UBOA: MATCHING SELECTED FORECASTING CLAIMS TO THEIR VALIDITY SCOPE

Anonymous authors
Paper under double-blind review

ABSTRACT

A validation-selected forecasting rule may reject a 5% improvement null without supporting a guarantee conditional on its selection history. UBOA makes this distinction explicit through an executable specification for matching statistical claims to evidence. Each request binds the selected rule, comparator, loss, evaluation law, error event, conditioning information, and coverage unit. Eight rules either license the unchanged conclusion or retain an explicit mismatch alongside separately scoped evidence. Relative soundness requires correct annotations and external premises, which the checker does not authenticate. In a 140-case contract challenge fixed before access to the checker implementation, the revised matcher made no false accepts or false rejects; coordinate equality produced 29/70 false accepts and 10/70 false rejects. We also instantiate existing concentration tools in two synthetic settings, preserving raw mean absolute error (MAE) or mean squared error (MSE), overlapping horizons, and the selected rule while supplying finite-sample history-conditional local control. A preregistered audit retained 80 histories and 327,680 continuations, with one false promotion, zero recorded first-true-null inclusion violations, and no Holm-adjusted evidence of excess error. Supplementary coefficients and counts support deterministic reconstruction of the certificates and aggregate audit quantities. Earlier simulations and six real applications retain their historical interpretations; the real applications do not have these conditional certificates.

1 INTRODUCTION

When does evidence about a forecasting result support the conclusion attached to it? A validation-selected rule may reject a later 5% improvement null, yet the resulting report can mean different things: an observed decision, a fixed-rule marginal result, or a claim conditional on the full selection history. The selected rule, comparator, loss, aggregation, evaluation law, and probability scope determine which interpretation is justified. Serial dependence and selection make these distinctions consequential.

Forecast-comparison tests address specified forecast losses (Diebold & Mariano, 1995; Giacomini & White, 2006); Reality Check and superior predictive ability (SPA) tests address candidate-pool questions under data snooping (White, 2000; Hansen, 2005); selective inference studies selected targets under explicit selection constructions (Berk et al., 2013; Taylor & Tibshirani, 2015; Lee et al., 2016; Fithian et al., 2014); and fixed-sequence methods control ordered families once suitable local tests are available (Westfall & Krishen, 2001; Wiens, 2003; Goeman & Solari, 2010). UBOA addresses the remaining reporting question: does the available evidence license this requested conclusion about this selected rule?

Core idea. Keep the request fixed and make every permitted evidence-to-conclusion step explicit. Eight rules compare the scientific object and probability scope, check their declared premises, and either license the conclusion or preserve an unsupported request alongside separately scoped evidence. This is our primary contribution: an executable claim/evidence specification with relative soundness and chronological non-implication examples. The checker verifies the supplied record, not the truth of its external annotations.

The specification also identifies how to supply full-history conditional control. Given the complete pre-evaluation history, an explicit conditional law, and valid numerical structural bounds, we instantiate

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.

local certificates for two contractive synthetic settings. Their guarantees compose by fixed sequence without changing the selected-realization target. An implementation-blind 140-case challenge evaluates the rule logic; a fresh preregistered audit evaluates the certificate implementation; historical simulations and six real applications illustrate reporting behavior and its limits. The real applications lack the required structural certificates and retain their historical interpretations. Anonymous reproducibility materials are available at https://anonymous.4open.science/r/UBOA-Repository-0113/.

2 Related Work

Inferential objects. Table 1 separates pool-level testing (White, 2000; Hansen, 2005), inference after explicit selection (Leeb & Pötscher, 2005; Berk et al., 2013; Lee et al., 2016; Fithian et al., 2014), independent-calibration risk testing (Angelopoulos et al., 2025), and composition of valid ordered local tests (Holm, 1979; Wiens, 2003; Goeman & Solari, 2010). UBOA checks whether their targets, laws, error events, conditioning and coverage match a requested conclusion.

Table 1: Adjacent methods answer different inferential questions. Compatibility requires a matched selected-realization claim and explicit validity premises.

Method family	Primary question and required inputs	Relation to UBOA
Reality Check / SPA	Does any member of a searched pool beat a benchmark under the pool null? **Required inputs:** candidate pool, benchmark, pool-null target, and dependence assumptions.	**Adjacent pool-level estimand.** A pool rejection is not recoded as the selected rule's terminal threshold.
Selective inference	What follows for a selected target under an explicit selection event or simultaneous construction? **Required inputs:** selected target, selection event or simultaneous construction, law, and conditioning information.	**Compatible only under exact target matching.** Target, selection information, law, and conditioning event must match the requested claim.
Learn then Test	Which configurations meet a risk bound under an independent calibration construction and familywise error-control procedure? **Required inputs:** independent calibration data, risk bound, and the familywise procedure.	**Independent framework.** Its guarantee is not transferred to dependent forecasting data here.
Fixed sequence	Which ordered claims reject when all local tests satisfy their premises? **Required inputs:** an ordered claim family and valid local tests.	**Used inside UBOA.** It provides classical composition, but no extra selection or law premise.
Structured evaluation	What context, provenance, quality practices, or comparability information accompanies a reported evaluation? **Required inputs:** provenance, quality-practice records, and comparability metadata.	**Complementary records.** UBOA checks a specific statistical implication; no comparative utility result is claimed.
UBOA	Is the requested claim about this selected realization licensed by the declared evidence and validity scope? **Required inputs:** selected realization, requested claim, declared evidence, and validity scope.	**Claim–evidence binding and terminal reporting policy.** UBOA does not replace pool, selective, or calibration procedures.

Evaluation records. BetterBench, FactSheets and Eval Factsheets describe evaluation practices and context (Reuel et al., 2024; Arnold et al., 2019; Bordes et al., 2025); Evaluation Cards adds implemented interpretive signals and extraction infrastructure (Ghosh et al., 2026). They supply

complementary records, whereas UBOA checks a statistical implication. Appendix L.5 compares their outputs and discusses the broader forecast-evaluation literature.

Conditional inference and concentration. Conditional post-selection prediction (Leeb, 2009) and bootstrap bands for expected forecast scores and skill (Pohle et al., 2026) address related but distinct objects. Our bounded certificate specializes established contractive-functional concentration (Dedecker & Fan, 2015; Dedecker et al., 2019); the Gaussian certificate uses quadratic moments and Cantelli's inequality. The explicit overlap coefficients preserve the selected forecasting comparison, rather than introducing a new general inequality.

3 Claim and Validity Scope

Let H denote the complete permitted pre-evaluation record and let $\mathcal{F} = \sigma(H)$. A selection rule S is $\mathcal{F}$-measurable and returns $s = S(H)$. The scientific target fixes comparator b, loss ℓ, index set I, nonnegative weights w summing to one, ordered thresholds $r_1 < \cdots < r_K$, and an evaluation-law kernel Q_H. Define

$$L_a(Q_H) = \mathbb{E}_{Q_H}\left[\sum_{i\in I} w_i \ell(a, Y_i)\right], \qquad \theta_{Q_H}(s, b) = 1 - \frac{L_s(Q_H)}{L_b(Q_H)}, \tag{1}$$

when $0 < L_b(Q_H) < \infty$. The threshold claim is $C_k(H) = \{\theta_{Q_H}(s(H), b(H)) > r_k\}$. For a full-history conditional target, Q_H is the future evaluation law conditional on $\mathcal{F}$; a marginal target instead uses a declared fixed law Q. If $L_b = 0$, the relative ratio is undefined even when the predictors are identical.

The complete request also records the error event, conditioning field, quantifier over laws, coverage unit, and whether support is finite-sample, asymptotic, or empirical. Fixed-rule marginal validity, repeated control of the selected procedure, validity conditional on the selected label, and validity conditional on the full history are therefore distinct requests. Changing b, ℓ, I, w, Q, or the ladder changes the scientific question.

Executable rules. A request $\mathcal{R} = (\tau, \mathcal{V}, c)$ specifies the scientific comparison τ, validity record $\mathcal{V}$, and exact conclusion c. Given evidence $\mathcal{E}$ and accepted external premises $\mathcal{A}$, the matcher checks coordinates and the applicable rule's conditions, then returns the unchanged request with a licensed conclusion or a mismatch reason.

Table 2 exposes the eight checks and their outputs; Appendix B supplies every side condition.

The matching step has two distinct components. Scientific coordinates determine what is being compared—the selected rule, comparator, loss, members, horizons, weights, evaluation law, and threshold ladder—whereas validity coordinates determine what probability statement is requested, including the error event, conditioning field, law quantifier, coverage unit, and support mode. Agreement on a numerical threshold or local method is therefore insufficient if either group changes; conversely, a weaker record can remain visible under M8 without being relabeled as support for the stronger request.

Relative soundness. If the accepted evidence assertions are correct, the external assumptions hold under the declared laws, and every applied rule satisfies its side conditions, every conclusion in a finite matcher derivation is true at exactly its recorded scientific and validity scope. This is soundness relative to supplied premises, not their authentication or a completeness claim; the proof is in Appendix B.

One record, two requests. Suppose validation selects ridge autoregression against a seasonal-naive comparator under MAE, with fixed horizons and weights. The evidence records rejection at 0%, 2%, and 5%, and retains the local method, endpoint, assumptions, selection history, and provenance. For the matching request "Was the 5% null rejected?", M2 returns that recorded decision. For "Is false-promotion probability at most .05 conditional on H?", the same evidence supplies no conditional bound: M7 returns unsupported and M8 preserves the observed decision separately. Neither the selected comparison nor its observed outcome changes.

Table 2: The executable claim–evidence checks. Scientific and validity coordinates must match, except for permitted conditioning coarsening. The output always retains the request; Appendix B gives the full side conditions.

Rule	Evidence and decisive check	Output
M1	Pointwise identity; finite positive denominator if a ratio is requested	Identity; ratio only if defined
M2	Matching observed record; method, endpoint, reached nodes and provenance retained	Recorded decision
M3	Matched finite design, denominator, selection and reporting unit	Empirical statement for that design
M4	Theorem scope and every required premise match	Instantiated theorem
M5	Same integrable error indicator and joint law; verified coarser conditioning	Bound at the coarser field
M6	Full-history local bounds, measurable truth events, potential tests and ordered stopping	Fixed-sequence error bound
M7	A required coordinate, premise or type is missing or inconsistent	Unsupported request with reason
M8	Weaker evidence exists under a different scope	Separate weaker record; request unchanged

For paired losses, UBOA uses

$$Z_t(r) = (1 - r)L_{b,t} - L_{s,t}. \tag{2}$$

Nonnegative comparator losses make the effect claims nested: 5% implies 2%, which implies a positive reduction. This nesting concerns effect size, not validity strength. Exact predictor identity yields *structural equivalence*, not a superiority test. With usable evidence, the procedure reports the last rejected threshold, or no supported positive reduction if the first null is not rejected. Unusable evidence leaves the request unresolved; an applicability failure can stop evaluation beforehand.

Figure 1 summarizes the resulting evidence-to-claim path: it separates the selected result and history, the evidence actually available, the additional ingredients required for a full-history conditional certificate, and the conclusions that can be licensed. The left-to-right arrows encode reasoning order rather than a universal strength ranking, so the same observed decision may support a literal report while leaving a stronger conditional request unsupported.

4 Promotion and Validity

Claim matching separates the local test from the composition of its conclusions. We first specify the historical test used for the earlier evidence, then state the local premise required for fixed-sequence control and instantiate it conditionally.

4.1 Selection, access, and the local test

Validation selects one rule from a fixed candidate pool under a fixed comparator, loss, scope, and deterministic tie break. Later evidence cannot reopen selection or alter the claim. Predictor identity terminates structurally. A nonidentical rule enters the statistical procedure only when its evidence requirements and applicability conditions are satisfied.

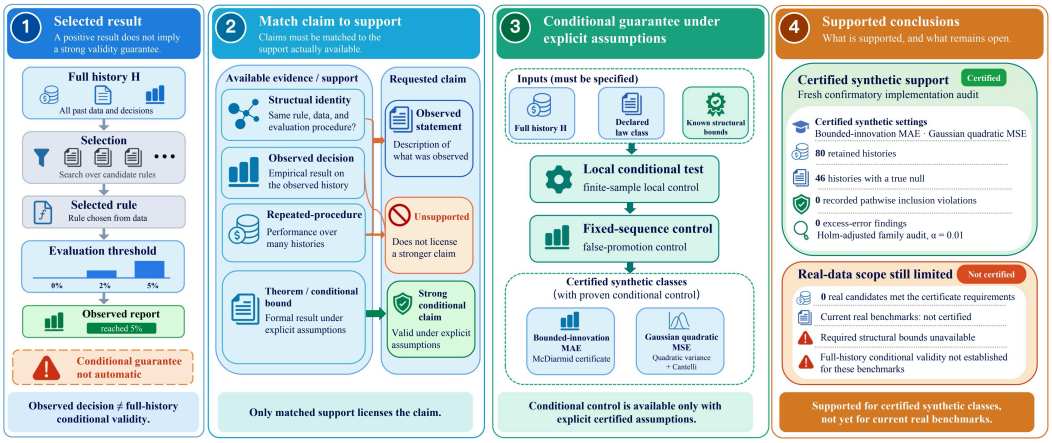


Figure 1: **From a selected result to its licensed validity scope.** UBOA checks a requested forecasting claim against the available evidence at its stated validity scope, retaining unsupported requests alongside separately scoped evidence. When the full pre-evaluation history H, a declared law class, and valid structural bounds are supplied, finite-sample conditional local control composes through the fixed sequence; we instantiate this for bounded-innovation MAE and Gaussian quadratic MSE. A fresh preregistered implementation audit retained 80 histories, including 46 with a finite first true null; it recorded zero pathwise inclusion violations and no Holm-adjusted evidence of excess false promotion. The theorem and assumptions supply the guarantee; current real benchmarks remain outside this certified scope.

