has 1,000 new-seed streams. Budgets 8 and 64 share prefixes, and the primary horizon is 2,048, with 512 a prespecified secondary horizon. Every method receives the same bounded exact score, with zero tolerance and boundary 20. The screen-selected official comparator is aKelly with the author's default .5 truncation.

PrPl, aKelly and LBOW are imported from `confseq` at commit `5ffe733c`. ONS uses the original function from `betting-paper-simulations` at commit `cb05eef4`, with its function body unchanged. The cited methods are established bounded-mean betting procedures (Waudby-Smith & Ramdas, 2024). Original LBOW can return zero wealth after a zero-residual prefix because of a numerical degeneracy; those results are retained with a flag and do not support a superiority claim.

The primary paired interval averages budgets within each replicate and then the four alternatives equally. It does not establish equivalence or noninferiority.

B.10 Controlled Acquisition Comparisons

The added acquisition study uses chunk size 32, cap 256 and quota one under balanced, moderately rare and rare cells. It includes 18 regular alternatives. OS is compared with fixed-256, a fixed budget rounded from its analytic expected sample count, and a fixed budget estimated from independent development draws. The ideal matcher knows predictive cell probabilities; the practical matcher incurs development cost. Average OS sample counts are 32.00, 33.14 and 107.51 in the three regimes. The practical match has maximum confirmation cost discrepancy 1.86%. Simultaneous paired intervals distinguish the ideal-budget and full-cap contrasts. The recorded comparison to the practical matcher remains descriptive under its reported cost mismatch.

B.11 Targets, Error Allocation and Acquisition Costs

The joint-path study has a historical partition T0, an independent 64-path-pilot partition T1, and a binary interaction T2. Each partition has fixed-256, quota-one OS with chunk 32 and cap 256, the valid envelope, and a development-based mean-cost matcher. T2 has an exact score and an envelope. Ten planned processes per model–series give 3,200 formal processes and threshold 128,000 under total error allocation .025. Different partitions define different components. The corresponding marginal audit allocates .025 over 1,920 processes, with threshold 76,800. Its two-sided quantile inequalities use strict and weak comparisons to retain ties. Grouping uses only past inputs.

The acquisition study uses 26,893,112 generated paths, 2,240 development acquisitions and 14,592 evaluation acquisitions. Costs include pilots, failed acquisitions and timing samples. Timings pair methods at the same context and aggregate repeated measurements before sequence or time-block resampling. GPU devices are shared, so the numbers describe these implementations and conditions rather than hardware-independent speed factors (Kayaalp et al., 2026).

C Additional Results

C.1 Score Choice, Effect Size and Outcome Budget

Table 8 complements detection rates with restricted mean crossing times. A value near 2,048 reflects late or absent crossings. Both summaries favor Pooled on the rare-state law and Monomial on the uniform law; no score is selected for individual laws.

Table 8: **Restricted mean crossing time and noncrossing rate.** Each time is $\mathbb{E}[\min(T_{\text{cross}}, 2048)]$ estimated over all 4,000 streams. Noncrossing percentages are reported separately. These are the same primary comparisons as Table 4.

		Restricted mean time		Noncrossing (%)	
Law	M	Monomial	Pooled	Monomial	Pooled
Uniform	8	1378.4	1394.1	31.50	33.60
Uniform	32	1003.3	1022.5	8.30	9.30
Asymmetric	8	1394.2	1404.4	34.85	33.88
Asymmetric	32	898.0	932.9	6.33	7.12
Correlated	8	1355.9	1366.4	32.23	32.40
Correlated	32	665.2	562.2	0.38	0.40
Rare states	8	2000.8	1844.5	87.48	64.72
Rare states	32	1858.9	730.2	47.10	0.25

