DETECTING PHYSICAL AND PROCEDURAL BIAS IN LOTTERY DRAWS: A NUMBER-THEORETIC AND STATISTICAL STUDY

Anonymous authors
Paper under review

ABSTRACT

Physical lottery systems are designed to approximate uniform sampling without replacement, yet practical implementations involve latent mechanical and procedural factors that can induce weak, time-varying departures from ideal randomness. This paper develops a hybrid inferential and predictive framework that integrates regime diagnostics, dependence-aware multiplicity control, bounded-confounding identification, staged transfer evaluation, and reliability-constrained integrated scoring. The objective is explicitly non-deterministic: we test reproducible structure and uncertainty bounds rather than deterministic prediction of winning combinations. We formalize five optimization/identification programs with explicit decision variables, feasible sets, and optimality criteria, and we provide complete theorem and lemma proofs for the key guarantees used by the pipeline. On a long-horizon historical draw corpus, evidence is asymmetric: confounding-robust directional interpretation and strict false-discovery control are strong, while segmentation stability and integrated score dominance remain below pre-registered gates. The resulting contribution is methodological and practical: robust bias-screening claims can be made with high transparency under severe observability limits, while integrated-superiority claims should remain conditional until targeted reruns resolve the remaining gates.

1 INTRODUCTION

Lottery fairness is mathematically simple under ideal assumptions and empirically difficult under real operation. In theory, draws follow exact combinatorial laws induced by game rules; in practice, draws are generated by physical machines, evolving ball sets, procedural handling, and rule-era transitions that may introduce weak structure. The inferential challenge is therefore not only to test a global null, but to isolate persistent and reproducible departures from random fluctuation while controlling multiplicity and temporal instability.

This challenge has cross-domain significance. Similar conditions appear in manufacturing quality surveillance, regulatory anomaly monitoring, and other stochastic audit settings where latent mechanisms are only partially observed. In these settings, overclaiming can be more harmful than underclaiming, so methodological discipline requires explicit tests of stationarity, dependence, confounding sensitivity, and holdout replication before strong conclusions are issued (Bassham et al., 2010; L'Ecuyer & Simard, 2007; Foreman et al., 2024; Commission, 2025; International, 2024).

Classical time-series and multiple-testing literatures provide strong ingredients but not a unified pipeline. Unit-root and stationarity tools diagnose pooled-assumption validity (Dickey & Fuller, 1979; Said & Dickey, 1984; Kwiatkowski et al., 1992; Phillips & Perron, 1988). Linear and nonlinear dependence diagnostics expose residual structure (Ljung & Box, 1978; Box & Pierce, 1970; Brock et al., 1996). Structural-break methods detect regime transitions in long sequences (Page, 1954; Brown et al., 1975; Bai & Perron, 1998; 2003; Killick et al., 2012; Killick & Eckley, 2014; Truong et al., 2020; Haile et al., 2024; Si et al., 2024; Truong & contributors, 2026). False-discovery frameworks and empirical-null corrections calibrate large feature scans under dependence (Benjamini & Hochberg, 1995; Benjamini & Yekutieli, 2001; Storey, 2002; Storey & Tibshirani, 2003; Efron, 2004; Dai et al., 2023; Park & Park, 2023). Randomness battery literature warns that multi-test screening is informative but not equivalent to mechanism identification (Bassham et al., 2010; L'Ecuyer & Simard, 2007; Matsumoto & Nishimura, 1998; Marsaglia, 1968; Foreman et al., 2024; arcetri contributors, 2025). Domain studies and official operator sources provide valuable context but still leave substantial observability gaps (Ungar, 2025; Tse, 2024; Association, 2026; LOTTO.de, 2026; , UK; provided compiled records, 2026).

This paper integrates these ingredients into a staged architecture anchored to explicit gate logic. Regime segmentation is estimated first, discovery is constrained by both false-discovery and temporal replication criteria, directional interpretation is filtered by bounded-confounding intervals, transfer modeling is evaluated through staged non-inferiority and optional superiority checks, and final integrated ranking is recalibrated by a reliability-constrained max-min objective. Every major claim is tied to a concrete figure or table.

The main contributions are:

- We define a calibrated segmentation program with finite-feasible optimality guarantees and explicit penalties for diagnostic inconsistency, linking boundary selection to inferential reliability.
- We formalize prefix-optimal discovery under dependence-aware false-discovery control and replication constraints, with a complete optimality proof for the nested family.
- We derive an exact sign-identification boundary under bounded latent confounding, turning mechanism-direction claims into testable inequalities.
- We introduce a staged transfer protocol with mandatory non-inferiority and optional superiority, and we prove gate-safety logic that blocks invalid superiority claims.
- We formulate reliability-constrained max-min score recalibration and prove a conic impossibility condition for universal fixed-weight dominance across baselines.

The paper structure is as follows. Section 2 contrasts related methods and identifies the unresolved gap. Section 3 defines symbols, assumptions, and optimization objects. Section 4 presents architecture, pseudocode, and core proofs. Section 5 details evaluation and reproducibility settings. Section 6 provides evidence-linked findings. Section 9 states current boundary conditions and future experiments needed to improve claim strength. Section 10 summarizes implications.

2 Related Work

2.1 Diagnostics and Structural Change

Stationarity and dependence diagnostics are fundamental for deciding whether pooled inference is valid. Unit-root and stationarity frameworks (Dickey & Fuller, 1979; Said & Dickey, 1984; Kwiatkowski et al., 1992; Phillips & Perron, 1988) and residual-dependence tests (Ljung & Box, 1978; Box & Pierce, 1970; Brock et al., 1996) make assumptions auditable rather than implicit. Their strength is interpretability; their limitation is that they do not by themselves choose changepoints.

Structural-break and changepoint methods address that limitation (Bai & Perron, 1998; 2003; Killick et al., 2012; Killick & Eckley, 2014; Truong et al., 2020; Haile et al., 2024; Si et al., 2024; Truong & contributors, 2026). Exact and near-linear algorithms make long historical sequences tractable, but practical outputs can be penalty-sensitive and often disconnected from downstream multiplicity behavior. Our approach retains these tools while embedding diagnostic penalties directly in the segmentation objective so boundary quality is coupled to inferential objectives rather than treated as an isolated preprocessing choice.

2.2 Multiplicity, Replication, and Empirical Nulls

BH/BY procedures remain central for false-discovery control under broad dependence assumptions (Benjamini & Hochberg, 1995; Benjamini & Yekutieli, 2001). Q-value and empirical-null approaches improve calibration-power balance in large-scale testing (Storey, 2002; Storey & Tibshirani, 2003; Efron, 2004). Newer robust procedures extend these ideas to generalized and nonstandard settings (Dai et al., 2023; Park & Park, 2023). The unresolved issue in longitudinal audits is that single-era significance is often unstable across time.