For the historical self-normalized procedure, at a reached threshold set $X_t = Z_t(r)$ and test $H_r : \mathbb{E}[X_t] \le 0$. For n paired observations,

$$W_n = n^{-2} \sum_{t=1}^{n} \left[\sum_{j=1}^{t} (X_j - \bar{X}_n) \right]^2, \qquad T_n = \frac{\sqrt{n}\,\bar{X}_n}{\sqrt{W_n}}. \tag{3}$$

The upper-tail reference is the Brownian functional

$$\frac{B(1)}{\left[\int_0^1 \{B(u) - uB(1)\}^2 du \right]^{1/2}}. \tag{4}$$

Here B is standard Brownian motion; the ratio is not standard normal (Shao, 2010b; 2015). The historical procedure used a nominal local significance level of 0.05 with the stored Monte Carlo critical value 5.2988297532, rejecting at equality. Deterministic quadrature gives an ideal-reference upper 0.95 quantile of approximately 5.32268; Appendix L.4 documents this discrepancy without altering historical decisions. A nonfinite input or nonpositive normalizer cannot promote. Its use still depends on validity under the declared law.

4.2 What fixed-sequence composition establishes

Define each potential local test $\phi_j(Y)$ on the full evaluation sample, even when operational stopping would not reach it. The procedure tests the ordered nulls $\theta \le 0$, $\theta \le .02$, and $\theta \le .05$ and stops at the first valid nonrejection. Control at level α requires the actual local-bound premise below, not merely a nominal significance label. A reached invalid node terminates the whole application request as unresolved: prior decisions remain in the procedure record, no public positive promotion is reported, and later nodes are absent.

Proposition (first true null). For a fixed ordered family, let $J = \min\{j : H_j \text{ is true}\}$, with $J = \infty$ when no null is true, and define $\phi_\infty = 0$. If $\Pr\{\phi_J(Y) = 1\} \le \alpha$, then the probability of promoting through any true null is at most α.

Proof. Any false promotion requires rejection of H_J, so the false-promotion event is contained in $\{\phi_J(Y) = 1\}$. When $J = \infty$, the false-promotion event is empty. No independence among the local tests is required. The premise concerns the potential test rather than a rejection probability conditioned on having reached J. This is the classical fixed-sequence argument (Goeman & Solari, 2010).

For a selected family, let $T_j = \{H_j \text{ is true}\}$ be $\mathcal{F}$-measurable. Suppose the family and $J(H)$ are $\mathcal{F}$-measurable and, for every j,

$$1_{T_j}\mathbb{E}[\phi_j(H, Y) \mid \mathcal{F}] \le \alpha 1_{T_j} \quad \text{a.s.} \tag{5}$$

under the joint law of (H, Y), conditioning on $\mathcal{F} = \sigma(H)$. The same event inclusion gives conditional false-promotion probability at most α; the tower property then gives the corresponding repeated-procedure bound. Fixed-rule marginal levels alone do not imply Equation 5.

Two sequential constructions expose the missing implications. First draw $A \sim \text{Bernoulli}(\alpha)$; afterwards draw $V \sim \text{Unif}(0, 1)$ independently and set $U = \alpha V$ if $A = 1$, and $U = \alpha + (1 - \alpha)V$ otherwise. Then U is marginally uniform, yet $\Pr(U \le \alpha \mid A = 1) = 1$. With two independent copies, choose the first candidate whose earlier $A_j = 1$, using candidate one if neither does. Selection never reads the later U_j, each fixed test has level α, and the selected test rejects with probability $1 - (1 - \alpha)^2$. These analytic constructions diagnose missing implications, not the empirical cohort (Fithian et al., 2014; Foygel Barber et al., 2021); Appendix C gives the calculations.

4.3 Finite-sample conditional certificates

Equation 5 can be supplied without changing the selected-realization target when the full-history law provides numerical structural bounds. Let $W_r = \sum_i w_i\{(1 - r)L_{b,i} - L_{s,i}\}$ and $\mu_r^0 = \mathbb{E}[W_r \mid \mathcal{F}]$. We use two explicit settings.

Bounded innovations. For bounded contractive dynamics, replacing future innovation j changes the complete overlapping functional by at most $c_{j,r}$; with $C_r = \sum_j c_{j,r}^2$, the strict threshold $\sqrt{C_r \log(1/\alpha)/2}$ gives conditional local level α.

Gaussian quadratic losses. For Gaussian linear–quadratic MSE, $W_r = \varepsilon^\top A_r \varepsilon + 2g_r^\top \varepsilon + k_r$ and $V_r = 2\,\text{tr}(A_r^2) + 4\|g_r\|_2^2$; the strict Cantelli threshold $\sqrt{V_r(1 - \alpha)/\alpha}$ gives the same local guarantee. Appendix K gives assumptions, overlap coefficients, and proofs. These are scope-preserving instantiations of established concentration tools (Dedecker & Fan, 2015; Dedecker et al., 2019), not a distribution-free guarantee. Since each node truth event remains $\mathcal{F}$-measurable, the first-true-null proposition composes either valid local certificate without additional independence assumptions.

The historical connected-prefix comparison reuses the same Boolean decisions, so its agreement is a representation identity rather than independent confidence-set evidence (Appendix D).

5 Evaluation Design

Evidence roles. The matcher challenge tests contract conformance, not expert-review utility. The fresh audit tests certificate implementation and informativeness; its mathematical guarantee comes from the theorem and structural assumptions. Historical simulations and real applications retain their original inferential scope.

Matcher challenge. The 140 agent-authored cases and reference labels (70 permissible, 70 impermissible) were fixed before access to checker source or outputs. Revised and original matchers are compared with field-presence and coordinate-equality baselines.

Fresh conditional-certificate audit. The preregistered ten-cell study retains eight histories per cell and 4,096 independent future continuations per history, totaling 80 histories and 327,680 continuations. Every history was fixed before any continuation. Histories with a true ladder null enter per-history exact binomial audits combined by Holm at family level .01.

Historical repeated-procedure study. The 144 cells cross six process families, two sample lengths, six signals, and Gaussian or standardized-t_5 innovations. Validation selects one of six coefficients under squared loss. Boundary cells use 5,000 replications and other cells 2,000, totaling 432,000. Reporting policies and a connected-prefix representation are evaluated on the same saved records (Appendix E); invalid-input checks are separate.

External process families. The unchanged self-normalized test is evaluated on six designated families and two boundary probes at lengths 256, 512 and 1,024. Fractional long memory and a near-unit-root AR(1) with deterministic drift are outside-scope probes (Appendix F).

Six-application cohort. Six public UCI applications were fixed in outcome-blind order. The protocol specified 60/20/20 chronological splits, comparators, losses, horizons, applicability checks and 1,024 future origins per eligible application. Pools include simple statistical rules and simplified DLinear-/PatchTST-class candidates, not standard architecture reproductions (Zeng et al., 2023; Nie et al., 2023). The local method was fixed before validation and applications were not replaced. Appendix H records the protocol and unresolved preprocessing and screen-design details; saved outcomes are not a complete replay of all intended rules.

6 RESULTS

6.1 CLAIM-MATCHING CONFORMANCE

On the fixed 140-case challenge, the revised matcher made no false accepts or false rejects (Table 3). The original matcher falsely accepted 9/70 impermissible cases. Coordinate equality falsely accepted 29/70 and rejected 10/70 permissible cases; field presence accepted every impermissible case. Rule-specific side conditions, provenance, premise verification, and legal conditioning coarsening therefore add behavior beyond coordinate matching on this set. Appendix L.6 gives construction and error details for this controlled challenge.

Table 3: Implementation-blind contract challenge fixed before access to the checker implementation. False accepts (FA) use the 70 impermissible cases; false rejects (FR) use the 70 permissible cases. Field-presence and coordinate-equality baselines intentionally omit inference-rule side conditions. This controlled evaluation does not measure human-review utility or population generalization.

System	FA / 70	FR / 70	Exact accuracy
Revised matcher	0	0	100.0%
Original matcher	9	0	93.6%
Coordinate equality	29	10	72.1%
Field presence	70	0	50.0%

6.2 CONDITIONAL ERROR AND INFORMATIVENESS

The preregistered audit yielded no Holm-adjusted excess-error finding and showed substantial variation in informativeness. All 80 histories were retained; 46 had a finite first true null. The stored results cover 327,680 continuations, with zero invalid records and zero recorded first-true-null inclusion violations. One false promotion in one Gaussian-MSE negative history (1/4,096) remains in the record. The prespecified exact binomial/Holm audit at family level .01 rejected 0/46 histories. These results test failure detection; the theorem and declared assumptions supply the inequalities.

The supplement reconstructs 240 history–threshold means and cutoffs, their overlap coefficients or quadratic matrices, and the 46 audit probabilities. This checks coefficients and aggregates, not unsaved individual future paths. Additional, explicitly post hoc 99% simultaneous binomial upper bounds range from 0.206% to 0.266% over the 46 eligible histories. These bounds cover only the sampled histories and are not a preregistered equivalence analysis. The original audit has only 23.8% first-step detection probability at true error .06 (Appendix L.3).

Usefulness varies sharply: terminal 5% reports occur in 86.87% of bounded-MAE and 98.59% of Gaussian-MSE continuations in the strongest cells, but in none of the bounded-MAE 8% continuations and only 15/32,768 of the Gaussian nominal-8% continuations. Table 4 retains every cell.

Table 4: All ten fresh conditional-audit cells: eight histories and 32,768 continuations per row. True-null histories enter the primary audit. Truth ranges are history-conditional; terminal 5% frequencies are descriptive, not universal power. The sole false promotion occurred at the zero threshold in a Gaussian negative history.

Setting	$\theta^0(H)$ range (%)	True-null histories	False promotions	Terminal 5% reports (%)
MAE negative	-0.078	8	0	0.000
MAE boundary 2%	2.000	8	0	0.000
MAE boundary 5%	5.000	8	0	0.000
MAE 8%	8.000	0	0	0.000
MAE 40%	40.000	0	0	86.874
MSE negative	$[-1.108, -1.092]$	8	1	0.000
MSE nominal 2%	1.990–2.011	8	0	0.000
MSE nominal 5%	4.976–5.043	6	0	0.000
MSE nominal 8%	7.962–8.241	0	0	0.046
MSE nominal 40%	39.876–40.479	0	0	98.593

6.3 Historical error–recovery trade-offs

The 432,000 historical replications comprise 111,775 structural selections and 320,225 statistical evaluations. The latter contain 3,629 false promotions (1.1333%), with a maximum cellwise rate of 39/1,249 (3.1225%). Exact recovery is 56.6111% over all replications, positive recovery 40.8919% over positive-truth replications, and underclaim 57.4006% over statistical evaluations. Table 10 and Figure 2 in Appendix E give counts and comparisons. The pooled rates describe the unequal prespecified cell mixture, not conditional control for every history.

Descriptive mapping recovers a positive claim in 88.9727% of positive-truth cases but falsely promotes 16.8200% of statistical evaluations. Always abstaining eliminates false promotion by making no positive claim. Removing ordered stopping changes pooled false promotion from 1.1333% to 1.1336% and produces two incoherent later-node states. Relative to fixed-sequence reporting on the same Boolean decisions, UBOA additionally enforces claim identity, admissibility, identity-based decisions, and explicit unsupported requests; statistical power is unchanged.

Design-sensitivity results are reported in Appendix G; changing comparator, loss, scope, ladder, or candidate pool can alter the scientific claim and does not authorize post-outcome tuning.

6.4 Dependence stress tests

Across six designated challenge families, five have near-nominal null rejection. The seasonal family rejects 10.15% at $n = 256$, 6.80% at $n = 512$, and 5.60% at $n = 1024$, creating a short-sample warning. At $n = 1024$ and standardized effect 0.10, GARCH and Markov-variance power is 0.826 and 0.839; the other designated families range from 0.293 to 0.545. Long-memory and near-unit-root/drift probes fall outside current support.

6.5 Real-data reporting scope

Table 14 preserves the outcome-blind order and all six entries. In four applications, every predeclared node was reached and the 5% null was rejected. Their observed reductions range from 29.954% to 64.820% and use different losses; they are descriptive reductions rather than lower confidence bounds. Beijing PM2.5 selected the comparator itself, and Bike Sharing failed the pre-evaluation applicability screen, so both entries remain in the cohort. These historical decisions do not become confirmation of

Table 5: External process-family scope under the specified local test.

Disposition	Process families
Supported empirically	ARMA, GARCH, Markov variance, threshold AR, heavy-tailed AR
Warning	Seasonal process at short and medium sample lengths
Outside current scope	Fractional long memory; near-unit-root AR(1) with deterministic drift

certificates developed after their outcomes were observed. Appendix H records the local assumptions, screen diagnostics and unresolved reconstruction details.

7 Discussion and Limitations

Statistical scope. Finite-sample history-conditional control requires a declared future law and valid numerical structural bounds available before evaluation. The current real applications do not supply such certificates and therefore remain outside this certified scope.

Empirical coverage. The fresh audit studies one contraction coefficient, three horizons, and two loss/innovation settings. Its two-rule pools have equal conditional expected risk across the selected signs, so the experiment does not evaluate selection among candidates with unequal conditional risks. The finite-sample certificates are also conservative at moderate effects. These results establish behavior for the prespecified design rather than sensitivity across contraction strength, sample length, candidate separation, or horizon overlap.

Evidence and implementation. The matcher checks supplied annotations and premises but does not independently authenticate their provenance or truth; the controlled challenge evaluates contract conformance rather than downstream human-auditing performance. The supplement reconstructs saved coefficients and aggregate summaries, whereas individual future paths were not retained. Historical cohort replay is also partial because some original screening and preprocessing records were not preserved; accordingly, those real-data outcomes are not used as confirmatory evidence for the conditional guarantee. Appendix H and Appendix J document the remaining historical details.