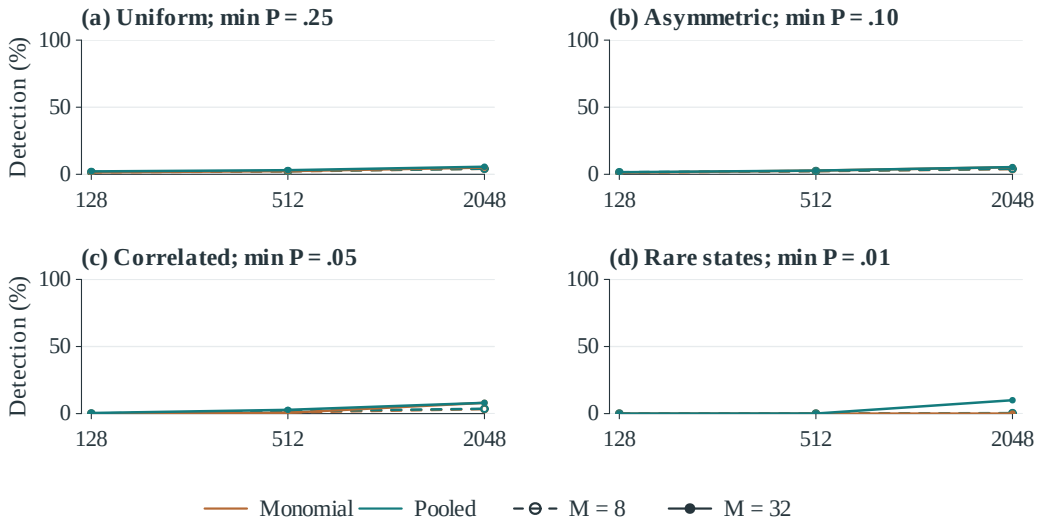


Figure 7: **Detection for a smaller component.** The same four laws, paired methods and budgets as Figure 3, with component $-.005$ and tolerance zero. Each endpoint uses all 4,000 confirmation streams per law. Color identifies the score and line type the forecast budget, as in the main figure.

The smaller component in Figure 7 reduces detection at the same budgets and horizons. Exact normalized means and variances for the strong component appear in Figure 4. At zero tolerance, the two Monomial bounds coincide. For nonzero tolerance, tightening the bound changes the range but has little effect on detection in this comparison. The maximum null crossing rate is 3.70%, with no Holm-significant excess among 60 tests. This checks the implementation; the score identity supplies the validity guarantee. Section 5.4 reports the separate native count replay, where Pooled has no average detection gain.

C.2 Native Controlled Results

The initial study and eight-source extension use separate confirmation families, as specified in Appendix B.3.

Eight-source extension. Table 9 reports the largest null crossing rate for each model and source, and the detection range across both contexts. Asterisks mark excesses resolved in the planned family. Each of the 64 model–context pairs has pointwise exact binomial intervals and score means with Monte Carlo standard errors. Component magnitudes differ across contexts, so these detection ranges do not rank models at a common effect size.

Table 9: **Native controls by model and source.** The three null columns report the maximum crossing percentage across two contexts and two null laws (four cells per entry). The last column gives the Monomial detection range across the two target-tilt contexts. $*$ marks an excess significant in the full 384-check Holm family. All 64 model–context pairs contribute, including low-power cases.