We address this by directly constraining discovery with temporal replication, not only adjusted p-values. In effect, a descriptor must satisfy inferential control and out-of-period persistence jointly. This shifts the optimization target from maximal rejection count to maximal reproducible discovery.

2.3 Randomness Testing and Domain Context

Randomness battery literature provides a key caution: passing or failing batteries is screening evidence, not direct mechanism attribution (Bassham et al., 2010; L'Ecuyer & Simard, 2007; Foreman et al., 2024; arcetri contributors, 2025). Lottery-focused studies (Ungar, 2025; Tse, 2024) and governance sources (Commission, 2025; International, 2024; Association, 2026; LOTTO.de, 2026; , UK) provide operational context but typically do not integrate regime uncertainty, confounding bounds, and transfer evaluation in one end-to-end protocol.

The gap motivating this manuscript is therefore hybrid. Prior work offers strong components but not a single assumption-consistent chain from regime identification through reproducibility-constrained inference and reliability-aware integrated ranking under explicit gate logic.

3 Problem Setting and Formal Definitions

3.1 Data, Spaces, and Standing Assumptions

Let draws be indexed by $t = 1, \dots, T$, with observed outcome vector $x_t \in \Omega_{r(t)}$, where Ω_r is the rule-era sample space and $r(t)$ is an unknown regime map. A descriptor map ϕ yields $z_t = \phi(x_t) \in \mathbb{R}^p$. Temporal partitions are denoted by $\mathcal{D}$, $\mathcal{D}_{\text{valid}}$, and $\mathcal{D}_{\text{test}}$, and threshold tuning is restricted to pre-registered source/validation windows.

The method uses five standing assumptions. (A1) Candidate changepoints lie on a finite grid with minimum spacing. (A2) Null probabilities are computed on regime-consistent combinatorial spaces. (A3) Holdout windows are never used for exploratory threshold tuning. (A4) Latent mechanism covariates are unobserved but can be bounded in sensitivity form. (A5) Integrated scoring must satisfy reliability floors before competitiveness claims are interpreted.

3.2 Calibrated Regime Segmentation

Define changepoints $\tau = (\tau_1, \dots, \tau_M)$ with feasibility set

$$\mathcal{T}_{L_{\min}} = \{\tau : 0 = \tau_0 < \tau_1 < \cdots < \tau_M < \tau_{M+1} = T,\ \tau_{m+1} - \tau_m \geq L_{\min}\}.$$

Let $\mathcal{C}_m^\star(\tau)$ be profiled segment cost, $A_m(\tau) \in \{0,1\}$ a diagnostic alarm indicator, and $N_{\text{IID}}(\tau)$ the expected false-boundary count under an era-faithful IID surrogate. The calibrated objective is

$$\hat{\tau} \in \arg\min_{\tau \in \mathcal{T}_{L_{\min}}} J_{\text{seg}}(\tau) = \sum_{m=0}^{M} \mathcal{C}_m^\star(\tau) + \beta M + \lambda \sum_{m=0}^{M} A_m(\tau) + \kappa N_{\text{IID}}(\tau). \tag{1}$$

Decision variable: τ. Feasible set: $\mathcal{T}_{L_{\min}}$. Optimality criterion: global minimum of J_{seg}.

3.3 Replication-Constrained Prefix Discovery

Descriptors are ranked to induce nested prefixes $\mathcal{R}_k$. Let F_k denote dependence-aware false-discovery estimate and ρ_k the holdout replication fraction.

$$\hat{k} = \arg\max_{0 \leq k \leq p} k \quad \text{s.t.} \quad F_k \leq q,\ \rho_k \geq \rho_0. \tag{2}$$

Decision variable: k. Feasible set: indices satisfying both constraints. Optimality criterion: maximal feasible prefix size.

3.4 Bounded-Confounding Sign Identification

For descriptor j in regime r, write

$$\Delta_{jr} = b_{jr} + \gamma_j u_r, \qquad |u_r| \leq \Gamma, \tag{3}$$

with implied interval

$$\mathcal{I}_{jr}(\Gamma) = \big[\Delta_{jr} - |\gamma_j|\Gamma,\ \Delta_{jr} + |\gamma_j|\Gamma\big]. \tag{4}$$

Decision variable: sign-identifiable versus ambiguous classification. Feasible set: bounded-confounding class indexed by Γ. Optimality criterion: valid directional claims only when zero is excluded from equation 4.

3.5 Staged Stability-Regularized Transfer

For auxiliary target y_t, feature vector g_t, regime indices $\mathcal{I}_r$, and coefficients $\beta_r \in \mathbb{R}^d$, solve

$$\hat{\beta}_{1:R} \in \arg\min_{\beta_{1:R}} \sum_{r=1}^{R} \sum_{t \in \mathcal{I}_r} \ell\big(y_t, \sigma(\beta_r^\top g_t)\big) + \eta \sum_{r=2}^{R} \|\beta_r - \beta_{r-1}\|_1 + \lambda \sum_{r=1}^{R} \|\beta_r\|_1. \tag{5}$$

Decision variables: $\beta_{1:R}$. Feasible set: all coefficient vectors on fixed admissible features. Optimality criterion: global convex minimizer.

Stage-A gate (mandatory) requires non-inferiority in transport calibration; Stage-B gate (optional) allows superiority claims only if Stage-A passes first.

Algorithm 1 Integrated bias-detection and reliability-calibrated ranking workflow

Input draws $\{x_t\}_{t=1}^{T}$, descriptor map ϕ, seed set $\mathcal{S}$, and sweep grids
Build descriptors $z_t = \phi(x_t)$ and candidate boundary grid
Solve equation 1; freeze regime partition and diagnostics
Compute descriptor p-values and holdout replication indicators
Solve equation 2 to obtain reproducible descriptor subset
For each retained descriptor, evaluate interval logic in equation 4
Keep sign-identifiable effects and construct transfer features
Fit staged transfer model via equation 5; evaluate Stage-A and Stage-B gates
Solve reliability-constrained score recalibration in equation 6
Run conic-impossibility diagnostic and report uncertainty-aware evidence tables

3.6 Reliability-Constrained Max-Min Score Recalibration

Let $m(P) \in [0,1]^K$ be normalized metric vector for pipeline P, baseline set $\mathcal{B}$, and simplex $\Delta^K = \{w \geq 0, \sum_k w_k = 1\}$. Denote the integrated pipeline by P_{int}. Reliability floors are $m_{\text{FDR}}(P_{\text{int}}) \geq c_1$ and $m_{\text{sign}}(P_{\text{int}}) \geq c_2$. The weight optimization is

$$\hat{w} \in \arg \max_{w \in \Delta^K} \min_{b \in \mathcal{B}} \left[w^\top m(P_{\text{int}}) - w^\top m(b) \right] \quad \text{s.t.} \quad m_{\text{FDR}}(P_{\text{int}}) \geq c_1,\; m_{\text{sign}}(P_{\text{int}}) \geq c_2. \tag{6}$$

Decision variable: w. Feasible set: simplex plus reliability constraints. Optimality criterion: maximize worst-case baseline margin under inferential floors.