8 Conclusion

UBOA makes evidence-to-claim licensing explicit for selected forecasting rules, preserving the requested scientific target and validity scope. The controlled challenge tests the eight-rule implementation; the two conditional constructions supply local guarantees that compose by fixed sequence. The fresh audit yielded no Holm-adjusted excess-error finding and exposed substantial conservatism. The real applications retain their historical interpretation outside the certified conditional scope.

Ethics statement

This work uses public, non-human-subject forecasting benchmarks and synthetic processes. It introduces no intervention, personal-data collection, or deployment decision.

AI use statement

Generative AI tools assisted experimental design and data analysis; conceptual and theoretical development; manuscript and figure revision; reference formatting; and authoring the contract-challenge cases and their reference-label rationales. The synthetic audit and implementation-blind challenge followed the prespecified timing described in the paper; historical real outcomes were not reused as fresh confirmation. The authors remain responsible for all claims, proofs, code, and disclosures.

Reproducibility statement

The accompanying anonymous code supplement contains the synthetic protocol and saved results, certificate reconstructions, original and revised matchers, the prespecified 140-case challenge and predictions, and an anonymous export of recovered historical evidence. Its root README lists deterministic verification commands and distinguishes coefficient and aggregate reconstruction from path replay. Future paths were not retained and raw-data-to-decision reproduction remains partial; Appendices J–L separate original records from later reconstruction.

References

Donald W. K. Andrews. Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation. *Econometrica*, 59(3):817, 1991. doi: 10.2307/2938229.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, et al. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi: 10.1214/24-AOAS1998. URL https://arxiv.org/abs/2110.01052.

Matthew Arnold, Rachel K. E. Bellamy, Michael Hind, et al. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development*, 63(4/5):6:1–6:13, 2019. doi: 10.1147/JRD.2019.2942288.

David H. Bailey and Marcos López de Prado. The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting, and Non-Normality. *The Journal of Portfolio Management*, 40(5): 94–107, 2014. doi: 10.3905/jpm.2014.40.5.094.

Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Christoph Bergmeir, Rob J. Hyndman, and Bonsoo Koo. A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, 120:70–83, 2018. doi: 10.1016/j.csda.2017.11.003.

Richard Berk, Lawrence Brown, Andreas Buja, et al. Valid post-selection inference. *The Annals of Statistics*, 41(2), 2013. doi: 10.1214/12-aos1077.

Florian Bordes, Candace Ross, Justine T. Kao, et al. Eval factsheets: A structured framework for documenting AI evaluations. *arXiv preprint arXiv:2512.04062*, 2025. URL https://arxiv.org/abs/2512.04062.

Frank Bretz, Willi Maurer, Werner Brannath, et al. A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine*, 28(4):586–604, 2009. doi: 10.1002/sim.3495.

Vitor Cerqueira, Luis Torgo, and Igor Mozetič. Evaluating time series forecasting models: an empirical study on performance estimation methods. *Machine Learning*, 109(11):1997–2028, 2020. doi: 10.1007/s10994-020-05910-7.

Todd E Clark and Michael W McCracken. Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics*, 105(1):85–110, 2001. doi: 10.1016/s0304-4076(01)00071-9.

Jérôme Dedecker and Xiequan Fan. Deviation inequalities for separately Lipschitz functionals of iterated random functions. *Stochastic Processes and their Applications*, 125(1):60–90, 2015. doi: 10.1016/j.spa.2014.08.001.

Jérôme Dedecker, Paul Doukhan, and Xiequan Fan. Deviation inequalities for separately Lipschitz functionals of composition of random functions. *Journal of Mathematical Analysis and Applications*, 479(2):1549–1568, 2019. doi: 10.1016/j.jmaa.2019.07.011.

Janez Demsar. Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

Francis X. Diebold and Roberto S. Mariano. Comparing Predictive Accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263, 1995. doi: 10.1080/07350015.1995.10524599.

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

Alex Dmitrienko and Ajit C. Tamhane. Gatekeeping procedures with clinical trial applications. *Pharmaceutical Statistics*, 6(3):171–180, 2007. doi: 10.1002/pst.291.

Alex Dmitrienko, Brian L. Wiens, Ajit C. Tamhane, et al. Tree-structured gatekeeping tests in clinical trials with hierarchically ordered multiple objectives. *Statistics in Medicine*, 26(12):2465–2478, 2007. doi: 10.1002/sim.2716.

Jesse Dodge, Suchin Gururangan, Dallas Card, et al. Show Your Work: Improved Reporting of Experimental Results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2185–2194, 2019. doi: 10.18653/v1/d19-1224.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, et al. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015. doi: 10.1126/science.aaa9375.

David Edwards and Jesper Madsen. Constructing multiple test procedures for partially ordered hypothesis sets. *Statistics in Medicine*, 26(28):5116–5124, 2007. doi: 10.1002/sim.2905.

William Fithian, Dennis Sun, and Jonathan Taylor. Optimal Inference After Model Selection. *arXiv preprint arXiv:1410.2597*, 2014. URL https://arxiv.org/abs/1410.2597.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, et al. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021. doi: 10.1093/imaiai/iaaa017.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, et al. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z.

Avijit Ghosh, Anka Reuel, Jenny Chim, et al. Evaluation cards: An interpretive layer for AI evaluation reporting. *arXiv preprint arXiv:2606.09809v2*, 2026. URL https://arxiv.org/abs/2606.09809v2.

Raffaella Giacomini and Halbert White. Tests of Conditional Predictive Ability. *Econometrica*, 74 (6):1545–1578, 2006. doi: 10.1111/j.1468-0262.2006.00718.x.

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.

Jelle J. Goeman and Aldo Solari. The sequential rejection principle of familywise error control. *The Annals of Statistics*, 38(6), 2010. doi: 10.1214/10-aos829.

David J. Hand. Classifier Technology and the Illusion of Progress. *Statistical Science*, 21(1), 2006. doi: 10.1214/088342306000000060.

Peter Reinhard Hansen. A Test for Superior Predictive Ability. *Journal of Business & Economic Statistics*, 23(4):365–380, 2005. doi: 10.1198/073500105000000063.

Moritz Hardt and Jonathan Ullman. Preventing False Discovery in Interactive Data Analysis Is Hard. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pp. 454–463, 2014. doi: 10.1109/focs.2014.55.

Peter Henderson, Riashat Islam, Philip Bachman, et al. Deep Reinforcement Learning That Matters. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018. doi: 10.1609/aaai.v32i1.11694.

Hansika Hewamalage, Klaus Ackermann, and Christoph Bergmeir. Forecast evaluation for data scientists: common pitfalls and best practices. *Data Mining and Knowledge Discovery*, 37(2): 788–832, 2023. doi: 10.1007/s10618-022-00894-5.

Sture Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. doi: 10.2307/4615733.

Mohammad F. Huque and Mohamed Alosh. A flexible fixed-sequence testing method for hierarchically ordered correlated multiple endpoints in clinical trials. *Journal of Statistical Planning and Inference*, 138(2):321–335, 2008. doi: 10.1016/j.jspi.2007.06.009.

Rob J. Hyndman and Anne B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688, 2006. doi: 10.1016/j.ijforecast.2006.03.001.

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, 2023. doi: 10.1016/j.patter.2023.100804.

Timothy Kassis, Vinayak Agarwal, Yuhuan He, et al. Scientific Agent Skills: A Library of Procedural Knowledge for Research Agents. arXiv preprint arXiv:2609.00065v2, 2026.

Nicholas M. Kiefer and Timothy J. Vogelsang. Heteroskedasticity-Autocorrelation Robust Standard Errors Using The Bartlett Kernel Without Truncation. *Econometrica*, 70(5):2093–2095, 2002. doi: 10.1111/1468-0262.00366.

Nicholas M. Kiefer and Timothy J. Vogelsang. A NEW ASYMPTOTIC THEORY FOR HETEROSKEDASTICITY-AUTOCORRELATION ROBUST TESTS. *Econometric Theory*, 21(06), 2005. doi: 10.1017/s0266466605050565.

Hans R. Kunsch. The Jackknife and the Bootstrap for General Stationary Observations. *The Annals of Statistics*, 17(3), 1989. doi: 10.1214/aos/1176347265.

Guokun Lai, Wei-Cheng Chang, Yiming Yang, et al. Modeling Long- and Short-Term Temporal Patterns with Deep Neural Networks. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 95–104, 2018. doi: 10.1145/3209978.3210006.

Jason D. Lee, Dennis L. Sun, Yuekai Sun, et al. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3), 2016. doi: 10.1214/15-aos1371.

Hannes Leeb. Conditional predictive inference post model selection. *arXiv preprint arXiv:0908.3615*, 2009. URL https://arxiv.org/abs/0908.3615.

Hannes Leeb and Benedikt M. Pötscher. MODEL SELECTION AND INFERENCE: FACTS AND FICTION. *Econometric Theory*, 21(1):21–59, 2005. doi: 10.1017/s0266466605050036.

Jing Lei, Max G'Sell, Alessandro Rinaldo, et al. Distribution-Free Predictive Inference for Regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018. doi: 10.1080/01621459.2017.1307116.

Bryan Lim, Sercan Ö. Arık, Nicolas Loeff, et al. Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764, 2021. doi: 10.1016/j.ijforecast.2021.03.012.

Ignacio N Lobato. Testing That a Dependent Process Is Uncorrelated. *Journal of the American Statistical Association*, 96(455):1066–1076, 2001. doi: 10.1198/016214501753208726.

S. Makridakis, A. Andersen, R. Carbone, et al. The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting*, 1(2):111–153, 1982. doi: 10.1002/for.3980010202.

Spyros Makridakis and Michèle Hibon. The M3-Competition: results, conclusions and implications. *International Journal of Forecasting*, 16(4):451–476, 2000. doi: 10.1016/s0169-2070(00)00057-1.

Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The M4 Competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, 34(4):802–808, 2018. doi: 10.1016/j.ijforecast.2018.06.001.

Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A Metric Learning Reality Check. In *Computer Vision – ECCV 2020*, Lecture Notes in Computer Science, pp. 681–699. Springer International Publishing, 2020. doi: 10.1007/978-3-030-58595-2_41.

Whitney K. Newey and Kenneth D. West. A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica*, 55(3):703, 1987. doi: 10.2307/1913610.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, et al. A Time Series Is Worth 64 Words: Long-term Forecasting with Transformers. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Jbdc0vTOcol.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, et al. Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program). *Journal of Machine Learning Research*, 22(164):1–20, 2021.

Marc-Oliver Pohle, Tanja Zahn, and Sebastian Lerch. Uncertainty quantification in forecast comparisons. *arXiv preprint arXiv:2605.03997*, 2026. URL https://arxiv.org/abs/2605.03997.

Dimitris N. Politis and Joseph P. Romano. The Stationary Bootstrap. *Journal of the American Statistical Association*, 89(428):1303–1313, 1994. doi: 10.1080/01621459.1994.10476870.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, et al. Do ImageNet Classifiers Generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 5389–5400, 2019.

Anka Reuel, Amelia Hardy, Chandler Smith, et al. BetterBench: Assessing AI benchmarks, uncovering issues, and establishing best practices. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0685.

David Salinas, Valentin Flunkert, Jan Gasthaus, et al. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3):1181–1191, 2020. doi: 10.1016/j.ijforecast.2019.07.001.

Xiaofeng Shao. Corrigendum: A Self-Normalized Approach to Confidence Interval Construction in Time Series. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(5):695–696, 2010a. doi: 10.1111/j.1467-9868.2010.00754.x.

Xiaofeng Shao. A Self-Normalized Approach to Confidence Interval Construction in Time Series. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(3):343–366, 2010b. doi: 10.1111/j.1467-9868.2009.00737.x.

Xiaofeng Shao. A simple test of changes in mean in the possible presence of long-range dependence. *Journal of Time Series Analysis*, 32(6):598–606, 2011. doi: 10.1111/j.1467-9892.2010.00717.x.

Xiaofeng Shao. Self-Normalization for Time Series: A Review of Recent Developments. *Journal of the American Statistical Association*, 110(512):1797–1817, 2015. doi: 10.1080/01621459.2015.1050493.

Leonard J. Tashman. Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting*, 16(4):437–450, 2000. doi: 10.1016/s0169-2070(00)00065-0.

Jonathan Taylor and Robert J. Tibshirani. Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25):7629–7634, 2015. doi: 10.1073/pnas.1507583112.

Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *CVPR 2011*, pp. 1521–1528, 2011. doi: 10.1109/cvpr.2011.5995347.

Sudhir Varma and Richard Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7(1), 2006. doi: 10.1186/1471-2105-7-91.

Gaël Varoquaux. Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage*, 180:68–77, 2018. doi: 10.1016/j.neuroimage.2017.06.061.

Runmin Wang and Xiaofeng Shao. Hypothesis testing for high-dimensional time series via self-normalization. *The Annals of Statistics*, 48(5), 2020. doi: 10.1214/19-aos1904.

Peter H. Westfall and Alok Krishen. Optimally weighted, fixed sequence and gatekeeper multiple testing procedures. *Journal of Statistical Planning and Inference*, 99(1):25–40, 2001. doi: 10.1016/s0378-3758(01)00077-5.

Halbert White. A Reality Check for Data Snooping. *Econometrica*, 68(5):1097–1126, 2000. doi: 10.1111/1468-0262.00152.

Brian L. Wiens. A fixed sequence Bonferroni procedure for testing multiple endpoints. *Pharmaceutical Statistics*, 2(3):211–215, 2003. doi: 10.1002/pst.64.