		Maximum null crossing (%)			Monomial detection
Model	Source	Monomial	OLS	Half-count	range (%)
Aurora	Weather	2.34	7.81	6.64	96.88–99.61
Aurora	Electricity	2.34	5.08	5.47	96.48–97.27
Aurora	Traffic	2.73	6.64	5.08	96.88–97.66
Aurora	Solar	1.95	7.42	10.55	87.11–99.22
Aurora	Exchange	1.56	1.95	1.17	99.61–99.61
Aurora	ETTh1	2.73	6.25	3.52	98.44–99.61
Aurora	US Births	1.56	8.98	10.94*	76.56–89.84
Aurora	Sunspot	3.12	10.55	3.52	93.36–98.44
Sundial	Weather	2.34	5.08	86.33*	0.00–90.23
Sundial	Electricity	2.34	5.86	22.27*	48.44–100.00
Sundial	Traffic	1.56	8.20	17.58*	65.23–87.11
Sundial	Solar	3.91	8.59	5.08	99.22–100.00
Sundial	Exchange	3.52	4.69	1.95	97.66–99.61
Sundial	ETTh1	2.73	3.91	3.12	98.83–99.61
Sundial	US Births	3.52	1.95	3.12	99.61–100.00
Sundial	Sunspot	1.95	4.69	44.53*	21.88–98.44
Moirai	Weather	1.17	1.56	94.53*	0.00–100.00
Moirai	Electricity	3.12	7.03	2.73	96.88–100.00
Moirai	Traffic	2.73	5.08	3.52	98.83–98.83
Moirai	Solar	3.52	6.64	4.69	94.14–97.66
Moirai	Exchange	2.34	1.17	1.95	100.00–100.00
Moirai	ETTh1	1.95	4.30	3.12	98.83–99.22
Moirai	US Births	2.34	4.30	2.34	99.22–100.00
Moirai	Sunspot	1.95	8.20	14.84*	64.84–98.05
Moirai-MoE	Weather	2.34	4.69	54.30*	12.11–98.44
Moirai-MoE	Electricity	1.95	3.91	3.12	99.22–100.00
Moirai-MoE	Traffic	1.95	7.81	14.84*	71.88–94.53
Moirai-MoE	Solar	1.95	9.38	14.45*	84.77–100.00
Moirai-MoE	Exchange	2.34	1.56	1.95	100.00–100.00
Moirai-MoE	ETTh1	2.73	2.34	3.52	99.61–100.00
Moirai-MoE	US Births	5.08	3.12	3.12	99.61–99.61
Moirai-MoE	Sunspot	1.17	2.73	66.02*	0.78–100.00

Initial study. All Monomial null means are within 1.85 Monte Carlo standard errors of zero. The maximum absolute empirical OLS mean is .02142; its maximum null crossing rate, 6.05%, is not significant in the planned family. Two half-count null cells have significant excesses. Monomial alternative crossing rates range from 10.74% (Sundial–Weather) to 100%; those alternatives have different unknown magnitudes. The half-count diagnostic detects all alternatives but fails its null

control in two cells, so its alternative rate is not a valid power advantage. The 72 method–law comparisons include score means, Monte Carlo standard errors and exact binomial crossing intervals.

A fixed diagnostic replication. The largest standardized Monomial mean in the extension is 3.85 Monte Carlo standard errors (Sundial, Solar, context 1, nuisance tilt), with mean .00030268. A diagnostic replication uses one prespecified new seed, 1,024 streams of 256 updates and the original pilot thresholds at this context. The repeated Monomial means are $-.00002898$ under $Q = P$ and $-.00004381$ under the nuisance tilt; stream-based Monte Carlo standard errors are .00003718 and .00003992. Crossing rates are 1.37% and 2.05%. The original mean displacement is not reproduced in this diagnostic. Because the context was selected after inspection, this replication remains separate from the original 384-check family. It uses 2,737,152 fresh paths and does not replace the original result.

C.3 Forecast Quality and Acquisition Costs

Table 10 reports the quality of the eight forecast interfaces on the shared historical windows. The model and preprocessing details are in Appendix B.4.

Table 10: **Forecast quality on shared historical windows.** All models complete all 14 datasets. Scaled pinball and MASE are averaged over series, datasets and eight domain families in succession; ETT, weather and electricity variants share their respective family weight. Coverage refers to the nominal 80% marginal interval. The table measures the forecast interfaces before auditing.

Model	Evaluated interface	Scaled pinball	MASE	Coverage (%)
Chronos-2	Issued quantiles	.2673	.8058	79.46
TimesFM 3.0	Issued quantiles	.2854	.8486	81.78
Toto 2.0-2.5B	Issued quantiles	.2848	.8494	81.66
Moirai 2.0-R-small	Issued quantiles	.2884	.8618	80.00
Aurora	256 native paths	.3997	1.0688	49.66
Sundial	256 native paths	.3248	.8734	42.76
Moirai 1.0-R-small	256 native paths	.9632	3.0643	74.41
Moirai-MoE 1.0-R-small	256 native paths	.3552	1.0587	80.55

For T0, the development cost match drifts by more than 10% in 222 of 320 series comparisons, with model-specific median absolute deviations of 13.98–24.52%. Thus positive evidence differences against that matcher are descriptive under an imperfect cost match. For T1, every main acquisition stops after 32 inference paths. Its saved prefixes and wealth are identical to fixed-32, including the common 64-path pilot. Including the pilot, this saves 70% of paths relative to fixed-256. Because fixed-32 gives the same scores, the saving is attributable to the smaller batch; the comparison provides no evidence of an adaptive gain.