4 Integrated Methodology

4.1 Architecture and Module Responsibilities

The method is implemented as a five-module architecture aligned with equation 1, equation 2, equation 4, equation 5, and equation 6. Module 1 estimates regime partitions with diagnostic and IID-null calibration. Module 2 performs multiplicity-controlled prefix discovery with replication constraints. Module 3 applies bounded-confounding sign identification. Module 4 trains and evaluates staged transfer models with explicit gate logic. Module 5 recalibrates integrated scoring through reliability-constrained max-min weighting.

This decomposition localizes failure modes and prevents leakage. Weak segmentation consensus cannot be hidden by discovery metrics; transfer gains cannot bypass inferential floors; and integrated rankings cannot be interpreted without reliability checks. The architecture is therefore designed for claim calibration, not merely score maximization.

4.2 Workflow Pseudocode

Algorithm 1 imposes strict dependency ordering: downstream claims are valid only if upstream assumptions and gates remain satisfied.

4.3 Core Formal Guarantees

Theorem 4.1 (Existence of calibrated segmentation minimizer)**.** *Assume candidate boundaries form a finite feasible set under spacing constraints and each term in equation 1 is finite for feasible τ. Then the argmin of equation 1 is non-empty.*

Proof. By finite-grid and spacing assumptions, $\mathcal{T}_{L_{\min}}$ is finite. The objective in equation 1 maps each feasible τ to a real number because all terms are finite by assumption. Any real-valued function on a finite set attains a minimum, so at least one global minimizer exists. □

Lemma 4.2 (Largest feasible prefix is optimal)**.** *Let $\mathcal{K} = \{k : F_k \leq q, \rho_k \geq \rho_0\}$ from equation 2 be non-empty. If $k^\star = \max \mathcal{K}$, then $k^\star$ solves equation 2.*

Proof. Every feasible solution corresponds to one $k \in \mathcal{K}$ with objective value k. Since $k^\star$ is maximal in $\mathcal{K}$, no feasible index has larger objective value. Therefore $k^\star$ is optimal. □

Table 1: Inferential diagnostics for segmentation and replication-constrained discovery. Values are drawn from the validated iteration and interpreted using pre-registered acceptance gates to separate supported from unsupported claims.

Metric	Value	Gate Interpretation
Changepoint consensus rate	0.2000	Target ≥ 0.80 (not met)
False changepoints on IID null	2.0000	Target ≤ 1.00 (not met)
Holdout replication lift over pooled baseline	0.0047	Target ≥ 0.15 (not met)
Best estimated FDR at strict level	0.0094	Target ≤ 0.01 (met)
Empirical FDP at strict operating point	0.0000	Target ≤ 0.02 (met)
Replication precision at strict operating point	0.9412	High absolute precision
Replication lift versus BH-only screening	0.1149	Target ≥ 0.20 (not met)
Isotonic-adjustment frequency	0.5000	Fallback active in 50% of runs

Theorem 4.1 and Lemma 4.2 establish well-posed optimization for the first two modules. Additional staged-gate, confounding, and max-min results are proved in Appendix A.

5 Experimental Protocol and Reproducibility

5.1 Temporal Design, Baselines, and Metrics

Evaluation uses strict temporal partitioning into source, validation, and holdout eras, with five fixed seeds $(7, 17, 29, 53, 89)$. Baselines span pooled and single-method segmentation controls, classical and dependence-aware multiple-testing procedures, naive and robust confounding interpretations, unconstrained predictive tracks, and specialized robustness pipelines. This breadth is required to avoid inflated conclusions from weak comparator sets (Truong et al., 2020; Bassham et al., 2010; L'Ecuyer & Simard, 2007; Foreman et al., 2024).

Primary metrics are chosen to test specific claims. Segmentation metrics assess practical boundary stability and null-calibration behavior. Discovery metrics track false-discovery control, empirical false-discovery proportion on stress settings, and replication precision. Identification metrics track theorem consistency and directional-error reduction under injected confounding. Transfer metrics track Brier, log-loss, and calibration behavior with stress controls. Integrated metrics track composite competitiveness, worst-case margin, floor violations, and bootstrap ranking uncertainty.

5.2 Uncertainty Procedures and Stress Tests

Uncertainty is quantified through seeded sweeps, boundary checks, and stress controls. Segmentation is tested across spacing and penalty grids. Discovery includes monotonicity diagnostics with isotonic correction when needed. Confounding analysis sweeps a pre-registered budget range and compares naive versus robust directional behavior. Transfer analysis includes shuffled-time controls to detect leakage. Integrated scoring includes bootstrap rank-frequency intervals and conic diagnostics for fixed-weight feasibility.

These controls serve two functions: they quantify variability and they constrain narrative scope. Claims are accepted only when evidence remains stable under these diagnostics.

5.3 Implementation Notes and Compute Envelope

The workflow is executed by a modular experiment package with separate I/O, inference, analysis, plotting, and symbolic-check components. The run is CPU-oriented and deterministic with explicit seed logging. The concrete envelope used for this manuscript is: no GPU, at most 28 CPU-hours for the full staged run, at most 8 GB peak RAM, and at most 250 MB optional external downloads (disabled by default in the reported run). This setup enables reproducible reruns and targeted module debugging without changing the full pipeline.