Haixu Wu, Jiehui Xu, Jianmin Wang, et al. Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting. In *Advances in Neural Information Processing Systems*, volume 34, pp. 22419–22430, 2021.

Ailing Zeng, Muxi Chen, Lei Zhang, et al. Are Transformers Effective for Time Series Forecasting? *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9):11121–11128, 2023. doi: 10.1609/aaai.v37i9.26317.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, et al. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):11106–11115, 2021. doi: 10.1609/aaai.v35i12.17325.

A Formal Validity-Scope Record

Let H be the random pre-evaluation record and $\mathcal{F} = \sigma(H)$ the sigma-field it generates. The record is not the sigma-field: metadata are functions of H, whereas conditional guarantees are statements relative to $\mathcal{F}$ or a declared subfield. A request is $\mathcal{R} = (\tau, \mathcal{V}, c)$, where

$$\tau = (S, s, b, \ell, I, w, Q, r_{1:K})$$

contains the selection rule and realization, comparator, loss, members, horizons, weights, evaluation-law kernel, and threshold ladder. The validity record $\mathcal{V}$ contains the error event, conditioning field, law quantifier or class, coverage unit, support mode, sample regime, and selection mechanism. The final coordinate c is the exact requested conclusion.

Evidence records are typed as structural identity, observed decision, finite-design empirical record, theorem, asymptotic theorem, conditional bound, or local-validity family. They are not ordered on one strength scale: an empirical record cannot become a theorem, and an asymptotic result cannot become finite-sample exactness.

B Evidence Matcher and Relative Soundness

Write $\mathcal{E}; \mathcal{A} \vdash \mathcal{R} \Rightarrow c$ when evidence records $\mathcal{E}$ and explicitly accepted external assumptions $\mathcal{A}$ license conclusion c for the unchanged request. Scientific coordinates (selected object, comparator, loss, members, horizons, weights, law, and thresholds) and validity coordinates (event, conditioning, law quantifier, coverage, support mode, sample regime, and selection mechanism) must match exactly, except for the conditioning coarsening in M5.Table 6 states the eight evidence-matching rules and the mechanically checked conditions for each licensed output.

Table 6: The eight evidence-matching rules. Every output retains the unchanged scientific and validity coordinates unless the rule explicitly coarsens conditioning.

Rule	Evidence input	Mechanically checked conditions	Licensed output
M1	Pointwise identity	Same predictors, outcomes, loss, and scope; finite positive comparator loss for a numeric ratio	Predictor and paired-loss identity; ratio only when defined
M2	Observed record	Same scope; method, endpoint, reached nodes, assumptions, and provenance retained	Literal reached-node decision; no probability guarantee
M3	Finite design	Same mixture, denominator, sample range, selection mechanism, and unit	Exact empirical statement for that design
M4	Theorem or asymptotic result	Target, law, event, conditioning, coverage, selection, regime, support mode, and all verified premises match	Instantiated theorem at its original scope
M5	Conditional bound	Same integrable event and joint law; verified $\mathcal{G} \subseteq \mathcal{H}$	$\mathbb{E}[\phi \mid \mathcal{H}] \leq \alpha$ implies $\mathbb{E}[\phi \mid \mathcal{G}] \leq \alpha$
M6	Local-validity family	Potential tests, measurable family/truth events, matching full-history bounds, ordered stopping, and invalid policy	Fixed-sequence false-promotion bound for the fixed request
M7	Failed match	Any required coordinate, premise, or type check is absent or inconsistent	Unsupported request with stable mismatch reasons
M8	Weaker evidence	M7 returned and a weaker record exists under a different scope	Unchanged request plus separately labeled weaker evidence

Table 7: Evidence requirements are matched to the requested conclusion. Missing premises remain visible; they are not replaced by a weaker target and called established.

Requested conclusion	Required justification	Available evidence	Admissible report
Predictor identity	Pointwise equality of selected rule and comparator	Deterministic identity check	Structural equivalence; no superiority test
Fixed-rule marginal level	Valid local test under the declared marginal law	Asymptotic self-normalized method plus stated assumptions	Test decision under those assumptions
Repeated selected procedure	Bound under the joint selection–evaluation law	Finite simulation design	Empirical operating characteristic for that design
Full-history conditional level	Almost-sure local bound given the declared history	Finite-sample certificates in two explicit synthetic settings; absent for current real applications	Conditional control only in the certified settings; real applications remain outside scope

For M5, if $\mathcal{G} \subseteq \mathcal{H}$ under the same joint law and the same integrable error indicator obeys $\mathbb{E}[\phi \mid \mathcal{H}] \leq \alpha$ almost surely, then the tower property gives $\mathbb{E}[\phi \mid \mathcal{G}] \leq \alpha$. The rule cannot reverse direction or change the event, estimand, joint law, or coverage unit while averaging.

For M6, define all potential tests on the joint sample space. Let $T_j = \{H_j \text{ is true}\} \in \mathcal{F}$ and assume

$$1_{T_j}\mathbb{E}[\phi_j \mid \mathcal{F}] \leq \alpha 1_{T_j} \quad \text{a.s. for every } j. \tag{6}$$

Set $J = \min\{j : T_j\}$, with $J = \infty$ when this set is empty, and $\phi_\infty = 0$. The family and truth events must be $\mathcal{F}$-measurable, the request fixed, and the procedure ordered. A reached invalid node makes the whole application request unresolved: earlier node results remain in the procedure record, no public positive promotion is returned, and later nodes are absent.

Table 7 summarizes how these rules distinguish four common requests by the justification they require and the report that can be licensed.

Proposition (relative non-amplification). Suppose the accepted evidence assertions are semantically correct, the external assumptions in $\mathcal{A}$ are satisfied under the declared laws, and every applied rule satisfies its side conditions. Every conclusion in a finite matcher derivation is true with exactly its recorded scientific and validity scope.

Proof. M1 substitutes equal predictor outputs into the same loss; a numeric relative ratio additionally needs a finite positive denominator. M2 quotes a reached record. M3 retains the source design and denominator. M4 is theorem instantiation after every coordinate and premise matches. M5 follows from conditional-expectation monotonicity and the tower property for the same event and law. For M6, false promotion implies $\phi_J = 1$ pathwise. On $\{J = j\}$, Equation 6 bounds its conditional expectation; when $J = \infty$, the false-promotion event is empty and $\phi_J = 0$. M7 derives no positive conclusion, and M8 preserves rather than strengthens both request and weaker record. At derivation depth zero, accepted leaves are true by premise. If every conclusion through depth d is true at its recorded scope, the rule-specific argument and checked side conditions make every depth-$d + 1$ conclusion true without changing scope. Induction completes the proof. □

This is a conditional property of the specification, not a theorem that verifies a real law. The matcher is incomplete and may reject a true conclusion for which it has no rule. It does not establish selected-procedure, full-history conditional, continuous-inversion, audit-superiority, or cohort-wide simultaneous validity for the applications.

Reference algorithm. The metadata-only implementation validates the request and evidence formats, dispatches by evidence type, checks required coordinates and premises, and returns either an established conclusion with a rule identifier or an unsupported request with stable reasons, the unchanged request, and separately scoped weaker evidence. Twenty-three hand-authored reference cases were fixed before the original implementation. Subsequent deterministic audits exposed incorrect acceptance for missing empirical fields, premise checks, conclusion types, validity coordinates, observed-record provenance, and empty theorem-premise lists. The revised matcher closes these gaps while retaining the original checker and all historical test cases. Because the strengthened observed-record contract requires explicit provenance, one earlier test case that omitted provenance is intentionally reclassified as unsupported; the other 22 retain their outcomes and a provenance-complete positive control is accepted. Tests designed after inspecting errors are development checks, not independent evaluation. A separate implementation-blind contract challenge, authored and fixed before access to the checker implementation, is reported in Appendix L.6.

Table 8 applies the matcher to four requests built from the same saved 5% decision, illustrating how the admissible report changes with the requested validity scope.

Table 8: Four requests applied to the same saved 5% decision. No statistic is recomputed.

Request	Result	Reason
What decision was recorded at 5%?	Admissible by M2	Copy the reached decision, named method, endpoint, assumptions, and provenance.
What was the finite-design rejection frequency?	Admissible by M3 only with the design record	Retain mixture, numerator, denominator, range, unit, and selection mechanism.
Does the selected procedure control false promotion conditional on the full history?	Unsupported by M7	A recorded rejection and fixed-rule asymptotics do not supply this theorem; M8 may display the decision separately.
Is the familywise false-promotion probability across applications at most 0.05?	Unsupported by M7	Per-request evidence does not imply cohort-wide familywise control.

Terminal states. Structural identity means exact predictor equality and no superiority test. A pre-evaluation stop means an applicability condition failed before evaluation. A valid nonrejection is absence of support at the reached node, not proof of the null. Not reached means no later-node record exists. Invalid evidence makes the application request unresolved. An unsupported requested guarantee may coexist with a usable weaker record, but the weaker record is not relabeled as the requested conclusion.

C Sequential Counterexamples and Composition

Fix $0 < \alpha < 1$. First draw the history $A \sim \text{Bernoulli}(\alpha)$. Afterwards draw $V \sim \text{Unif}(0,1)$ independently and set

$$U = \begin{cases} \alpha V, & A = 1, \\ \alpha + (1-\alpha)V, & A = 0. \end{cases}$$

The conditional densities are $1/\alpha$ on $[0,\alpha]$ and $1/(1-\alpha)$ on $[\alpha,1]$; mixing with weights α and $1-\alpha$ gives density one on $[0,1]$. Thus U is marginally uniform, $\Pr(U \le \alpha) = \alpha$, and $\Pr(U \le \alpha \mid A = 1) = 1$. History precedes the future variable and future randomness remains, so chronology alone is not independence.

Now generate two independent copies (A_j, V_j) and the corresponding U_j. Using only the earlier history, select the first j with $A_j = 1$; choose $j = 1$ if both are zero. Each fixed U_j is marginally

uniform, but the selected future test rejects exactly when at least one history bit is one:

$$\Pr(U_S \le \alpha) = 1 - (1-\alpha)^2 = 2\alpha - \alpha^2 > \alpha.$$

At $\alpha = 1/20, 1/10, 1/2$, the selected probabilities are $39/400, 19/100, 3/4$. These analytic constructions establish the two non-implications. Their law differs from the application cohort, whose recorded decisions are unchanged; the construction also makes no general claim against chronological splitting.

D Local Test and Boundary Convention

For threshold r, the null is $H_r : \theta \le r$ and the alternative is $\theta > r$. The paired contrast is $X_t = (1-r)L_{b,t} - L_{s,t}$. Equation 3 divides the signed sample mean by a within-sample partial-sum scale. The fixed implementation uses an upper-tail Brownian-functional reference, nominal local $\alpha = 0.05$, critical value 5.2988297532, and rejection at equality. Shao's result identifies the limiting Brownian functional (Shao, 2010b;a; 2015); the decimal is the historical empirical 0.95 quantile, not the exact limiting quantile. The accompanying anonymous code supplement contains the recovered generator source, declared 200,000-path configuration, and sorted reference array, whose linear empirical 0.95 quantile is 5.298829753159216 (Appendix L.4). Raw Brownian increments and the original generator execution transcript were not retained, so source-to-array execution provenance remains partial. A nonfinite input or nonpositive normalizer is unusable evidence. The asymptotic justification requires an appropriate functional central limit theorem for the fixed paired sequence; chronology alone does not establish it.

The connected-prefix comparison takes the largest prefix of the same Boolean ladder decisions used by fixed sequence. Its 432,000-record agreement is therefore a representation-consistency identity. It is not an independently computed confidence bound or a source of power. A continuous inversion would additionally require a connected lower rejection set, no threshold-specific invalidity, matched endpoints, and a justified monotonicity argument.

E Repeated-Procedure Design and Metrics

Each replication retains $2n$ observations after a 1,024-step burn-in and divides them into equal validation and later evaluation blocks, with $n \in \{128, 512\}$. The driver innovation u_t and response innovation ϵ_t are independent sequential draws from the same cell law: standard Gaussian or variance-one t_5, obtained by multiplying Student-t_5 by $\sqrt{3/5}$. The response and candidates are

$$y_t = \beta x_{t-1} + \epsilon_t, \qquad \widehat{y}_t(g) = g x_{t-1},$$

with $g \in \{0, .05, .10, .20, .30, .40\}$, comparator $g = 0$, and $\beta \in \{0, .05, .10, .18, .26, .40\}$. Validation selects the smallest MSE, with numeric g order as the tie break. Table 9 gives the exact six synthetic driver processes and their normalizations used in this repeated-procedure study.

Under the declared synthetic evaluation law, $\mathbb{E}[x^2] = \mathbb{E}[\epsilon^2] = 1$ and the two innovations are independent. Hence $L(g) = 1 + (\beta - g)^2$ and the selected-realization relative reduction is

$$\theta(\beta, g) = \frac{2\beta g - g^2}{1 + \beta^2}.$$

This derivation applies to the represented synthetic marginal law. In particular, the seeded-phase block-variance process is a deterministic alternating variance regime, not a general stationary-law assertion.

The Cartesian design has $6 \times 2 \times 6 \times 2 = 144$ cells. Cells with $\beta \in \{0, .10\}$ have 5,000 replications; all others have 2,000, totaling 432,000. Constant and nonfinite invalid-evidence checks are deterministic and outside these stochastic cells. False promotion divides by statistical evaluations, exact recovery by all replications, positive recovery by positive-truth replications, and underclaim by statistical evaluations.

Table 11 compares these reporting policies on the same saved records: the separate-test policy produced two incoherent later-node states, while ignoring structural identity changed 111,775 terminal interpretations. Other deterministic checks violate outcome isolation, fixed claim identity, or unusable-input handling. They test the specification and are excluded from stochastic denominators.

