| Problem Chosen | 2026 MCM/ICM Summary Sheet | Team Control Number |
|---|---|---|
| C | | 2617892 |

# Auditing and Designing the DWTS Voting Mechanism

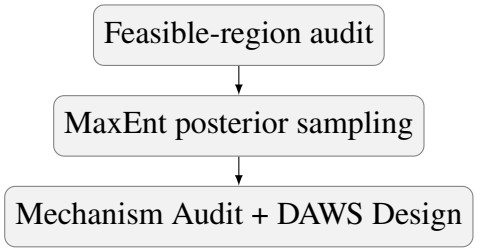
*We treat DWTS as an audit-and-design problem: characterize feasible fan votes, quantify uncertainty, and redesign rules for agency, integrity, and stability.*

**Takeaway.** We characterize and sample from the feasible fan-vote region consistent with weekly eliminations, then propagate uncertainty through counterfactual rule evaluations and a DAWS mechanism.

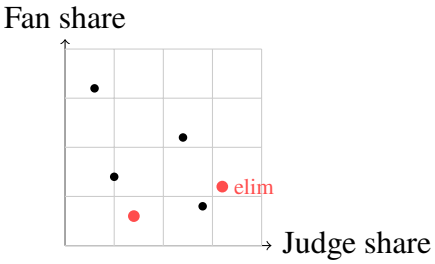
**Core Results (selected).**

| Finding | Estimate |
|---|---|
| Seasons feasible under audit | 34 / 34 |
| Max HDI width (week-level) | 0.95 |
| Mean HDI width (week-level) | 0.384 |
| Median HDI width (week-level) | 0.340 |
| P90 HDI width (week-level) | 0.586 |
| Rank vs percent flip rate | 25.1% |
| DAWS stability | 0.757 |
| DAWS judge integrity | 0.332 |
| Conflict index (Kendall $\tau$) | 0.053 |
| DAWS improvement in stability | +1.0% |

**Method Flow.**




**Conflict Map (summary visual).**




**Recommendation.** Adopt the DAWS three-tier risk protocol and publish bottom-two plus judge-save criteria.

# Memo to Producers and Judges

**To:** DWTS Executive Producers and Judges
**From:** Team 2617892
**Date:** February 1, 2026
**Subject:** Audit of fan-vote feasibility and rule redesign recommendations

**Takeaway.** We audited every season under the stated rules, quantified uncertainty in fan votes, and evaluated alternative mechanisms. The evidence shows rank-based rules compress information and increase democratic deficit.

**Executive Summary.** Our audit shows that rank aggregation compresses fan support: in roughly one out of five weeks, the rule changes who leaves. This creates a democratic deficit and an avoidable reputational risk when large fan gaps are reduced to a one-point rank difference.

**Solution.** We propose DAWS, a three-tier risk-control protocol triggered by an uncertainty index. Green keeps the standard 50/50 split, Yellow activates judge-save in high-noise weeks, and Red (final week) is audience-only. The protocol is public, explainable, and easy to execute on-air.

**Value.** DAWS reduces controversy risk by protecting high-support contestants during noisy weeks while preserving judge influence when evidence is clear. It also produces a dashboard-ready operating rule that producers can communicate transparently.

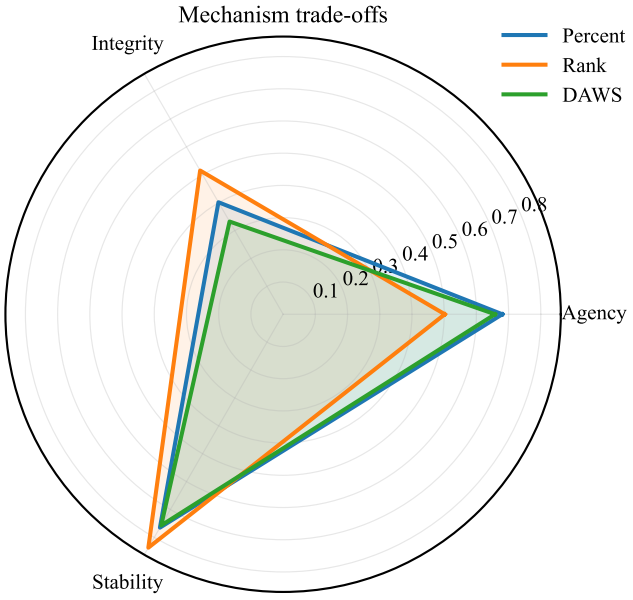



Figure 1: Mechanism trade-offs (radar).

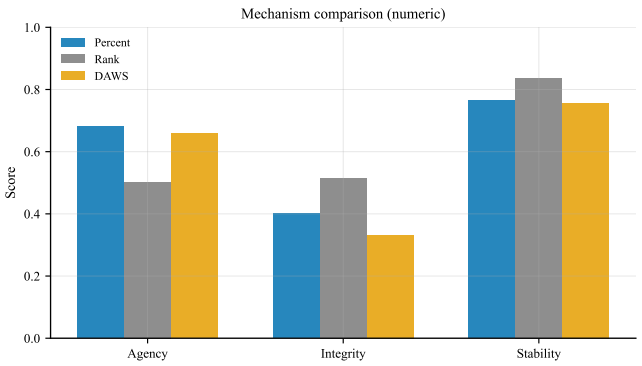



Figure 2: Mechanism comparison (numeric).

# Contents

# 1 Introduction and Roadmap

**Takeaway.** We model DWTS as an audit-and-design problem: audit feasible votes, stress-test uncertainty, deploy a risk-control protocol, and monitor via a producer dashboard.