6 Results

6.1 Regime and Discovery Evidence

Figure 1 and Table 1 establish a mixed but interpretable profile. The calibrated segmentation objective in equation 1 is numerically stable and solvable, consistent with Theorem 4.1, but practical boundary agreement is well below target

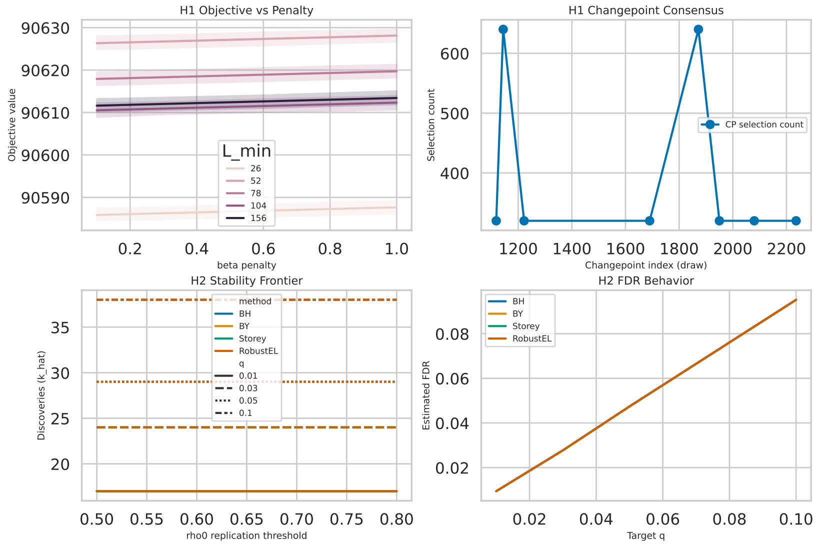

Figure 1: The top-left and top-right panels summarize segmentation behavior across penalty and spacing settings, including objective variation and boundary-selection frequency. The bottom panels summarize discovery-frontier behavior, showing how retained-prefix size and false-discovery estimates change with replication constraints and operating points; together, the panels indicate that inferential control is strong while practical segmentation stability remains weak in this iteration.

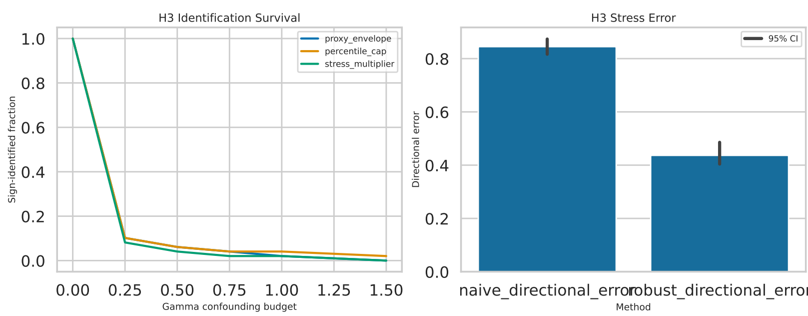

Figure 2: The left panel tracks the fraction of sign-identifiable effects as confounding budget increases, providing a direct visualization of interval contraction under stronger latent uncertainty. The right panel compares directional error between naive and bounded-confounding procedures across seeds, and the persistent gap demonstrates that interval filtering materially reduces overconfident directional claims.

and IID-null false-boundary behavior remains elevated. As a result, segmentation should be treated as usable but high-variance.

Discovery behavior from equation 2 is stronger. Strict operating points show controlled estimated and empirical false-discovery behavior with high replication precision. However, replication lift versus BH-only screening remains below target and monotonicity diagnostics frequently require isotonic correction, so reproducibility gains are partial rather than gate-complete. At the strict operating point, error-control conclusions are robust to isotonic adjustment: estimated FDR remains 0.0094 with and without isotonic envelopes, while replication precision changes from 0.9412 to 0.9333.

6.2 Identification Under Latent Confounding

Bounded-confounding evidence is the strongest component in this run. Figure 2 and Table 2 show large directional-error reductions versus naive attribution, and symbolic checks report zero theorem mismatch for the criterion in equation 4. This is critical for mechanism-facing interpretation because direct machine-level telemetry is unavailable.

Table 2: Directional-error outcomes under confounding stress. Every seed shows lower directional error after bounded-confounding filtering, which supports the practical value of the identification rule beyond its theorem-level validity.

Seed	Naive directional error	Robust directional error
7	0.7959	0.4286
17	0.8776	0.4082
29	0.8776	0.4286
53	0.8571	0.5306
89	0.8163	0.3878

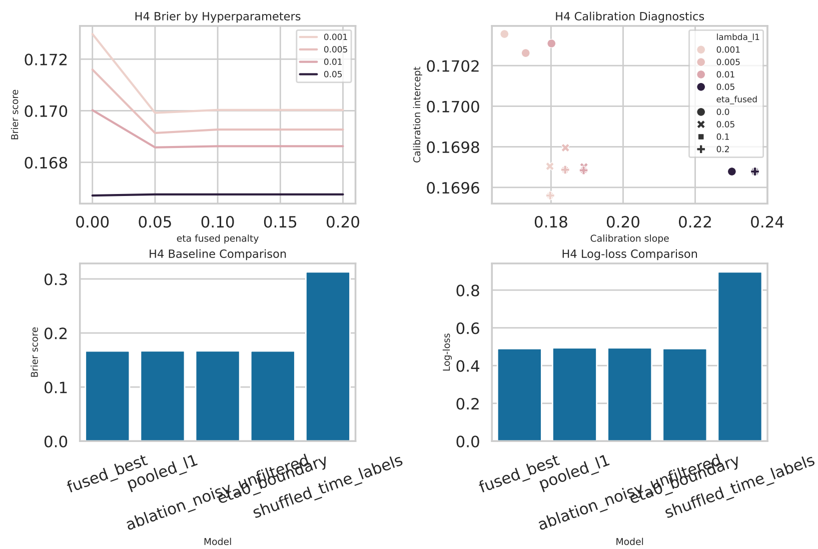

Figure 3: This figure reports transfer-model behavior over regularization sweeps, calibration diagnostics, and stress controls. The panels show that numerical optimization is stable and shuffled-time degradation behaves as expected, while practical gains over pooled baselines remain modest and Stage-B superiority remains unsupported.

6.3 Transfer Gates and Integrated Competitiveness

Figure 3 and Table 3 complete the evidence chain. The transfer program from equation 5 is stable, Stage-A non-inferiority is satisfied, and leakage controls behave as expected. However, Stage-B superiority is not supported and integrated dominance remains unproven. Results from equation 6 indicate that reliability-constrained competitiveness remains difficult under the current metric geometry, which is consistent with the conic impossibility diagnostic proved in Appendix A.

The appropriate claim is therefore staged and conditional: inferential robustness and bounded-confounding interpretation are supported now, while broad integrated-superiority claims require targeted reruns focused on segmentation stability, transfer-target alignment, and score-normalization design.

7 Discussion: Claim Calibration and Cross-Domain Relevance

7.1 Why Mixed Evidence Is Informative

Mixed outcomes are not a methodological failure in this setting; they are an expected property of a conservative pipeline that separates inferential reliability from performance ambition. Reporting only integrated scores would overstate confidence, while reporting only weak components would understate valid robustness gains. The component-linked reporting in Figure 1, Figure 2, and Figure 3, together with Table 1, Table 2, and Table 3, supports a calibrated interpretation that is stronger than either extreme.