Table 9: Exact prespecified synthetic drivers. Here u_t has variance one and unavailable lags are omitted. Every retained series has a 1,024-step burn-in except where initialization is stated.

Driver	Definition and normalization
Mild AR	$x_0 = u_0$; $x_t = .2x_{t-1} + \sqrt{1 - .2^2}\, u_t$.
Strong AR	$x_0 = u_0$; $x_t = .8x_{t-1} + \sqrt{1 - .8^2}\, u_t$.
Seasonal AR	For $t < 12$, $x_t = \sqrt{1 - .65^2}\, u_t$; afterwards $x_t = .65x_{t-12} + \sqrt{1 - .65^2}\, u_t$.
Multiscale AR	$x_t = .35x_{t-1} + .40x_{t-12} + su_t$, where $s = (\sum_{j=0}^{19999} \psi_j^2)^{-1/2}$, $\psi_0 = 1$, and $\psi_j = .35\psi_{j-1} + .40\psi_{j-12}$.
Block variance	$x_t = \sigma_t u_t / \sqrt{8.5}$; seeded-phase alternating 32-step blocks have $\sigma_t \in \{1, 4\}$. The normalization averages these deterministic regimes.
Markov variance	$p_{00} = .97$, $p_{11} = .92$, scales $(1, 4)$, stationary $\pi_0 = 8/11$, and $x_t = \sigma(S_t)u_t / \sqrt{\pi_0 + 16(1 - \pi_0)}$.

Table 10: Repeated-procedure operating characteristics under the declared selected-rule truth and prescribed design mixture. Denominators are statistical evaluations for false promotion and underclaim, positive-truth replications for positive recovery, and all replications for exact recovery and unresolved/invalid outcomes.

Metric	Count / denominator	Rate
False promotion	3,629 / 320,225	1.1333%
Exact terminal recovery	244,560 / 432,000	56.6111%
Positive recovery	97,171 / 237,629	40.8919%
Underclaim	183,811 / 320,225	57.4006%
Unresolved / invalid	0 / 432,000	0%

Table 11: Selected-realization policies evaluated on the same saved records. Rates retain their specified denominators.

Policy	False promotion	Exact recovery	Positive recovery	Underclaim	Scientific role
UBOA fixed sequence	1.1333%	56.6111%	40.8919%	57.4006%	Target procedure
Connected-prefix representation	1.1333%	56.6111%	40.8919%	57.4006%	Same Boolean decisions
Separate nominal tests	1.1336%	56.6111%	40.8927%	57.4003%	Removes ordered stopping
Direct descriptive mapping	16.8200%	75.6282%	88.9727%	16.0587%	Noninferential ceiling
Always abstain	0%	25.8738%	0%	100%	Conservative floor
Identity-agnostic testing	1.1333%	30.7373%	40.8919%	57.4006%	Removes structural semantics

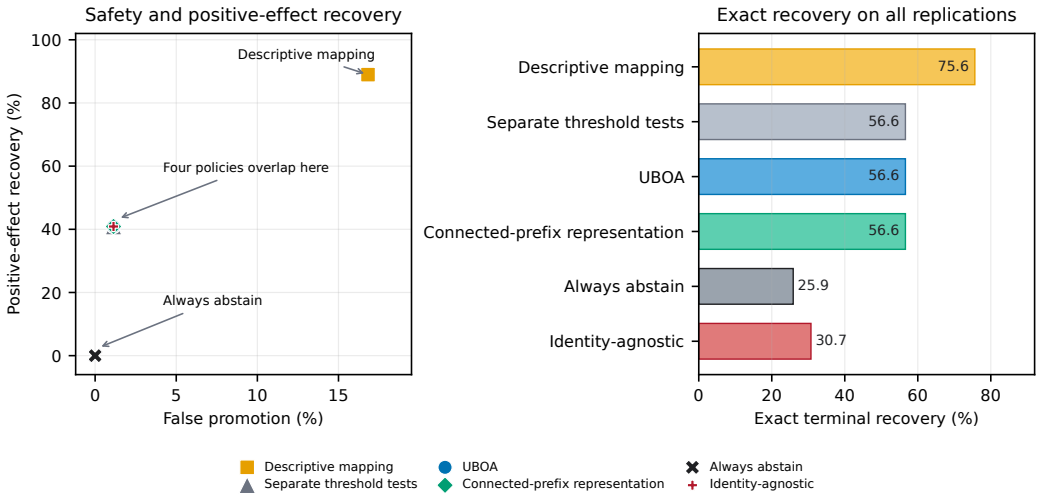


Figure 2: Pooled behavior on the same 432,000 saved records. The left-panel annotations identify overlapping points; the right panel distinguishes all policies by exact recovery. Descriptive mapping has higher positive-effect recovery and 16.8200% false promotion; always abstaining has neither. UBOA and the connected-prefix representation overlap exactly because they reuse the same Boolean decisions; this is a representation check, not independent inferential corroboration. Metrics use the denominators stated in Table 10.

F External Process Challenge

Here AR denotes autoregressive, ARMA autoregressive moving-average, and GARCH generalized autoregressive conditional heteroskedasticity. Table 12 gives the exact definitions and interpretive roles of the eight process families. The challenge crosses eight families, $n \in \{256, 512, 1024\}$, and $\delta \in \{0, .005, .01, .02, .03, .05, .07, .10\}$. Zero-shift cells have 2,000 replications and positive-shift cells 1,000, giving 192 cells and 216,000 evaluations.

Table 13 reports the valid-null rejection rates and the separately labeled out-of-scope stress responses. At $n = 1024$ and effect .10, power is .545 for ARMA, .293 for seasonal AR, .826 for GARCH, .839 for Markov variance, .361 for threshold AR, and .416 for heavy-tailed AR. The fractional-noise rows are outside the admitted law class. The zero-added-shift near-unit-root probe already has deterministic drift, so its rejection rate is not a Type-I-error or size estimate.

G Claim Sensitivity

The sensitivity study has 32 base worlds and 32,000 replications. Same-claim variations alter sample length, dependence strength, variance geometry, or the selection seed while retaining the realized rule. Claim-changing variations alter comparator, loss, scope, ladder, or candidate pool. Comparator, loss, ladder, and pool changes altered truth depth in 56.25%, 50.00%, 25.00%, and 15.625% of base worlds. The scope change altered the estimand and the terminal record, but the balanced construction produced exactly 0.000% discrete truth-depth change. A zero bar in Figure 3 is therefore a result, not a missing value. These saved sensitivity results do not permit post-outcome tuning.

H Cohort Protocol and Scope Audit

The six public UCI datasets remain in the order shown in Table 14. Table 15 records their source identifiers and the claim-defining frequency, target, comparator, loss, lookback, horizons, membership, and aggregation fields. Their source metadata record the Creative Commons Attribution 4.0 (CC BY 4.0) license. Chronological row identities define $\lfloor .6N \rfloor$ training, a validation prefix through $\lfloor .8N \rfloor$, and the final 20% evaluation suffix. Beijing uses $N = 35{,}064$ rows per station. Household Power first splits minute rows, then forms consecutive complete 60-row hourly means.

Table 12: External process challenge. Every retained stochastic path is divided by its positive root-mean-square magnitude without demeaning; a constant standardized shift δ is then added.

Family	Prespecified definition	Interpretive role
ARMA(2,1)	$x_t = .55x_{t-1} - .20x_{t-2} + e_t + .35e_{t-1}$; Gaussian; 1,024 burn-in	In-scope valid null
Seasonal AR	$(1 - .35L)(1 - .55L^{12})x_t = e_t$; Gaussian; 1,536 burn-in	In-scope null with short-sample warning
GARCH(1,1)	$x_t = \sqrt{h_t}e_t$, $h_t = .05 + .08x_{t-1}^2 + .87h_{t-1}$, $h_0 = 1$; 1,024 burn-in	In-scope valid null
Markov variance	$p_{00} = .97$, $p_{11} = .92$, scales $(1, 4)$, stationary initialization	In-scope valid null
Threshold AR	Difference of two independent paths, each coefficient .35 after a negative lag and .75 otherwise; 1,024 burn-in	In-scope valid null
Heavy-tailed AR	$x_t = .6x_{t-1} + e_t$ with variance-one t_7 innovations; 1,024 burn-in	In-scope valid null
Fractional noise	$d = .35$, $w_0 = 1$, $w_k = w_{k-1}(k - 1 + d)/k$, truncated at lag 1,024	Out-of-scope assumption stress
Near-unit-root AR(1) with drift	$x_t = .995x_{t-1} + \sqrt{1 - .995^2}e_t$, zero start and no burn-in; add $.002t/\sqrt{n}$ before δ	Out-of-scope, non-null drift stress

Table 13: Valid-null size diagnostics and out-of-scope stress responses at nominal level 0.05. The second panel is not a Type-I-error or size assessment.

Family	Scope	$n = 256$	$n = 512$	$n = 1024$
Valid-null size diagnostics				
ARMA(2,1)	in	.050	.048	.048
Seasonal AR	warning	.102	.068	.056
GARCH(1,1)	in	.053	.053	.048
Markov variance	in	.044	.051	.050
Threshold AR	in	.053	.053	.051
Heavy-tailed AR	in	.058	.051	.053
Non-null or out-of-scope stress response; not size				
Fractional noise	out	.175	.164	.151
Near-unit-root AR(1) with drift	out	.210	.190	.148

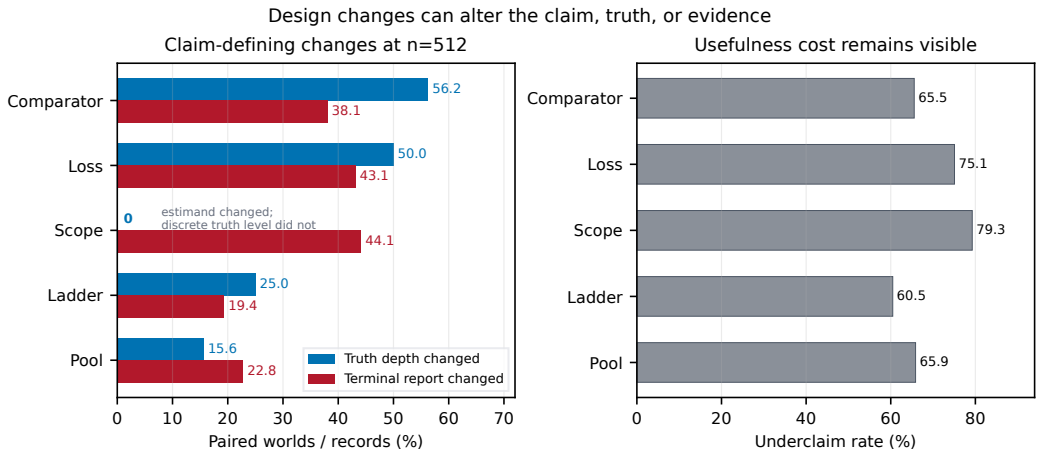


Figure 3: Design sensitivity on the saved study at $n = 512$. Left: percentages of paired worlds with changed truth depth (blue) and paired records with changed terminal reports (red) after changing each claim-defining field. Right: the corresponding underclaim rates. For scope, truth-depth change is exactly 0.000% in this balanced construction even though the estimand changes; the zero is a result, not a missing value. These retrospective sensitivity summaries do not authorize post-outcome tuning.

Table 14: Prospective cohort in its outcome-blind prespecified order. In the rule/comparator column, the first line is the selected rule and the second is the comparator. The four 5% null rejections are observed test decisions, not full-history conditional guarantees; no simultaneous cohort-wide control is claimed.

Application	Selected rule / comparator	Loss	Before evaluation	Procedure outcome	Reduction
Appliances	Ridge autoregression Seasonal naive (lag 144)	MAE	Eligible	5% null rejected	29.954%
Beijing PM2.5	Last value Last value	MAE	Identity	Structural equivalence; no test	N/A
Bike Sharing	Ridge autoregression Seasonal naive (lag 24)	MAE	Stopped	21/24 primary cells $\leq .065$; short-cell max $.078 > .075$; not evaluated	N/A
Household Power	Ridge autoregression Seasonal naive (lag 24)	MSE	Eligible	5% null rejected	56.984%
Air Quality CO	Ridge autoregression Last value	MSE	Eligible	5% null rejected	42.753%
Metro Traffic	Ridge autoregression Seasonal naive (lag 24)	MSE	Eligible	5% null rejected	64.820%

Table 15: Public application identities and claim-defining protocol fields. UCI identifiers and DOIs identify the source releases. The evaluation-design column retains lookback, horizons, members, and aggregation in that order.

Application and UCI source	Frequency; target	Comparator; loss	Evaluation design
Appliances, ID 374, DOI 10.24432/C5VC8G	10 minute; Appliances	seasonal lag 144; MAE	Lookback 144; horizons $(1, 6, 18)$; one series; mean across horizons per origin
Beijing Multi-Site Air Quality, ID 501, DOI 10.24432/C5RK5G	hourly; PM2.5	last value; MAE	Lookback 48; horizons $(1, 6, 24)$; 12 stations; equal mean over members and horizons
Bike Sharing, ID 275, DOI 10.24432/C5W894	hourly; cnt	seasonal lag 24; MAE	Lookback 48; horizons $(1, 6, 24)$; one series
Household Power, ID 235, DOI 10.24432/C58K54	minute source to hourly; Global active power	seasonal lag 24; MSE	Lookback 48 hours; horizons $(1, 6, 24)$; consecutive complete 60-row means; incomplete tail discarded
Air Quality, ID 360, DOI 10.24432/C59K5F	hourly; CO(GT)	last value; MSE	Lookback 48; horizons $(1, 6, 24)$; one series
Metro Interstate Traffic, ID 492, DOI 10.24432/C5X60B	hourly; traffic volume	seasonal lag 24; MSE	Lookback 48; horizons $(1, 6, 24)$; one series

Training origins run from the lookback through the last in-range training target and are capped at 4,096 deterministic evenly spaced integer indices. Validation uses stride-one origins beginning at the training boundary. The later evaluation starts at $\lfloor .8N \rfloor + \max(horizons)$ and uses the first 1,024 complete stride-one origins. Per-origin loss averages fixed horizons and, for Beijing, the 12 members.