C.4 High-Budget Comparison with the Projection Envelope

For states $b_j = j/4$, $j = 0, \dots, 4$, take $\phi = (1, b, b^2)$ and $f = b^3$. With three-state subsets I, put $d_I = \det(\Phi_I)^2$ and $r_I(y) = \prod_{j \in I}(b_y - b_j)$. Cauchy–Binet gives

$$r_P(y) = \frac{\sum_I d_I \prod_{j \in I} P_j \, r_I(y)}{\sum_I d_I \prod_{j \in I} P_j}.$$

This weighted average has absolute range bound $R = 3/8$ and maximum interpolation-residual spread $D = 3/8$. Simultaneous Clopper–Pearson intervals with failure $\delta/5$ per state, $\delta = .01$, define a rectangular confidence set. Its minimum residual $L_C(y)$ is found at one of 32 corners: fixing other coordinates gives a linear-fractional function with endpoint extrema. If a corner has zero denominator, use the global lower envelope $\min_I r_I(y)$. The box can include points outside the simplex, which only enlarges the uncertainty set.

On coverage, $L_C(y) \le r_P(y)$; off coverage, the excess is at most D. Hence the bounded corrected score satisfies

$$\mathbb{E}_{C,Y} \frac{L_C(Y) - D\delta}{R + D\delta} \le \frac{\Delta_f(P, Q)}{R + D\delta}.$$

It has nonpositive conditional mean under the component null. The per-update correction accounts for confidence failure in expectation; stakes are chosen before the current counts.

The protocol fixes $P = (1, 2, 3, 4, 5)/15$ and three outcome laws: $Q_j = P_j\{1 + .2(b_j - \mathbb{E}_P b)\}$, $Q = (1/12, 1/15, .3, .2, .35)$, and $Q_j = P_j\{1 + .75 r_P(b_j)/\|r_P\|_\infty\}$. The first two components are zero; the third is 0.010546875. For each budget $M \in \{3, 31, 127, 511, 2047\}$ in Table 11, 128 streams use 4,096 outcomes, with additional prespecified endpoints 512 and 2,048. Forecast counts and uniform outcome random numbers are shared across methods and laws. The eight positive bets are .005, .01, .025, .05, .1, .2, .4 and .8; their e-processes form an equal mixture with crossing boundary 20 at level .05 per method.

The exact count numerator is $\sum_I d_I \prod_{j \in I} C_j r_I(y)/(M)_3$. Its prespecified normalizer bounds both this numerator and the forecast-side companion. For $M \le 32$ we enumerate all count vectors. For larger budgets we bound the corresponding cubic on the simplex using its degree-32 Bernstein coefficients, then multiply by $M^2/\{(M-1)(M-2)\}$ to recover the count normalization. Bounds are inflated by 10^{-11} for numerical roundoff. This normalizer is conservative at zero tolerance. The independent numerical checks cover 135 rows, enumerating count laws at budgets 3, 7 and 12 under 13 predictive distributions; the largest identity error is 2.98×10^{-17}.

Table 11: **Same-null comparison on the five-state target.** Alternative crossing percentages at $T = 4{,}096$, with median crossing times among detected streams in parentheses; 128 streams per configuration. Exact 95% binomial intervals are [0,2.84]% for 0/128 and [97.16,100]% for 128/128. Both finite-forecast methods have 0/128 crossings in all ten null configurations. "–" means no crossing.

	Unknown P, shared forecast budget		Known P
M	Volume score	Projection envelope	Oracle
3	0 (–)	0 (–)	100 (36)
31	100 (1,755.5)	0 (–)	100 (38)
127	100 (1,793)	0 (–)	100 (44)
511	100 (1,766)	0 (–)	100 (35.5)
2,047	100 (1,771)	100 (978.5)	100 (44)

C.5 OS Acquisition Costs

Table 12: **Fewer paths do not imply faster native acquisition.** Four native models across all 14 datasets. T0 is the declared historical partition; T1 uses an independent 64-path pilot to define another partition. OS uses chunks of 32, quota one and cap 256. Time is relative to batched fixed-256; values above one are slower. Aggregation gives equal weight to series, datasets and domain families in succession.