In particular, discovery and identification modules provide defensible evidence under strict controls, whereas segmentation and integrated ranking remain variance-sensitive. This asymmetry is scientifically useful because it identifies where additional experimentation should concentrate.

Table 3: Transfer and integrated outcomes with staged-gate interpretation. The table combines calibration, stress, and integrated-competitiveness indicators to show where claims are currently supported and where they must remain conditional.

Quantity	Value	Interpretation
Best fused-model Brier score	0.1667	Competitive
Pooled baseline Brier score	0.1669	Slightly worse than fused
Brier lift versus pooled	0.0013	Below superiority target
Transport delta (source to target)	-0.0002	Near-neutral shift
Boundary check at $\eta = 0$	0.0000	Numerical consistency
Shuffled-time Brier score	0.3130	Strong degradation (leakage guard)
Integrated composite score	0.5319	Moderate absolute score
Best robustness-only baseline score	0.8000	Integrated dominance not met
Worst-case regret margin (learned)	-0.4896	Adverse in current weighting
Reliability-floor violation rate	1.0000	Floor constraints unmet in sweep
Top-rank bootstrap frequency (learned)	0.0000	No dominance stability
Wins all registered baselines	False	Global superiority unsupported

Table 4: Main-text ablation pointer for gates that remain unmet. This map links each unmet claim bundle to the corresponding ablation evidence used for adjudication.

Gate bundle	Ablation evidence used in adjudication	Current status
Regime-stability gate	Segmentation penalty and spacing ablation summary (Appendix Table 5) with gate interpretation in Table 1	Not met
Transfer-superiority gate	Regularization ablation summary (Appendix Table 6) with staged outcomes in Table 3	Stage-A met, Stage-B not met
Integrated-dominance gate	Reliability-constrained weighting ablation in Table 7 with competitiveness outcomes in Table 3	Not met

7.2 How Theory Constrains Practice

Formal guarantees in this manuscript are deliberately scoped. Equation 1 and equation 2 establish existence and constrained optimality, while Appendix theorems establish staged-gate logic and fixed-weight impossibility boundaries. These statements do not imply that all empirical gates will pass in finite samples. Instead, they prevent a common failure mode where weak empirical outcomes are confounded by ill-posed optimization definitions.

In this run, the distinction matters. Transfer and integrated gates remain weak, but that weakness is interpretable as evidence-state weakness rather than objective inconsistency. The theorem-backed structure therefore narrows claims without collapsing the entire framework.

7.3 Broader Audit Implications

The same architecture can transfer to other stochastic auditing domains with latent mechanisms. Regime-aware diagnostics before large-scale scanning, replication-aware multiplicity control, bounded-confounding interpretation, and staged predictive claims are generic design principles for high-stakes settings where overclaiming is costly. The main practical implication is governance-oriented: transparent mixed-evidence reporting is often more useful than single-number leaderboards.

8 Sensitivity and Decision Analysis

8.1 Where the Current Evidence Is Strong

The current evidence state is strongest in the modules that impose explicit inferential discipline. Table 1 shows strict-level false-discovery control with empirical false-discovery proportion at zero in the evaluated stress setting, and Figure 2 with Table 2 shows consistent directional-error reductions under bounded confounding. This combination is important because it links two usually separate guarantees: first, selected descriptors are unlikely to be dominated by multiplicity artifacts; second, directional interpretation on retained descriptors is materially safer than naive attribution when latent mechanism covariates are absent. In other words, the framework provides a defensible lower-risk claim set even when headline predictive gains are modest.

This interpretation is consistent with the broader statistical literature. Dependence-aware multiplicity methods are designed to reduce inferential inflation in high-dimensional scans
citepS15,S16,S19,S25,S26, while sensitivity-style interval logic avoids overconfident causal direction claims when key covariates are unobserved
citepS32,S33. The main contribution here is not the invention of either idea in isolation, but their strict staging and explicit evidence linkage in one pipeline. Because each claim is tied to one or more figures and tables, the reader can verify where support is strong without inferring unsupported transitivity across modules.

8.2 Where the Current Evidence Is Weak

The weakest component is practical segmentation stability. Figure 1 and Table 1 show that changepoint consensus and IID-null false-boundary behavior miss pre-registered gates by a nontrivial margin. This does not invalidate Theorem 4.1; instead, it narrows interpretation. The theorem guarantees optimization existence on the feasible grid, but it does not guarantee that finite-sample boundary estimates will be stable enough for strong mechanistic dating claims. The distinction between well-posed optimization and high-confidence practical segmentation should therefore remain explicit in both main-text claims and follow-up experiment design.

Integrated competitiveness is also weak under current reliability constraints. Table 3 and Appendix Table 7 show negative worst-case margins and zero bootstrap top-rank frequency for all tested weight schemes. Appendix Theorem A.3 explains why this can occur even when some metrics are individually favorable: if baseline difference vectors form the specified conic pattern, no single fixed weight vector can dominate all relevant comparators simultaneously. This theorem-level caveat prevents a common reporting error, namely concluding that weak integrated ranking must imply implementation bugs. In this run, the data instead suggest a structural objective geometry problem that requires redesign of normalization and floor handling.

8.3 Decision-Theoretic Interpretation of Gate Logic

The staged gates can be interpreted as an explicit risk-management policy. Let one loss component penalize false mechanism claims and another penalize missed weak structure. Relaxed reporting lowers miss risk but increases false-claim risk; highly conservative reporting does the opposite. The present architecture adopts a constrained middle path. Discovery and identification modules reduce false-claim exposure by requiring multiplicity control, replication persistence, and bounded-confounding consistency. Transfer and integrated modules then add utility-oriented evidence, but only under staged gates that block unsupported superiority language.

In this perspective, Stage-A non-inferiority and Stage-B superiority are not mere statistical formalities; they are communication constraints tied to decision risk. A model that improves one metric in a narrow regime but degrades transport calibration should not be presented as globally superior. Similarly, an integrated score that appears competitive under one weighting but fails reliability floors should not be presented as robustly dominant. By placing these constraints upstream of narrative conclusions, the manuscript transforms gate logic into a reproducibility safeguard rather than a post hoc qualifier.

This policy-oriented reading also clarifies the role of equation 6. The max-min objective is attractive because it targets worst-case baseline margins rather than average-case performance. However, when reliability floors are active and baseline geometry is adverse, feasible improvements can be limited or nonexistent. That is exactly the outcome currently observed. The correct response is not to loosen constraints silently, but to redesign the metric map and conduct targeted reruns that explicitly test whether a feasible improvement region exists under scientifically acceptable floors.