Nonnumeric entries become missing. Air Quality also converts comma decimals and treats values at most -199 as missing. The implementation forward-fills causally after an initial run, but fills an initial missing run from the first later finite value. This initial fill is not strictly past-only. The protocol's rule to defer an application with more than 20% missing validation targets has no recovered denominator or pre-fill enforcement record; the inspected implementation only rejects remaining nonfinite values after fill.

Every pool contains its comparator, the last-three mean, ridge autoregression, and a DLinear-class candidate. A tiny PatchTST-class candidate appears for Appliances, Bike, Air Quality, and Metro Traffic; its panel-data and missing-value requirements exclude Beijing, and the prespecified CPU budget excludes Household Power.Table 16 records the recoverable candidate implementations and the fixed fitting and selection fields.

The study protocol describes tie-breaking by candidate identifier within 10^{-12}, whereas the inspected execution code sorts exact floating scores and uses candidate identifier only after exact equality. This discrepancy could change a near-tie reproduction and is not retroactively harmonized.

For a selected nonstructural application, let $z_t = L_{b,t} - L_{s,t}$ and $c_t = z_t - \bar{z}$. The prior descriptor is lag-one correlation, seasonal correlation at $k = \min(\text{seasonal lag}, \max(1, \lfloor n/4 \rfloor))$, and the ratio of maximum to minimum nonoverlapping $q = \max(8, \lfloor n/8 \rfloor)$ block variance. Undefined correlations become zero. Four generators combine clipped lag-one and seasonal correlations with Gaussian or variance-one t_5 innovations: AR, seasonal lag 24, alternating 32-step block variance, and an equal-weight multiscale AR/seasonal form. Each retained path is divided by its root-mean-square magnitude without centering.

The historical summary reports $4 \times 2 \times 3 = 24$ primary null cells at $n = 1024$, with 2,000 replications per cell, and six short cells at $n = 512$ with 1,000 replications. The four generators and two innovation laws are named above, but the third factor, its levels, and the six short-cell combinations are not

Table 16: Recoverable candidate implementations and fixed training/selection fields.

Candidate	Exact implementation	Fit and selection record
Ridge autoregression	Closed form on $[1, \text{raw lookback}]$; penalty diagonal $(0, 1, \ldots, 1)$; slope penalty 1.0; no standardization	Up to 4,096 deterministic evenly spaced windows
DLinear-class	One affine map from lookback to the fixed horizon vector; not the original decomposition architecture	20 epochs; scalar train-window standardization
Tiny PatchTST-class	Patch length 8, stride 4, width 16, one two-head encoder layer, feed-forward width 32, zero dropout, mean pooling	12 epochs; present in four pools only
Neural fit	Adam at .001, batch 128, deterministic Torch algorithms, no shuffle, last epoch, inference batch 512	Fixed deterministic seeds; validation averages the primary loss over horizons, origins, and members
Selection	Lowest floating validation score, then candidate identifier after an exact score tie	This differs from the protocol's stated 10^{-12} tie tolerance

specified in the retained manuscript description (Table 17). The saved screen decisions are therefore reported as historical diagnostics, not as a fully reconstructable grid. Thirty-two positive-effect cells are diagnostic and cannot rescue a null-screen failure. The stated eligibility rule is

$$N_{\text{invalid}} = 0, \quad \max p_{\text{null}} \leq .075, \quad \#\{p_{\text{null}} \leq .065\} \geq 23, \quad \max p_{\text{short}} \leq .075.$$

The archived protocol records .065 and .075 as prespecified empirical tolerance bands, not as cutoffs derived from a multiplicity theorem. At 2,000 Bernoulli replications with nominal probability .05, these lie about 3.1 and 5.1 Monte Carlo standard errors above .05; for the 1,000-replication short cells, .075 is about 3.6 such standard errors above .05. The 23/24 clause permits at most one primary cell outside the .065 band. These calculations describe the scale of the prespecified rule rather than retroactively deriving it; the constants are study-specific and would need prospective justification in another application. Table 18 reports the resulting prospective local-validity diagnostics and eligibility decision for each nonstructural application. Bike had 21/24 primary rates at most .065 and short-cell maximum .078, so it was stopped before evaluation; its primary maximum .074 and invalid count zero satisfied their clauses. Beijing was structurally resolved and did not enter the screen.

Table 17: What the retained description specifies about the historical eligibility screen. The third factor and short-cell combinations are not determined by the reported counts; they are not inferred from which applications passed.

Component	Recorded definition or remaining gap
Four generators	AR; seasonal lag 24; alternating 32-step block variance; equal-weight multiscale AR/seasonal form.
Two innovation laws	Gaussian and variance-one t_5.
Third factor (three levels)	Factor identity and level values are not specified in the retained description.
Primary-cell budget	24 cells; $n = 1024$; 2,000 replications per cell.
Short-cell budget	Six cells; $n = 512$; 1,000 replications per cell; exact combinations are not specified.
Positive-effect diagnostics	32 cells; cannot override a null-screen failure.

Four statistical entries were evaluated once and rejected all three threshold nulls. Beijing selected last value against last value and was not evaluated. Bike was stopped before evaluation. Figure 4 summarizes these six terminal procedure states in the prespecified cohort order. The cohort was not

Table 18: Prespecified prospective local-validity diagnostics. Invalid counts unusable records; $\leq .065/24$ counts primary null cells with empirical rejection rate at most .065 out of 24 cells. Max and Short max are the largest rejection rates over the primary and six prespecified short null cells, respectively. Eligibility requires zero invalid records, at least 23 qualifying primary cells, and both maxima at most .075. Beijing PM2.5 was resolved by structural identity and does not enter this screen.

Application	Invalid	$\leq .065/24$	Max	Short max	Decision
Appliances	0	23	.0680	.0650	Eligible
Bike Sharing	0	21	.0740	.0780	Stopped
Household Power	0	24	.0615	.0660	Eligible
Air Quality CO	0	24	.0640	.0670	Eligible
Metro Traffic	0	24	.0585	.0630	Eligible

reordered or supplemented after outcomes. The validity-scope record was applied retrospectively and changes interpretation, not the stored decisions.

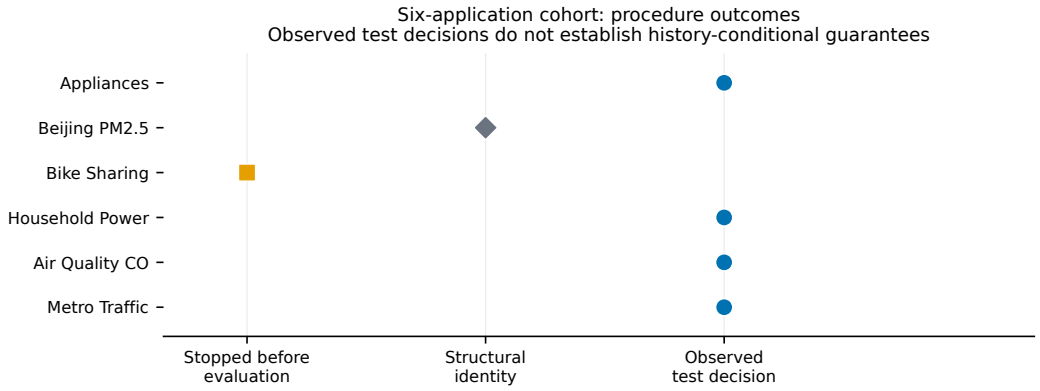


Figure 4: Cohort outcomes in the prespecified order. Position encodes procedure outcome, not effect magnitude. Four observed rejections remain separate from the unestablished history-conditional guarantee for these real applications. Each eligible sequence used 1,024 chronological origins; the structural and stopped entries were not evaluated.

I Earlier Real-Data Record

Traffic and Electricity selected the comparator and ended in structural identity. Weather and ETTm2 selected the mean of the last three observations over last value, rejected the zero-reduction null, and stopped when the 2% null was not rejected; the 5% node was not reached. Their descriptive MAE reductions were .874% and .862%.

The earlier protocol initially planned a prewhitened heteroskedasticity-and-autocorrelation-consistent (HAC) test. Benchmark-matched calibration after validation but before evaluation showed that it was inadequate for the fixed process envelope, so the self-normalized test was adopted while selections, comparators, losses, scopes, thresholds, and evaluation blocks stayed fixed and unopened. The later cohort fixed the method before validation.

These saved records also exercise the mismatch rules on a real protocol history. Consider two requests about a final Weather or ETTm2 decision. First, asking the observed reached-node decision to establish false-promotion control conditional on the full validation history returns M7: the available observed record and fixed-rule asymptotic premise do not establish that requested guarantee; M8 can display the observed decision separately without relabeling it. Second, asking to copy the final decision as though it had been produced by the originally planned prewhitened HAC method also fails

the observed-record side condition that the method be retained, because the recorded final method is the self-normalized test. Neither mismatch invalidates the prospectively amended decision; each prevents a stronger or historically inaccurate description of it. The later prospective cohort fixes the final method before validation and therefore does not have this method-identity mismatch.

J Reproducibility and Reporting Context

The accompanying anonymous code supplement contains the original and revised metadata matchers, the prespecified implementation-blind contract challenge and predictions, the conditional-study protocol, saved histories and aggregate results, deterministic certificate reconstruction, and an anonymous export of recovered historical evidence. Deterministic reconstruction computes every history-specific certificate and table entry without generating new random paths. These records permit certificate and aggregate-audit verification, but not independent pathwise replay: future innovations and individual terminal decisions were not retained.

Historical raw-data-to-decision reproduction remains partial. Under the saved validation scores, the smallest top-two candidate gap is 0.0181474, so changing only exact-tie versus 10^{-12} tie handling does not alter any of the six recorded winners. The six checked target prefixes have no leading missing run, so the future-assisted initial-fill branch was not triggered there. However, the original 20% missing-target denominator and enforcement log remain unrecovered; retrospectively derived masks do not replace that record. Household Power used consecutive complete 60-row groups rather than calendar-hour groups; evaluating the latter would require new preprocessing and refitting. The class-inspired neural candidates are not default reproductions of DLinear or PatchTST. The historical environment recorded NumPy 1.26.4 and Torch 2.11.0, but not environment-wide thread count or bitwise cross-hardware portability.

For the historical Brownian reference, the supplement exports the recovered generator source, declared configuration, sorted 200,000-value reference array, and 16 reached-node paired-loss records. All 16 statistics recompute exactly, and none lies between the historical empirical cutoff 5.298829753159216 and the ideal continuous-reference quantile 5.322679900530264; replacing only the cutoff would therefore change none of those 16 recorded decisions. Raw Brownian increments, unreached-node statistics, and the original generator execution transcript remain unavailable. Original protocol and result files are retained separately from later reconstructions and matcher corrections. Matching file contents do not establish when or how often an experiment ran, annotation truth, or stochastic assumptions (Dodge et al., 2019; Pineau et al., 2021; Kassis et al., 2026).

K Finite-Sample Conditional Certificates and Fresh Audit

The bounded construction specializes established martingale concentration for contractive iterated random functions (Dedecker & Fan, 2015; Dedecker et al., 2019); it is not a new general inequality. This appendix records the assumptions behind Section 4.3. The object is unchanged from the main text: for a fixed threshold r, the selected rule, comparator, raw loss, members, horizons, weights, and evaluation scope are all determined by the full pre-evaluation record H, with $\mathcal{F} = \sigma(H)$, and

$$W_r = \sum_{i \in I} w_i \{(1-r)L_{b,i} - L_{s,i}\}, \qquad \mu_r^0 = \mathbb{E}[W_r \mid \mathcal{F}].$$

For every law in the declared class, assume the conditional distributions exist and the stated bounds hold almost surely. The null is $\mu_r^0 \le 0$. The constructions below therefore do not replace the selected-realization future-block target by a marginal, selected-label, repeated-procedure, or predictable-average target.

K.1 Bounded-innovation contractive setting

Condition on $\mathcal{F}$. Let the future state satisfy $Z_t = F_t(Z_{t-1}, \varepsilon_t)$ with future innovations independent conditional on $\mathcal{F}$ and Z_0 fixed by H. Suppose the declared law supplies numerical, $\mathcal{F}$-measurable constants $0 \le \rho_t < 1$ and $D_t < \infty$ such that, on the reachable state space,

$$d(F_t(z,e), F_t(z',e)) \le \rho_t d(z,z'), \qquad \sup_{z,e,e'} d(F_t(z,e), F_t(z,e')) \le D_t.$$

For term $i = (o_i, h_i, k_i)$, let the paired-loss contrast $\psi_{i,r}$ satisfy

$$|\psi_{i,r}(z,u) - \psi_{i,r}(z',u')| \le A_{i,r}d(z,z') + B_{i,r}d(u,u')$$

with known $\mathcal{F}$-measurable coefficients. Replacing future innovation j then changes the complete overlapping functional by at most

$$c_{j,r} = D_j \sum_i w_i \Bigg[A_{i,r}\mathbf{1}\{j \le o_i\} \prod_{k=j+1}^{o_i} \rho_k + B_{i,r}\mathbf{1}\{j \le o_i + h_i\} \prod_{k=j+1}^{o_i+h_i} \rho_k \Bigg]. \tag{7}$$

where an empty product is one. This sum is taken over all overlapping origins and horizons before squaring, so shared innovations are not treated as independent loss terms. For $C_r = \sum_j c_{j,r}^2 > 0$ and $x > 0$, the conditional bounded-difference argument gives

$$\Pr\{W_r - \mu_r^0 \ge x \mid \mathcal{F}\} \le \exp(-2x^2/C_r).$$

Thus the strict local test

$$\phi_r^{\mathrm{bd}} = \mathbf{1}\left\{ W_r > \sqrt{\frac{C_r}{2}\log\frac{1}{\alpha}} \right\} \tag{8}$$

satisfies Equation 5. In the fresh audit this construction is instantiated with $Y_t = .5Y_{t-1} + \varepsilon_t$, Rademacher innovations, raw MAE, horizons $\{1,3,6\}$, and the structural support $|Y_t| \le 2$; no empirical maximum is used as a certificate.