	T0 partition			T1 partition
Model	Paths saved (%)	Quota success (%)	Time ratio	Time ratio
Aurora	54.77	72.82	4.075	1.090
Sundial	35.99	46.12	5.309	.994
Moirai	62.59	75.21	2.181	1.027
Moirai-MoE	47.71	61.13	4.041	.941

Earlier four-cell study. The earlier study uses 1,000 streams, 128 updates, five caps $\{8, 16, 32, 64, 96\}$ and threshold 20. Predictive cell masses are $(.55, .25, .15, .05)$, or $(.80, .10, .08, .02)$ for the rare-cell case. Normal target draws are clipped to $[0, 1]$; their analytic

expectations, feedback shifts and shared-prefix seeds are fixed by the protocol. No component audit crosses under either equality or a pure cell-frequency shift. At cap 96, fixed batch and both OS quotas detect all strong within-cell and rare-cell alternatives. In weak cancellation, 58.3% fewer paths accompanies detection falling from 50.5% to 46.0%. The 225-row comparison includes all caps, failure costs and the 31-path pilot, including 100 rows using the distinct outcome-matching access model of Appendix A.10.

An exploratory four-origin Moirai-MoE runtime probe records .3085, 13.9135 and 2.3882 seconds per origin for batched fixed-96, sequential fixed-96 and quota-two sequential OS, respectively. The larger 14-source measurements in Table 12 supersede this small probe for assessing interface cost.

C.6 Two-Cell Comparison at Comparable Costs

Two cells with low/high target values give predictive laws $(.56, .14, .06, .24)$ and $(.72, .18, .02, .08)$ and within-cell means $(.2, .8)$. The outcome laws separately vary cell frequency and target means. Table 13 reports all eight alternative designs at the same cap and at approximately matched expected path costs. Each law uses 2,000 streams and 512 updates. Component-null crossing rates range from 1.35% to 3.60%. The 80 method–law comparisons report path costs and pointwise Wilson intervals (Wilson, 1927). Per-stream paired crossings were not retained, so these marginal counts cannot support paired intervals.

Table 13: **OS and fixed batches at comparable acquisition costs.** Two-cell experiment: p_0 is predictive cell mass; r is quota. All eight designs are shown. M_{near} is the analytic expected OS cost rounded to an integer. Detection is at $\Delta_f = .10$, $T = 512$, $\alpha = .05$, with 2,000 paired streams per design. The target and cells are fixed; no pilot is used.

			Paths per origin		Detection (%)		
p_0	Cap	r	OS mean	M_{near}	OS	Fixed near	Fixed cap
0.7	8	1	3.57	4	94.60	92.00	98.45
0.7	8	2	6.06	6	95.10	97.30	98.45
0.7	24	1	3.77	4	95.75	92.10	99.80
0.7	24	2	7.10	7	99.00	98.50	99.80
0.9	8	1	5.80	6	78.05	76.25	85.75
0.9	8	2	7.61	8	39.55	85.75	85.75
0.9	24	1	9.32	9	95.45	88.70	98.75
0.9	24	2	16.31	16	95.45	97.05	98.75

C.7 Detection at the Minimum Exact Budget

Table 14: **The exact-budget threshold does not determine power.** Detection percentages for known alternatives. The asymmetric design uses 4,000 streams with 512 updates. The reference-law study has 112 alternatives, each with 4,000 streams and 512 updates. Rows have different effect sizes and should be compared across budgets within each row.