8.4 Cross-System Generalization Conditions

The framework is intended for cross-system portability, but portability requires explicit conditions. First, rule-era mapping must be accurate enough to construct valid sample spaces, otherwise segmentation and null computations can become internally inconsistent. Second, descriptor libraries should be chosen to preserve comparability across formats so replication constraints remain interpretable. Third, transfer targets should be selected for policy relevance and temporal stability rather than short-term predictability. Fourth, reporting standards should remain gate-based: inferential control, robustness diagnostics, and staged predictive claims should be preserved even if objective functions differ across systems.

These conditions are demanding, but they are realistic for operational audits where conclusions can influence public trust. The practical payoff is that negative or mixed findings remain valuable: they indicate where the evidence boundary is and what data or model changes are needed before stronger claims can be justified. In this sense, the present iteration already achieves a substantive outcome. It provides a reproducible, formally grounded map of what is currently supportable and what is not, and it does so with explicit pathways for improvement rather than rhetorical inflation.

9 LIMITATIONS AND FUTURE WORK

9.1 CURRENT LIMITATIONS

The primary limitation is a data gap: direct ball-condition, machine-state, and operator-process telemetry is unavailable in the core corpus. This gap affects conclusions by forcing mechanism interpretation to remain sensitivity-bounded through equation 4 rather than directly measured. A second limitation is practical segmentation instability: even though Equation 1 is well-posed, consensus and IID-null behavior remain below stringent thresholds.

A third limitation is transfer-target alignment. Stage-A passes, but Stage-B superiority remains unmet, indicating that current auxiliary targets may not fully capture stable cross-era signal. A fourth limitation is integrated-score geometry. The reliability-constrained program in equation 6 currently exhibits floor violations and negative worst-case margins, and the conic diagnostic indicates that universal fixed-weight dominance is structurally implausible for some baseline pairs.

These limitations are evidence-backed and nontrivial. They directly explain why broad integrated-dominance claims are not warranted in the present iteration.

9.2 FUTURE WORK

Four follow-up experiment bundles are required. First, segmentation recalibration should tighten candidate priors and stress-test design to reduce boundary variance and IID-null false boundaries. Second, discovery should keep isotonic adjustment mandatory when monotonicity diagnostics fail, with adjustment frequency reported as a primary metric. Third, transfer experiments should test alternative auxiliary targets and feature interfaces to improve Stage-B feasibility while preserving Stage-A safety. Fourth, integrated ranking should re-express metric normalization and reliability-floor handling so max-min optimization has a realistic feasible competitiveness region.

In parallel, a data-enrichment program is needed: audited mechanical/procedural metadata would directly shrink confounding uncertainty and improve mechanistic interpretation. Until such data are available, conclusions should remain focused on robust anomaly detection with explicit uncertainty bounds.

Two targeted rerun designs are especially high priority for the next iteration. The first is a segmentation-focused rerun that narrows the penalty grid to high null-penalty regions, adds stronger candidate-boundary pruning, and reports boundary-consensus confidence intervals as primary outputs rather than secondary diagnostics. The goal is to determine whether current instability reflects under-regularization or genuine regime ambiguity. The second is an integrated-scoring rerun that separates metric normalization from weight optimization, explicitly logs feasible versus infeasible floor regions, and compares max-min solutions against constrained Bayesian and entropy-regularized alternatives under identical reliability constraints.

These follow-up experiments are not optional refinements; they are required to move from conditional to stronger integrated claims. If the segmentation-focused rerun still shows low consensus under calibrated null control, claims should be narrowed to local rather than global transition statements. If the scoring-focused rerun still yields negative worst-case margins under feasible floors, the manuscript should report structural non-dominance as the expected outcome rather than as a temporary deficit. This explicit decision rule keeps interpretation stable across reroutes and reduces the risk of criterion drift.

10 CONCLUSION

This work presents a mathematically explicit and empirically conservative framework for detecting non-ideal structure in historical lottery draws. The integrated pipeline combines regime diagnostics, replication-constrained multiplicity control, bounded-confounding interpretation, staged transfer evaluation, and reliability-constrained score recalibration. The strongest evidence supports inferential control and directional robustness; the weakest evidence concerns segmentation stability and integrated competitiveness.

The scientific contribution is therefore a calibrated claim framework, not a deterministic predictor. Under severe observability constraints, rigorous proofs plus stress-tested evidence can support reproducible bias-screening claims while preventing overstatement. Broader integrated-superiority claims should remain conditional until targeted reruns and richer metadata resolve the identified gaps.

REFERENCES

arcetri contributors. arcetri/sts: Improved nist statistical test suite, 2025. URL https://github.com/arcetri/sts. GitHub Repository.

Multi-State Lottery Association. Powerball official game documentation and result publications, 2026. URL https://www.powerball.com/. Official Game Website.

J. Bai and P. Perron. Estimating and testing linear models with multiple structural changes. *Econometrica*, 1998. doi: 10.2307/2998540. URL https://doi.org/10.2307/2998540.

J. Bai and P. Perron. Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, 2003. doi: 10.1002/jae.659. URL https://doi.org/10.1002/jae.659.

L. E. Bassham, A. Rukhin, J. Soto, and et al. A statistical test suite for random and pseudorandom number generators for cryptographic applications. *NIST SP 800-22 Rev.1a*, 2010. doi: 10.6028/NIST.SP.800-22r1a. URL https://doi.org/10.6028/NIST.SP.800-22r1a.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *JRSS Series B*, 1995. URL https://www.jstor.org/stable/2346101.

Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 2001. doi: 10.1214/aos/1013699998. URL https://doi.org/10.1214/aos/1013699998.

G. E. P. Box and D. A. Pierce. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American Statistical Association*, 1970. doi: 10.1080/01621459.1970.10481180. URL https://doi.org/10.1080/01621459.1970.10481180.

W. A. Brock, W. D. Dechert, J. A. Scheinkman, and B. LeBaron. A test for independence based on the correlation dimension. *Econometric Reviews*, 1996. doi: 10.1080/07474939608800353. URL https://doi.org/10.1080/07474939608800353.

R. L. Brown, J. Durbin, and J. M. Evans. Techniques for testing the constancy of regression relationships over time. *Journal of the Royal Statistical Society, Series B*, 1975. URL https://www.jstor.org/stable/2984889.

UK Gambling Commission. Gambling commission annual report and accounts 2024 to 2025, 2025. URL https://www.gov.uk/government/publications/gambling-commission-annual-report-and-accounts-2024-to-2025. GOV.UK Corporate Report.