K.2 Gaussian quadratic-MSE setting

Condition on $\mathcal{F}$ and collect all future scalar innovations in $\varepsilon \sim N(0, I)$. For the linear forecasts used in the audit, the complete overlapping raw-MSE contrast has the exact representation

$$W_r = \varepsilon^\top A_r \varepsilon + 2g_r^\top \varepsilon + k_r,$$

where $A_r = A_r^\top, g_r, k_r$ are determined by H, the comparator, the selected mapping, and the fixed origin–horizon weights. No diagonal approximation is made: off-diagonal entries of A_r retain overlap from shared innovations. Gaussian moments give

$$\mu_r^0 = k_r + \operatorname{tr}(A_r), \qquad V_r = \operatorname{Var}(W_r \mid \mathcal{F}) = 2\operatorname{tr}(A_r^2) + 4\|g_r\|_2^2.$$

For $V_r > 0$, Cantelli's inequality gives the strict local test

$$\phi_r^{\mathrm{quad}} = \mathbf{1}\left\{ W_r > \sqrt{\frac{1-\alpha}{\alpha}V_r} \right\}, \tag{9}$$

which satisfies Equation 5. The audit instantiation uses $Y_t = .5Y_{t-1} + \varepsilon_t$, raw MSE, horizons $\{1,3,6\}$, and the exact Gaussian moments. The comparison remains conditional on the realized evaluation-start state in H; stationary labels used to construct the design do not define truth.

K.3 Fixed-sequence inheritance

Let $T_r = \{\mu_r^0 \le 0\}$. The event T_r, certificate coefficients, threshold $b_r(H)$, and the precommitted map $y \mapsto \mathbf{1}\{W_r(H,y) > b_r(H)\}$ are determined by $\mathcal{F}$. The realized outcome $\phi_r(H,Y)$ depends on future data and is measurable with respect to $\mathcal{F} \vee \sigma(Y)$, not generally $\mathcal{F}$. On T_r,

$$\mathbf{1}_{T_r}\mathbb{E}[\phi_r \mid \mathcal{F}] \le \mathbf{1}_{T_r}\Pr\{W_r - \mu_r^0 > b_r(H) \mid \mathcal{F}\} \le \alpha\mathbf{1}_{T_r}.$$

The first-true-null index is $\mathcal{F}$-measurable, so the inclusion from Section 4 gives conditional family false-promotion probability at most α, without independence among nodes. Missing or nonfinite bounds make the request unresolved. If a valid bound has $C_r = 0$ or $V_r = 0$, $W_r = \mu_r^0$ conditionally almost surely; the strict zero threshold then has zero rejection probability on the null. The positive-scale inequalities above are not evaluated by dividing by zero.

K.4 PREREGISTERED IMPLEMENTATION AUDIT

The fresh study fixes ten cells, eight outer histories per cell, and 4,096 independent future continuations per history. All 80 histories, including their selected label and exact history-specific truth, are retained before any inner continuation is generated. For each history with a finite first true null, the primary audit counts future continuations on which the fixed sequence rejects any true null and tests whether that conditional rejection probability exceeds .05. The 46 exact binomial audits are combined by Holm at family level .01. A separate deterministic check requires every false-promotion path to reject the first true null. Table 19 summarizes the retained histories, continuations, inclusion checks, Holm audit outcomes, and the single retained false promotion.

Table 19: Fresh preregistered conditional-certificate implementation audit. The family audit includes every history with a finite first true null. A nonrejection of the audit is not a proof of the theorem; the theorem supplies the guarantee under its stated structural assumptions.

Quantity	Value
Outer histories retained	80
Future continuations per history	4,096
Finite first-true-null histories	46
Invalid records	0
Recorded pathwise inclusion violations	0
Histories with a Holm excess-error finding ($\alpha_{\mathrm{fam}} = .01$)	0
Retained false promotions	1 in one negative history

One false promotion occurred in a single Gaussian-MSE negative history among its 4,096 continuations; it is retained. No pathwise inclusion violation occurred, no record was invalid, and the Holm audit flagged no history. These outcomes and the deterministic reconstruction support the explicitly checked aspects of implementation conformance within the two declared settings; they do not prove the probability inequalities above. The same preregistration also reports informativeness rather than filtering on it: the certificates are informative in the strong-effect cells but can be conservative or uninformative at moderate effects.

The complete ten-cell design, history-specific coefficients and additional analyses are in Appendix L and the code supplement. The original pathwise counts are execution summaries; no independent path-by-path replay is inferred from their zero value.

K.5 REAL-TASK APPLICABILITY BOUNDARY

The already observed real applications cannot serve as confirmatory evidence for a certificate chosen after their evaluation outcomes were known. In addition, the required conditional innovation law, propagation bounds, and raw-loss support or moment envelopes are not available for those applications. A separate prespecified metadata-only search over an initial prespecified set of candidate real tasks found no task for which all required structural certificates were available before outcomes; the search was not expanded after that zero-eligible-task result. Thus the new finite-sample guarantee is reported for the explicit certified synthetic settings only.

L RECONSTRUCTION, AUDIT SENSITIVITY, AND TECHNICAL POSITIONING

L.1 COMPLETE SYNTHETIC DESIGN AND CONDITIONAL TRUTH

Both experiments use $Y_t = \rho Y_{t-1} + \varepsilon_t$, $\rho = 1/2$, and $Y_0 = 0$. The bounded setting uses independent Rademacher innovations and raw MAE; the Gaussian setting uses independent standard normal innovations and raw MSE. Innovations through time 256 form the burn-in. Validation origins are exactly $256, \ldots, 319$, with horizons $\{1, 3, 6\}$. The last validation target is Y_{325}. The full history includes the complete path through that time, selection scores, selected sign, and all design information. Evaluation begins at origin 325; future innovations start at time 326. There are 512 bounded-setting origins and 256 Gaussian-setting origins, always at stride one, with equal weights over the three

horizons and all origins. Thus no evaluation target is used for selection, although origins and targets within evaluation overlap.

The MAE rules are $f_\delta(o,h) = \rho^h Y_o + \delta$, with candidates $\delta \in \{-1/10, +1/10\}$. The MSE rules are $f_\gamma(o,h) = (\rho^h + \gamma)Y_o$, with candidates $\gamma \in \{-.1, +.1\}$. Selection minimizes mean validation loss; an exact tie chooses the negative candidate. Each of the ten cells has eight retained histories and 4,096 future continuations per history. Table 20 specifies every comparator. Rational MAE values are used for boundary classification, not rounded decimals.

Table 20: Complete comparator design. Each row has eight histories. MAE truth is exact for both selected signs. Gaussian percentages are design labels only; the realized history-specific truth, not that label, determines each null. Gaussian comparator decimals are the original input values.

Loss and setting	Comparator parameter	Truth or nominal label
MAE, negative	$\delta_b = 0$	$\theta^0 = -1/1280$
MAE, boundary 2%	$\delta_b = 289/784$	$\theta^0 = 1/50$
MAE, boundary 5%	$\delta_b = 5875/10336$	$\theta^0 = 1/20$
MAE, 8%	$\delta_b = 2747/3680$	$\theta^0 = 2/25$
MAE, 40%	$\delta_b = 541/328$	$\theta^0 = 2/5$
MSE, negative	$\gamma_b = 0$	negative
MSE, nominal 2%	$\gamma_b = 0.16971626701134285$	2% nominal
MSE, nominal 5%	$\gamma_b = 0.24185434428325728$	5% nominal
MSE, nominal 8%	$\gamma_b = 0.30019949209541941$	8% nominal
MSE, nominal 40%	$\gamma_b = 0.79009575500905815$	40% nominal

For MAE, let $E_h = \sum_{j=1}^h \rho^{h-j}\varepsilon_j$. Exact enumeration of its 2^h equally probable values gives

$$L_\delta^0 = \frac{1}{3}\sum_{h\in\{1,3,6\}} 2^{-h} \sum_{e\in\{-1,1\}^h} \left|\sum_{j=1}^h \rho^{h-j}e_j - \delta\right|.$$

The term $\rho^h Y_o$ cancels from forecast error, so this risk is independent of the initial state. Symmetry makes the two selected signs equal in expected risk, although their observed losses differ. For Gaussian MSE, write $y_* = Y_{325}$ and number evaluation origins $o = 0, \ldots, n-1$. With

$$q_h = \frac{1-\rho^{2h}}{1-\rho^2}, \qquad m_o(H) = \rho^{2o}y_*^2 + \frac{1-\rho^{2o}}{1-\rho^2},$$

we have $L_\gamma^0(H) = (3n)^{-1}\sum_{o,h}\{q_h + \gamma^2 m_o(H)\}$ and $\theta^0(H) = 1 - L_s^0(H)/L_b^0(H)$. Here too the candidate signs have equal expected risk; the risk varies with H through y_*. These designs do not test selection among candidates with unequal conditional risks.

L.2 Explicit overlap coefficients and quadratic matrices

For bounded MAE, valid coefficients for the original paired contrast are

$$A_{i,r} = (2-r)\rho^{h_i}, \qquad B_{i,r} = 2-r, \qquad D_j = 2.$$

With $w_i = 1/(3n)$, substitution into the bounded construction gives

$$c_{j,r} = \frac{2(2-r)}{3n}\sum_{o=0}^{n-1}\sum_{h\in\{1,3,6\}}\left[\mathbf{1}_{\{j\le o\}}\rho^{o+h-j} + \mathbf{1}_{\{j\le o+h\}}\rho^{o+h-j}\right], \quad 1 \le j \le n+5.$$

The two terms respectively account for dependence through the forecast input and the target. This is a valid potentially loose bound, not a claim of minimal sensitivity.

For Gaussian MSE let $T = n+5$, $1 \le j \le T$, and define vectors

$$a_{o,j} = \mathbf{1}_{\{j\le o\}}\rho^{o-j}, \qquad z_{o,h,j} = \mathbf{1}_{\{o<j\le o+h\}}\rho^{o+h-j}.$$

The error of rule γ is $v_{\gamma,o,h}^{\top}\varepsilon + d_{\gamma,o}$, where $v_{\gamma,o,h} = z_{o,h} - \gamma a_o$ and $d_{\gamma,o} = -\gamma\rho^o y_*$. Therefore

$$A_r = \sum_{o,h} w_{o,h}\{(1-r)v_{b,o,h}v_{b,o,h}^{\top} - v_{s,o,h}v_{s,o,h}^{\top}\},$$
$$g_r = \sum_{o,h} w_{o,h}\{(1-r)d_{b,o}v_{b,o,h} - d_{s,o}v_{s,o,h}\},$$
$$k_r = \sum_{o,h} w_{o,h}\{(1-r)d_{b,o}^2 - d_{s,o}^2\}.$$

These formulas retain every shared innovation. A deterministic reconstruction separately calculates L_b^0, L_s^0 and $k_r + \mathrm{tr}(A_r)$, requiring agreement within $10^{-10}(1+|\mu_r^0|)$, and checks all stored thresholds and Gaussian variance values. The accompanying anonymous code supplement contains 240 history–threshold records and 123 coefficient files: three shared MAE files and 120 history-specific Gaussian files. The matrices are reconstructed from saved histories after the study, not mislabeled as original pre-execution exports.

L.3 Per-history audit, sensitivity, and secondary upper bounds

For each of the $m = 46$ histories having a true ladder null, let K_h be the number of false promotions in $M = 4096$ continuations. The original one-sided audit uses

$$p_h = \Pr\{\mathrm{Bin}(M, .05) \geq K_h\}$$

and Holm adjustment at family level $\eta = .01$ over all 46 histories. Forty-five histories have $K_h = 0$; one Gaussian negative history has $K_h = 1$. Their probabilities are respectively 1 and $1 - .95^{4096}$, the latter also rounding to 1 in double precision. All adjusted probabilities round to 1. Every one of the 80 histories, including the 34 without a true ladder null, is retained in the supplementary table; all denominators are 4,096.

At the first Holm step, the smallest count exceeding the rejection requirement is 257, or 6.2744% of the continuations. The single-history probability of crossing that first-step cutoff is $\Pr\{\mathrm{Bin}(4096, p) \geq 257\}$. It is 1.76%, 23.84%, 72.96%, and 99.90% at $p = .055, .060, .065, .075$, respectively. This characterizes the strictest Holm threshold, not the exact joint power for every multi-history alternative; later Holm steps use different cutoffs. Nonrejection alone is not evidence that every history has error at most .05.

As a separately labeled *post hoc descriptive analysis*, a one-sided Bonferroni Clopper–Pearson upper bound is

$$U_h = \mathrm{Beta}^{-1}\left(1 - \frac{.01}{46}; K_h + 1, M - K_h\right).$$

Conditional on the selected set of histories and the stated independent-continuation design, this fixed construction has simultaneous coverage at least 99% over that finite set. It yields $U_h = .00205692$ when $K_h = 0$ and .00266067 when $K_h = 1$. These bounds concern the 46 sampled histories only. Their later addition is not a preregistered equivalence analysis, does not select a new success margin, and does not certify arbitrary unobserved histories. It adds finite-grid information to the original failure-detection audit without replacing it.