Design	$M = 3$	$M = 8$	$M = 32$	$M = 64$
Asymmetric binary interaction	2.48	15.68	33.65	–
112 model-derived reference alternatives	–	3.48	–	8.03

The four-state binary experiment uses uniform P, the same all-high target and nuisance basis, and $q = 3$. The null is $Q_0 = (.15, .25, .25, .35)$; the alternative is $Q_1 = (.25, .15, .15, .45)$ with $\Delta_f = .10$. The Gram determinant is $1/16$, so the Volume score has alternative mean .00625 at both $M = 3$ and $M = 8$. Each uses 2,000 streams of 512 origins, seed 20260925, and an equal mixture over both signs and positive stakes $\{.1, .25, .5, .75\}$. Normalization is the largest absolute realized score over every count vector and outcome state at the specified budget. At $M = 3$ and $M = 8$, detection is 64.00% and 100.00%, null crossing is 2.30% and 3.20%, and zero-score frequency is 90.65% and 28.89%. The empirical mean scores are .006294 and .006237. Their score bounds are

$1/6$ and $3/56$, respectively; the comparison changes normalization as well as the score distribution. Zero-score fractions use 1,024,000 origins under the alternative; crossing fractions use 2,000 streams.

A separate rational audit checks four Volume identities, two stopped-label identities and the primitive-degree examples $1, 3, 3, 7$. In the four-state design at $M = 2$, the exact-score constraints form a rank-40 system in 40 symmetrized score variables, leaving no nonzero exact mean coefficient. This gives an exact instance of Theorem 5.

D Historical Audits and Earlier Variants

To decompose a historical discrepancy, the raw and residual scores must share a mean coefficient. We construct such scores and simultaneous intervals for the resulting weighted historical components.

D.1 Scores for Weighted Historical Components

Let $f \in [0,1]$. For a K-cell partition and $M > K$, let $b_{\max} = \max_{\sum_j n_j = M} \prod_j n_j$ and use $B = \prod_j C_j / b_{\max} \in [0,1]$. When every cell is observed,

$$U = B\{f(Y) - \widehat{\mu}_{g(Y)}\}, \qquad \widetilde{m} = \frac{\sum_j (C_j - 1)\widehat{\mu}_j}{M - K}, \qquad W = B\{f(Y) - \widetilde{m}\}.$$

All three statistics vanish if a cell is unobserved. Product-count tilting shifts the multinomial count law by one observation per cell, yielding

$$\mathbb{E}U = c\Delta_f, \qquad \mathbb{E}W = cD, \qquad \mathbb{E}B = c \in [0,1], \qquad |U|, |W| \le B,$$

where $D = \mathbb{E}_Q f - \mathbb{E}_P f$. The fitted score is $W - U$. Using B times the ordinary sample mean would not give this common-weight decomposition.

For OS, multiplying the ordinary sample mean by the quota-success indicator generally changes its expectation. An additional independent forecast X_0 would give $W_{\mathrm{OS}} = A_r\{f(Y) - f(X_0)\}$ with the required coefficient, at the cost of one extra path. The historical study uses the fixed-budget product-count construction above.

At update t, average scores over n_t origins maturing together. Write $a_t = n_t^{-1} \sum_i c_{ti}\Delta_{ti}$, $d_t = n_t^{-1} \sum_i c_{ti} D_{ti}$ and $c_t = n_t^{-1} \sum_i c_{ti}$. Their cumulative predictable means are $A_T = \sum_t a_t$, $D_T = \sum_t d_t$ and $C_T = \sum_t c_t$. Dependence among origins in a group is allowed; conditional linearity and bounded group means suffice, under the declared sampling filtration. Conditional mean identities remain assumptions even when updates are nonoverlapping. A failed pilot is retained with zero scores and $c_{ti} = 0$.

For $C_T > 0$, the residual component is A_T/C_T, the raw discrepancy is D_T/C_T and the fitted component is $(D_T - A_T)/C_T$. These are equal-maturity, coefficient-weighted historical quantities. They are not the unweighted mean over all cached origins or a statement about the current origin. Ratios of the observed sums alone are descriptive estimates.