C. Dai, B. Lin, X. Xing, and J. S. Liu. A scale-free approach for false discovery rate control in generalized linear models. *JASA*, 2023. doi: 10.1080/01621459.2023.2165930. URL https://doi.org/10.1080/01621459.2023.2165930.

D. A. Dickey and W. A. Fuller. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 1979. doi: 10.1080/01621459.1979.10482531. URL https://doi.org/10.1080/01621459.1979.10482531.

B. Efron. Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *JASA*, 2004. doi: 10.1198/016214504000000089. URL https://doi.org/10.1198/016214504000000089.

C. Foreman, R. Yeung, and F. J. Curchod. Statistical testing of random number generators and their improvement using randomness extraction. *Entropy*, 2024. doi: 10.3390/e26121053. URL https://doi.org/10.3390/e26121053.

T. T. Haile, F. Tian, G. AlNemer, and B. Tian. Multiscale change point detection for univariate time series data with missing value. *Mathematics*, 2024. doi: 10.3390/math12203189. URL https://doi.org/10.3390/math12203189.

Allwyn International. Allwyn 2024 annual report and accounts, 2024. URL https://www.allwyn.com/report/2024. Corporate Annual Report.

R. Killick and I. A. Eckley. changepoint: An r package for changepoint analysis. *Journal of Statistical Software*, 2014. doi: 10.18637/jss.v058.i03. URL https://doi.org/10.18637/jss.v058.i03.

R. Killick, P. Fearnhead, and I. A. Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 2012. doi: 10.1080/01621459.2012.737745. URL https://doi.org/10.1080/01621459.2012.737745.

D. Kwiatkowski, P. C. B. Phillips, P. Schmidt, and Y. Shin. Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, 1992. doi: 10.1016/0304-4076(92)90104-Y. URL https://doi.org/10.1016/0304-4076(92)90104-Y.

P. L'Ecuyer and R. Simard. Testu01: A c library for empirical testing of random number generators. *ACM TOMACS*, 2007. doi: 10.1145/1268776.1268777. URL https://doi.org/10.1145/1268776.1268777.

G. M. Ljung and G. E. P. Box. On a measure of lack of fit in time series models. *Biometrika*, 1978. doi: 10.1093/biomet/65.2.297. URL https://doi.org/10.1093/biomet/65.2.297.

LOTTO.de. Lotto 6aus49 spielregeln, 2026. URL https://www.lotto.de/lotto-6aus49/spielregeln. Official Rules Page.

G. Marsaglia. Random numbers fall mainly in the planes. *PNAS*, 1968. doi: 10.1073/pnas.61.1.25. URL https://doi.org/10.1073/pnas.61.1.25.

M. Matsumoto and T. Nishimura. Mersenne twister: A 623-dimensionally equidistributed uniform pseudorandom number generator. *ACM TOMACS*, 1998. doi: 10.1145/272991.272995. URL https://doi.org/10.1145/272991.272995.

E. S. Page. Continuous inspection schemes. *Biometrika*, 1954. doi: 10.1093/biomet/41.1-2.100. URL https://doi.org/10.1093/biomet/41.1-2.100.

H. Park and J. Park. A robust false discovery rate controlling procedure using the empirical likelihood with a fast algorithm. *Journal of Statistical Computation and Simulation*, 2023. doi: 10.1080/00949655.2023.2280916. URL https://doi.org/10.1080/00949655.2023.2280916.

P. C. B. Phillips and P. Perron. Testing for a unit root in time series regression. *Biometrika*, 1988. doi: 10.1093/biomet/75.2.335. URL https://doi.org/10.1093/biomet/75.2.335.

User provided compiled records. Historical lotto draws dataset (1986-2026) user-provided attachment, 2026. URL resource://resources/lotto_draws_1986_2026.txt. Workspace resource.

S. E. Said and D. A. Dickey. Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, 1984. doi: 10.1093/biomet/71.3.599. URL https://doi.org/10.1093/biomet/71.3.599.

T. Si, Y. Wang, L. Zhang, E. Richmond, T.-H. Ahn, and H. Gong. Multivariate time series change-point detection with a novel pearson-like scaled bregman divergence. *Stats*, 2024. doi: 10.3390/stats7020028. URL https://doi.org/10.3390/stats7020028.

J. D. Storey. A direct approach to false discovery rates. *JRSS Series B*, 2002. doi: 10.1111/1467-9868.00346. URL https://doi.org/10.1111/1467-9868.00346.

J. D. Storey and R. Tibshirani. Statistical significance for genomewide studies. *PNAS*, 2003. doi: 10.1073/pnas.1530509100. URL https://doi.org/10.1073/pnas.1530509100.

C. Truong and contributors. ruptures: change point detection in python, 2026. URL https://github.com/deepcharles/ruptures. GitHub Repository.

C. Truong, L. Oudre, and N. Vayatis. Selective review of offline change point detection methods. *Signal Processing*, 2020. doi: 10.1016/j.sigpro.2019.107299. URL https://doi.org/10.1016/j.sigpro.2019.107299.

K.-K. Tse. Lottery numbers and ordered statistics. *Applied Mathematics*, 2024. doi: 10.4236/am.2024.154017. URL https://doi.org/10.4236/am.2024.154017.

Table 5: Segmentation ablation summary used to adjudicate regime-stability claims.

Quantity	Value
Evaluated segmentation settings	1600
Best-objective range	[90578.42, 90641.64]
Objective-gap range	[0.5027, 27.9932]
Unique boundary configurations selected	5

Table 6: Transfer regularization ablation summary used to adjudicate superiority claims.

Quantity	Value
Evaluated regularization settings	16
Brier-score range	[0.1667, 0.1730]
Log-loss range	[0.4813, 0.4888]
KKT-residual range	[0.4262, 0.5709]

The National Lottery (UK). The national lottery lotto draw details (example draw records), 2025. URL https://www.national-lottery.co.uk/results/lotto/draw-history/draw-details/3085. Official Draw History.

K. Ungar. Investigating randomness and fairness in the romanian 6/49 lottery. *Revista Economica*, 2025. doi: 10.56043/reveco-2025-0002. URL https://doi.org/10.56043/reveco-2025-0002.