Table 4 in the main text reports all ten cells.

The supplementary per-history table includes selected label, initial state, comparator and selected risks, exact MAE or computed Gaussian truth, first true null, terminal counts, numerator, denominator, raw and adjusted probabilities, and secondary upper bounds. A linked table provides all 240 means, thresholds, margins, and coefficient filenames. File-integrity records link these reconstructions to the saved histories and source. Reconstruction can cross-check false-promotion counts against terminal histograms and local counts, but cannot independently recover individual path decisions from aggregate counts. The zero pathwise violations are the original execution summary, not a subsequently observed replay result.

L.4 Deterministic reference check for the historical critical value

Let $Z = B(1)$ and $Q = \int_0^1 \{B(u) - uB(1)\}^2 du$. The Gaussian bridge is independent of Z. Its covariance eigenvalues give $Q \overset{d}{=} \sum_{k\geq 1} \xi_k^2/(\pi k)^2$, with independent standard normal ξ_k. Multiplying the normal-square Laplace transforms and using the product for $\sinh$ yields

$$\mathbb{E}[e^{-sQ}] = \left\{ \frac{\sqrt{2s}}{\sinh\sqrt{2s}} \right\}^{1/2}.$$

The angular Gaussian-tail identity $\overline{\Phi}(x) = \pi^{-1} \int_0^{\pi/2} \exp\{-x^2/(2\sin^2 t)\}\, dt$ follows by integrating the rotationally symmetric bivariate normal density over the half-plane beyond x. Thus, for $q > 0$,

$$\Pr\{Z/\sqrt{Q} > q\} = \frac{1}{\pi} \int_0^{\pi/2} \left\{ \frac{q/\sin t}{\sinh(q/\sin t)} \right\}^{1/2} dt.$$

Deterministic quadrature at 40 and 65 decimal working precision, cross-checked after the change of variable $u = \cot t$, gives an ideal continuous-reference 0.95 quantile 5.32267990053. The historical constant 5.2988297532 has ideal-reference upper-tail probability $.05057992335$, not exactly $.05$. An existing-evidence export now includes the historical generator source, its declared 200,000-path/grid/seed configuration, and the sorted 200,000-value terminal reference array; the array's linear empirical 0.95 quantile is 5.298829753159216. Raw Brownian increments and the original execution transcript were not retained, so the source-to-array execution trail remains partial. All 16 saved reached-node statistics recompute exactly and none lies in $[5.298829753159216, 5.322679900530264)$, so replacing only the cutoff would not alter those 16 recorded decisions. This does not address unreached nodes, reconstruct the original Monte Carlo uncertainty, or determine finite-sample error under dependent forecasting losses.

L.5 Closest-work comparison and demonstrated contribution

Broader forecast-evaluation context. Loss choice determines the forecast property being elicited, while time-aware evaluation governs the temporal target (Gneiting & Raftery, 2007; Hyndman & Koehler, 2006; Tashman, 2000; Bergmeir et al., 2018; Cerqueira et al., 2020). Forecast-comparison tests also depend on their own nulls and model relationships; nested-model comparisons, for example, require specialized predictive-accuracy theory (Clark & McCracken, 2001). Resampling and long-run-variance methods address dependence under different assumptions (Kunsch, 1989; Politis & Romano, 1994; Newey & West, 1987; Andrews, 1991; Kiefer & Vogelsang, 2002; 2005; Lobato, 2001). Self-normalization has its own asymptotic premises and extensions (Shao, 2010b;a; 2011; 2015; Wang & Shao, 2020). Table 21 compares the intended outputs of the closest reporting and inferential frameworks with the narrower evidence-to-conclusion role specified by UBOA.

Selection can bias validation estimates (Varma & Simon, 2006); adaptive reuse motivates holdout safeguards (Hardt & Ullman, 2014; Dwork et al., 2015); and multiplicity procedures target different error objects (Benjamini & Hochberg, 1995; Bailey & López de Prado, 2014; Edwards & Madsen, 2007; Dmitrienko & Tamhane, 2007; Dmitrienko et al., 2007; Huque & Alosh, 2008; Bretz et al., 2009). Distribution-free prediction and its conditional limits likewise show why a marginal guarantee cannot simply be relabeled as an arbitrary conditional one (Lei et al., 2018; Foygel Barber et al., 2021).

Benchmark evidence is sensitive to replication units, dataset shift, shortcuts, and implementation choices (Dietterich, 1998; Demsar, 2006; Varoquaux, 2018; Torralba & Efros, 2011; Geirhos et al., 2020; Henderson et al., 2018; Musgrave et al., 2020; Hand, 2006; Recht et al., 2019; Pineau et al., 2021; Kapoor & Narayanan, 2023). Forecasting competitions and modern architectures make precise realized comparisons especially important (Makridakis et al., 1982; Makridakis & Hibon, 2000; Makridakis et al., 2018; Lai et al., 2018; Salinas et al., 2020; Lim et al., 2021; Zhou et al., 2021; Wu et al., 2021; Zeng et al., 2023; Hewamalage et al., 2023). UBOA turns these concerns into an executable evidence-to-conclusion rule that can incorporate an appropriate local method.

L.6 Implementation-blind matcher contract challenge

A separate controlled challenge was authored from the public rule contract, formal appendix, published record format, and published representation conventions. The 140 cases (35 families, 70

Table 21: Related work by its intended output, not a claimed superiority ranking. Reporting frameworks are not converted into artificial statistical classifiers. The distinction concerns what the current paper specifies; practical accuracy or time savings over these systems have not been measured.

Work	Intended output	Relation to this paper
BetterBench (Reuel et al., 2024)	Assessment of benchmark development and usability practices	Supplies quality context; UBOA checks a named probability implication.
FactSheets (Arnold et al., 2019)	Service documentation and declarations of conformity	Complementary provenance; documentation alone does not instantiate a local conditional theorem.
Eval Factsheets (Bordes et al., 2025)	Structured descriptive evaluation taxonomy/questionnaire	Shares context and scope concerns; UBOA retains law, event, quantifier, and conditioning in each licensing rule.
Evaluation Cards (Ghosh et al., 2026)	Integrated metadata, interpretive signals, extraction/monitoring	Already operational, not merely a checklist; UBOA's narrower output is a typed statistical conclusion.
Conditional post-selection prediction (Leeb, 2009)	Conditional predictive distributions/intervals under a stated model	Demonstrates conditional post-selection inference, not a universal guarantee supplied by metadata.
Forecast comparison bands (Pohle et al., 2026)	Simultaneous expected-score and skill bands with bootstrap theory	Related relative-risk inference; its assumptions and target must match before its output can be licensed here.
Contractive-functional inequalities (Dedecker & Fan, 2015; Dedecker et al., 2019)	Martingale/deviation inequalities for separately Lipschitz path functionals	Technical basis, not a new result of this paper. Our explicit coefficients preserve forecast overlap and the selected mapping.
UBOA	Unchanged request plus licensed conclusion or explicit mismatch	Relative soundness assumes correct leaves; the controlled challenge measures contract conformance, not human-audit utility.

permissible and 70 impermissible) and their reference-label rationales were fixed before the checker source or outputs were read. The primary set contains no ambiguous cases. This construction is agent-authored rather than expert-adjudicated and therefore tests deterministic contract conformance, not end-user accuracy or time savings.

Table 3 reports the comparison in the main text.

The revised matcher rejects all 27 single-coordinate invalid cases and all other impermissible cases while licensing every permissible case. The original implementation's nine false accepts arise from validity-scope omissions in M1, missing observed-record provenance checks in M2, empty-premise theorem acceptance in M4, and one conclusion-type omission in M6. Coordinate equality misses side conditions and also rejects ten legal M5 coarsenings because the evidence conditioning intentionally differs from the requested coarser field. The revised input specification explicitly requires nonempty observed-record provenance and nonempty theorem premises. Strengthening the provenance contract intentionally reclassifies one historical observed-copy test case that lacked provenance; the original record is retained separately.

L.7 ALL RETAINED HISTORIES

Table 22: Complete retained-history audit index. The first-true-threshold column reports the threshold in percent; ∞ means no true ladder null. K is the false-promotion count, always out of 4,096. p_{adj} is the original Holm-adjusted probability; dashes denote inapplicability, not missing histories. U is the later, post hoc 99% simultaneous upper bound in percent. Initial states, selected parameters, all coefficients and thresholds are linked in the supplementary 80- and 240-row tables.

History	Setting	$\theta^0(H)$ (%)	First true threshold (%)	K	p_{adj}	U (%)
1	MAE negative	-0.07812	0	0	1.000	0.206
2	MAE negative	-0.07812	0	0	1.000	0.206
3	MAE negative	-0.07812	0	0	1.000	0.206
4	MAE negative	-0.07812	0	0	1.000	0.206
5	MAE negative	-0.07812	0	0	1.000	0.206
6	MAE negative	-0.07812	0	0	1.000	0.206
7	MAE negative	-0.07812	0	0	1.000	0.206
8	MAE negative	-0.07812	0	0	1.000	0.206
9	MAE boundary 2%	2.00000	2	0	1.000	0.206
10	MAE boundary 2%	2.00000	2	0	1.000	0.206
11	MAE boundary 2%	2.00000	2	0	1.000	0.206
12	MAE boundary 2%	2.00000	2	0	1.000	0.206
13	MAE boundary 2%	2.00000	2	0	1.000	0.206
14	MAE boundary 2%	2.00000	2	0	1.000	0.206
15	MAE boundary 2%	2.00000	2	0	1.000	0.206
16	MAE boundary 2%	2.00000	2	0	1.000	0.206
17	MAE boundary 5%	5.00000	5	0	1.000	0.206
18	MAE boundary 5%	5.00000	5	0	1.000	0.206
19	MAE boundary 5%	5.00000	5	0	1.000	0.206
20	MAE boundary 5%	5.00000	5	0	1.000	0.206
21	MAE boundary 5%	5.00000	5	0	1.000	0.206
22	MAE boundary 5%	5.00000	5	0	1.000	0.206
23	MAE boundary 5%	5.00000	5	0	1.000	0.206
24	MAE boundary 5%	5.00000	5	0	1.000	0.206
25	MAE 8%	8.00000	∞	0	—	—
26	MAE 8%	8.00000	∞	0	—	—
27	MAE 8%	8.00000	∞	0	—	—
28	MAE 8%	8.00000	∞	0	—	—
29	MAE 8%	8.00000	∞	0	—	—
30	MAE 8%	8.00000	∞	0	—	—
31	MAE 8%	8.00000	∞	0	—	—
32	MAE 8%	8.00000	∞	0	—	—
33	MAE 40%	40.00000	∞	0	—	—
34	MAE 40%	40.00000	∞	0	—	—
35	MAE 40%	40.00000	∞	0	—	—
36	MAE 40%	40.00000	∞	0	—	—
37	MAE 40%	40.00000	∞	0	—	—
38	MAE 40%	40.00000	∞	0	—	—
39	MAE 40%	40.00000	∞	0	—	—
40	MAE 40%	40.00000	∞	0	—	—
41	MSE negative	-1.09161	0	0	1.000	0.206
42	MSE negative	-1.09446	0	0	1.000	0.206
43	MSE negative	-1.09207	0	0	1.000	0.206
44	MSE negative	-1.09186	0	0	1.000	0.206
45	MSE negative	-1.09697	0	1	1.000	0.266
46	MSE negative	-1.10764	0	0	1.000	0.206
47	MSE negative	-1.09472	0	0	1.000	0.206
48	MSE negative	-1.09375	0	0	1.000	0.206

Table 22 continued

History	Setting	$\theta^0(H)$ (%)	First true threshold (%)	K	p_{adj}	U (%)
49	MSE nominal 2%	2.00965	5	0	1.000	0.206
50	MSE nominal 2%	2.00649	5	0	1.000	0.206
51	MSE nominal 2%	1.98993	2	0	1.000	0.206
52	MSE nominal 2%	1.98990	2	0	1.000	0.206
53	MSE nominal 2%	1.99732	2	0	1.000	0.206
54	MSE nominal 2%	2.00876	5	0	1.000	0.206
55	MSE nominal 2%	2.01118	5	0	1.000	0.206
56	MSE nominal 2%	1.99825	2	0	1.000	0.206
57	MSE nominal 5%	5.04336	∞	0	—	—
58	MSE nominal 5%	4.97817	5	0	1.000	0.206
59	MSE nominal 5%	5.00259	∞	0	—	—
60	MSE nominal 5%	4.98827	5	0	1.000	0.206
61	MSE nominal 5%	4.97554	5	0	1.000	0.206
62	MSE nominal 5%	4.99824	5	0	1.000	0.206
63	MSE nominal 5%	4.98930	5	0	1.000	0.206
64	MSE nominal 5%	4.97598	5	0	1.000	0.206
65	MSE nominal 8%	7.96436	∞	0	—	—
66	MSE nominal 8%	7.98800	∞	0	—	—
67	MSE nominal 8%	8.10622	∞	0	—	—
68	MSE nominal 8%	7.96455	∞	0	—	—
69	MSE nominal 8%	8.24099	∞	0	—	—
70	MSE nominal 8%	7.99897	∞	0	—	—
71	MSE nominal 8%	7.97359	∞	0	—	—
72	MSE nominal 8%	7.96207	∞	0	—	—
73	MSE nominal 40%	39.92879	∞	0	—	—
74	MSE nominal 40%	39.90442	∞	0	—	—
75	MSE nominal 40%	40.47902	∞	0	—	—
76	MSE nominal 40%	39.87622	∞	0	—	—
77	MSE nominal 40%	40.08294	∞	0	—	—
78	MSE nominal 40%	40.14961	∞	0	—	—
79	MSE nominal 40%	39.87616	∞	0	—	—
80	MSE nominal 40%	40.20560	∞	0	—	—