D.2 Bounded Confidence Sequences

For bounded adapted $X_t \in [l,u]$, put $Z_t = (X_t - l)/(u - l)$ and choose a predictable center $\widehat{Z}_t \in [0,1]$. With $\psi(\lambda) = -\log(1-\lambda) - \lambda$, $0 < \lambda < 1$, the inequality

$$\log(1 + \lambda z) \ge \lambda z - \psi(\lambda) z^2, \qquad |z| \le 1,$$

implies an exponential supermartingale for deviations of the sum from its conditional means. To see this, set $m_t = \mathbb{E}[Z_t \mid \mathcal{F}_{t-1}]$. The conditional expectation of

$$\exp\{\lambda(Z_t - m_t) - \psi(\lambda)(Z_t - \widehat{Z}_t)^2\}$$

is at most $\exp\{-\lambda(m_t - \widehat{Z}_t)\}\{1 + \lambda(m_t - \widehat{Z}_t)\} \le 1$. Apply the same argument to $1 - Z_t$ for the other tail. For one tail and one fixed λ, Ville's inequality gives the time-uniform radius

$$R_{X,T}(\lambda, \delta) = \frac{u - l}{\lambda} \left[\log(1/\delta) + \psi(\lambda) \sum_{t \le T} (Z_t - \widehat{Z}_t)^2 \right]. \tag{D.1}$$

Intersect sum intervals with $[Tl, Tu]$. A finite union bound over a fixed grid allows the smallest of these jointly valid radii to be used. This bounded-martingale construction follows standard confidence-sequence arguments; see Howard et al. (2021) for broader methods.

The implementation uses the previous observed mean as center, with midpoint at the first update, and grid $\{.01, .02, .05, .1, .2, .4, .6, .8\}$. It splits .05 over 12 traces, three scores (U, W, B), both tails and all eight grid points. For U, W, use $[l, u] = [-1, 1]$; for B, use $[0, 1]$. Dependence across traces is harmless to this union bound. All updates and final intervals use their own phase's past, not later feedback or other phases' outcomes.

On the simultaneous event, divide numerator and denominator intervals using $C_T > 0$ and the population range $[-1, 1]$. If the denominator lower bound is zero, use a one-sided bound when the numerator sign is established and otherwise retain the full range. The fitted numerator interval is $[D_L - A_U, D_U - A_L]$. Intervals at different times concern different averages and are not intersected as if the parameter were constant.

The fixed-protocol guarantee does not correct for choosing this reporting analysis after examining earlier research results. Its nominal .05 family is separate from the earlier 48-test family, and the two are not combined into a single .05 claim.

D.3 Persistence Targets in Pedestrian and Wind Data

Let θ_h and IQR_h be the pilot's .80 quantile and interquartile range at coordinate h. Define

$$b_h = \max\{.25\,\mathrm{IQR}_h, .001\}, \qquad a_h(x) = [1 + \exp\{-(x_h - \theta_h)/b_h\}]^{-1}.$$

The bounded persistence target is

$$f(x) = \frac{\exp[24^{-1}\sum_h \log\{.05 + .95a_h(x)\}] - .05}{.95}.$$

It is high when the soft-high values a_h remain high across the 24-step trajectory; it is a score, not an event probability.

The study reuses forecast paths for Chronos-T5-base and Moirai-MoE. After development on eight Solar data subsets per model, the target, pilot split, partitions and testing family were fixed before evaluation on either validation domain. The retrospective study includes every archived origin for Melbourne pedestrian counts and wind-speed trajectories (Godahewa et al., 2021; Ding, 2021).

Each origin has 1,023 forecast paths and one matured 24-step outcome. The first 255 paths define these quantities and a partition based on two features: $24^{-1}\sum_h a_h(x)$ and $23^{-1}\sum_{h=1}^{23} a_h(x)a_{h+1}(x)$. A binary tree of depth two or four gives $K = 4$ or $K = 16$ cells. At a node with n pilot paths, each feature proposes a split halfway between its two central distinct values. Each child must contain at least $\max\{8, \lfloor .1n \rfloor\}$ paths. Among admissible splits, choose the smallest child-count imbalance, breaking ties by feature order. An unavailable split makes the partition unavailable. The $K = 16$ partition nests the $K = 4$ partition. The remaining 768 paths supply the product-count score and companion. Failed designs and batches with an unobserved cell contribute zero scores and remain in the denominator. For each pedestrian system, the permitted feedback protocol gives daily phases with 129,001 origins/3,996 updates and 74 origins/74 updates; wind has 169,920 origins/708 updates. Every within-phase maturity gap is at least 24 hours. Group-mean scores give each maturity update equal weight. Separate phases use their own histories and evidence processes under the conditional sampling assumptions in Appendix A.11.