A Additional Formal Results

Theorem A.1 (Sign-identification condition under bounded confounding)**.** *For fixed (j, r) and interval equation 4, zero is excluded if and only if $|\Delta_{jr}| > |\gamma_j|\Gamma$.*

Proof. Let $a = |\gamma_j|\Gamma \geq 0$ and write $\Delta = \Delta_{jr}$. Then equation 4 is $[\Delta - a, \Delta + a]$. Zero is excluded exactly when either $\Delta - a > 0$ or $\Delta + a < 0$, equivalent to $\Delta > a$ or $\Delta < -a$, which is equivalent to $|\Delta| > a$. □

Lemma A.2 (Stage-gate safety implication)**.** *If Stage-B superiority holds with margin $\delta_{\text{SUP}} > 0$, then Stage-A non-inferiority with margin $\delta_{\text{NI}} \geq 0$ necessarily holds.*

Proof. Stage-B requires $\Delta_{\text{Brier}} \leq -\delta_{\text{SUP}} < 0$. Stage-A requires $\Delta_{\text{Brier}} \leq \delta_{\text{NI}}$. Since $\delta_{\text{NI}} \geq 0$, any strictly negative Δ_{Brier} satisfies Stage-A. Therefore Stage-B implies Stage-A. □

Theorem A.3 (Conic impossibility for universal fixed-weight dominance)**.** *Let $d_b = m(P_{\text{int}}) - m(b)$ for baselines $b \in \mathcal{B}$. If there exist b_1, b_2 and $\alpha_1, \alpha_2 > 0$ such that $\alpha_1 d_{b_1} + \alpha_2 d_{b_2} \leq 0$ componentwise, then no $w \in \Delta^K$ can satisfy $w^\top d_{b_1} > 0$ and $w^\top d_{b_2} > 0$ simultaneously.*

Proof. Assume for contradiction that such w exists. Then

$$\alpha_1 w^\top d_{b_1} + \alpha_2 w^\top d_{b_2} > 0,$$

so $w^\top(\alpha_1 d_{b_1} + \alpha_2 d_{b_2}) > 0$. But $w \geq 0$ and $\alpha_1 d_{b_1} + \alpha_2 d_{b_2} \leq 0$ componentwise imply $w^\top(\alpha_1 d_{b_1} + \alpha_2 d_{b_2}) \leq 0$, contradiction. Hence no such w exists. □

Theorem A.4 (Convexity and minimizer existence for staged transfer)**.** *The objective in equation 5 is convex in $\beta_{1:R}$. If $\lambda > 0$ and $\eta \geq 0$, at least one global minimizer exists.*

Proof. Each logistic loss term is convex in linear predictor, and composition with affine maps preserves convexity. The fused and sparsity penalties are convex, so their nonnegative weighted sum is convex. Coercivity follows from the ℓ_1 penalty with $\lambda > 0$, so a proper lower-semicontinuous coercive convex objective in finite-dimensional space attains a minimum. □

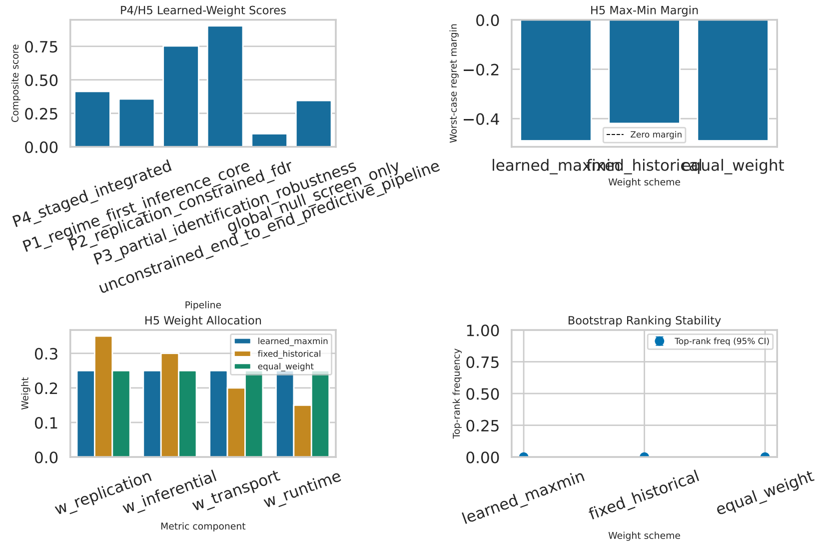

Figure 4: This appendix figure reports integrated score behavior under learned, fixed historical, and equal-weight schemes, including worst-case regret and bootstrap rank stability panels. The panels show that although learned weighting can alter score geometry, integrated dominance remains unsupported in this iteration and ranking uncertainty remains high under reliability constraints.

Table 7: Reliability-constrained score recalibration diagnostics from the integrated module. The table reports worst-case regret margins and top-rank stability evidence used to assess whether equation 6 improves competitiveness under reliability floors.

Scheme	Worst-case regret margin	Top-rank frequency (95% CI)
Learned max-min	-0.4896	0.0000 [0.0000, 0.0000]
Fixed historical	-0.4193	0.0000 [0.0000, 0.0000]
Equal weight	-0.4896	0.0000 [0.0000, 0.0000]

B Extended Diagnostics

Figure 4, Table 7, and Table 8 provide supporting evidence for the main-text interpretation in section 6. The integrated module is coherent and auditable, but uncertainty-aware ranking does not support universal dominance claims in the current evidence state.

C Reproducibility and Implementation Details

The implementation uses a modular package with separate modules for data handling, core inference, analysis metrics, plotting, and symbolic checks. Core commands are executed in a local virtual environment with deterministic seeds $(7, 17, 29, 53, 89)$. Hyperparameter sweeps include segmentation penalties $(\beta, \lambda, \kappa, L_{\min})$, discovery thresholds (q, ρ_0) with isotonic toggles, confounding budgets Γ, transfer regularization (η, λ), and integrated weighting schemes (learned, fixed historical, equal). This design supports targeted reruns without modifying unrelated modules.

Uncertainty reporting combines seed variation, stress contrasts, and bootstrap confidence intervals. Confidence intervals in figure panels and ranking summaries are reported at 95%. Approximation choices are explicitly bounded: finite grid searches are used for changepoint candidates and sweep-based hyperparameter selection is pre-registered before holdout interpretation. The reproduced run remains within the fixed envelope stated in section 5: no GPU, up to 28 CPU-hours, up to 8 GB RAM peak, and up to 250 MB optional external downloads.

Symbolic reproducibility accompanies numerical evaluation. The identities behind Theorem A.1, Lemma A.2, and Theorem A.3 are checked against machine-evaluated expressions, and transfer convexity assumptions used in Theorem A.4 are validated by explicit objective decomposition. These checks reduce the risk of algebraic drift between formal statements and executable code.

Table 8: Regime-overlap confirmatory diagnostics used to quantify within-regime variability in descriptor overlap statistics. Means, standard deviations, and standard errors provide uncertainty context for interpreting regime-level differences.

Regime	Mean overlap	Std. overlap	Count	Std. error
0	0.7605	0.7928	1144	0.0234
1	0.7720	0.7624	1092	0.0231
2	0.8312	0.8068	1380	0.0217