The original component-testing family has 48 tests: both signs, $K \in \{4, 16\}$ and $\tau \in \{0, .02\}$ in the six system/domain/phase combinations, with boundary $\log 960$. The interval family below has a separate nominal error allocation.

D.4 Simultaneous Historical Intervals

The 12 traces represent 597,990 system–origin pairs. Models and partitions reuse outcomes, and both partitions and every phase remain in the interval family. Update counts follow the maturity grouping above.

The intervals use the original targets, nuisance spaces and scores. Their nominal coverage has the post-selection limitation stated above: the reporting analysis was chosen after earlier results were available.

Table 15: **All historical component intervals.** Nominal simultaneous 95% intervals under the fixed retrospective protocol. Ped. M and Ped. S denote the large and short pedestrian phases. K is the number of cells. The intervals concern each trace's coefficient-weighted historical means.

			Historical mean: nominal simultaneous 95% interval		
Model	Phase	K	Raw	Fitted	Residual
Chronos	Ped. M	4	$[.0127, .0269]$	$[.0106, .0374]$	$[-.0105, .0021]$
Chronos	Ped. S	4	$[-.4019, .3923]$	$[-.7945, .7858]$	$[-.3935, .3926]$
Chronos	Ped. M	16	$[.0083, .0312]$	$[.0103, .0548]$	$[-.0237, -.0020]$
Chronos	Ped. S	16	$[-.8197, .8110]$	$[-1.0000, 1.0000]$	$[-.8111, .8117]$
Chronos	Wind	4	$[.0181, .1317]$	$[-.0239, .1715]$	$[-.0398, .0430]$
Chronos	Wind	16	$[.0002, .1523]$	$[-.0397, .2422]$	$[-.0900, .0399]$
Moirai-MoE	Ped. M	4	$[.0141, .0279]$	$[-.0151, .0113]$	$[.0165, .0293]$
Moirai-MoE	Ped. S	4	$[-.3624, .4345]$	$[-.7733, .8150]$	$[-.3804, .4110]$
Moirai-MoE	Ped. M	16	$[.0098, .0324]$	$[-.0210, .0233]$	$[.0090, .0309]$
Moirai-MoE	Ped. S	16	$[-.8170, .9244]$	$[-1.0000, 1.0000]$	$[-.8454, .8885]$
Moirai-MoE	Wind	4	$[.0038, .0987]$	$[-.0932, .0856]$	$[.0126, .0971]$
Moirai-MoE	Wind	16	$[-.0175, .1218]$	$[-.1336, .1366]$	$[-.0149, .1160]$

Residual intervals exclude zero in four traces: the large Chronos pedestrian phase with 16 cells, the large Moirai-MoE pedestrian phases with 4 and 16 cells, and Moirai-MoE wind with 4 cells. All four also have raw intervals excluding zero, as do three further traces. For the 16-cell Chronos pedestrian trace, the fitted and residual intervals have opposite signs, indicating cancellation in the weighted historical discrepancy. All four 74-update traces remain unresolved.

D.5 Earlier Variants and Negative Results

The earlier quantile archive contains 1,377 channels from nine datasets at horizons 24, 96, 192 and 336; its denominators remain separate from the 80-series study. A continuous-nuisance study compares an exact score with a localized conservative score. A selected quadratic-target case improves from 20.25% to 94.45% at $M = 2{,}048, T = 512$, but the broader path-maximum comparison gives 12 Holm-adjusted wins and 25 losses across 48 alternatives. The localized score has nonpositive null mean without the exact score–companion identity. Across 390 alternatives, predictable gating loses 3.845 percentage points to always using the localized score and 1.797 points to an equal-wealth mixture. It is excluded from PathTail.

The Berlin/Tokyo Moirai audit has six crossings in eight component streams and five for a rank diagnostic under a different null, so it does not establish a power ordering. It uses 1,218 outcome windows and 4,988,928 paths. Weather outcomes include reanalysis/model-derived values (Open-Meteo, 2026). The fixed-protocol guarantees exclude research selection; pretraining overlap and intervention benefits remain unresolved (Rossellini et al., 2025).