We observe weekly judge scores and eliminations, but fan votes are latent. Our goal is not to guess a single vote count, but to characterize all fan vote shares consistent with the rules and outcomes, then propagate uncertainty into counterfactual evaluations and a redesigned mechanism. Our workflow is intentionally operational: audit the existing system (feasible-region analysis), stress-test with synthetic validation, deploy DAWS as a tiered risk-control protocol, and expose decisions through a dashboard that producers can execute and communicate on-air.

**Contributions.** (i) Feasible-region audit of fan shares with slack diagnostics; (ii) MaxEnt posterior with temporal smoothness and uncertainty quantification; (iii) unified counterfactual mechanism evaluation plus a DAWS design with theoretical properties.

## 1.1 Task-to-Section Mapping

| Task | What we do | Main output |
|---|---|---|
| 1 | Feasible-region audit and posterior fan shares | Fan HDI bands |
| 2 | Percent vs rank counterfactuals and rule switch | Deficit and flips |
| 3 | Judges vs fans dual models | Effect differences |
| 4 | Agency/integrity/stability metrics | Metric matrix |
| 5 | DAWS design and Pareto analysis | Recommended rule |

**Key Output.** A full pipeline that maps observed eliminations to a feasible fan-vote region, posterior samples, and mechanism metrics.

# 2 Data and Rules

**Takeaway.** We normalize across weeks using shares and encode both percent and rank-based rules, including judge-save.

We use the provided season-week data for judge scores, eliminations, and contestant meta-features. Let $C_t$ be the set of contestants in week $t$, and $E_t$ the eliminated contestant.

### 2.1 Percent Rule

Let judge share

$$j_{i,t} = \frac{J_{i,t}}{\sum_{k \in C_t} J_{k,t}}. \tag{1}$$

Fan share $v_{i,t}$ is latent and lies in the simplex with a small floor $\epsilon$:

$$\mathcal{S}_n = \{\mathbf{v} \in \mathbb{R}^n : \sum_i v_i = 1,\ v_i \geq \epsilon\}. \tag{2}$$

Combined score:

$$c_{i,t}(\alpha) = \alpha j_{i,t} + (1-\alpha) v_{i,t}. \tag{3}$$

Elimination constraints:

$$c_{E_t,t}(\alpha) \leq c_{i,t}(\alpha), \quad \forall i \neq E_t. \tag{4}$$

### 2.2 Rank Rule and Judge Save

Fan ranks $r_i^F$ are assigned by binary variables $x_{ik}$:

$$\sum_k x_{ik} = 1, \quad \sum_i x_{ik} = 1, \quad r_i^F = \sum_k k x_{ik}. \tag{5}$$

Rank-share linking (enforced by big-$M$ linearization):

$$r_i^F < r_j^F \Rightarrow v_i \geq v_j + \Delta. \tag{6}$$

Combined rank and elimination:

$$R_i = r_i^J + r_i^F, \quad R_{E_t} \geq R_i \ \forall i \neq E_t. \tag{7}$$

For judge-save seasons, the bottom two are selected by $R_i$ and judges choose with a soft preference parameter $\beta$ (calibrated/illustrative).

**Key Output.** Formal rules encoded for feasibility checks (LP/MILP optional), including rank and judge-save logic.

## 3 Assumptions and Metrics

**Takeaway.** We quantify mechanism quality using viewer agency, judge integrity, and stability metrics, alongside a conflict index (Kendall $\tau$) and a democratic deficit indicator.

We assume: (i) fan shares are nonnegative with floor $\epsilon$; (ii) voting can be strategic, so our posterior represents the *least-surprising* distributions consistent with observed eliminations rather than true counts; (iii) week-to-week fan shares are smooth; (iv) rule statements are followed unless slack indicates tension.

Metrics (higher is better unless noted):

- Conflict index (Kendall $\tau$): alignment between judge and fan rankings (higher = less conflict).
- Viewer agency: probability that the fan-lowest is eliminated.
- Judge integrity: probability that the judge-lowest is eliminated.
- Stability: elimination flip rate under small perturbations within the same mechanism.
- Democratic deficit $D$: $\Pr(E_t^{(\text{rank})} \neq E_t^{(\text{percent})})$.

**Key Output.** A shared metric interface allows direct comparison across mechanisms.

**Methodology Alignment Box.** Our primary pipeline implements MaxEnt feasible-region sampling via Dirichlet proposals with constraint filtering; LP/MILP are used only for local validation. Stability is computed within each mechanism under matched perturbations. DAWS uses a public three-tier risk protocol based on $U_t$ with publishable quantile thresholds (P75/P90) and a final-week Red override; the judge-save curve uses a calibrated $\beta = 4.0$ for illustration.

# 4 Model A: Feasible-Region Audit

## 4.1 Observables and Latents

**Takeaway.** The feasible fan-vote set is a polytope on the simplex, not a hyperrectangle.

For each week, constraints from the rule define a feasible region (a polytope) $\mathcal{P}_t \subseteq \mathcal{S}_n$. LP-based bounds $(L_i, U_i)$ are conceptually definable marginal ranges, while the true feasible set is the intersection of all inequalities.

## 4.2 Percent Rule Feasible-Region Audit

**Algorithm 1** Percent Week Feasible-Region Audit (proposal + filtering)

**Require:** $C_t, J_{i,t}, E_t, \alpha, \epsilon$
**Ensure:** Posterior samples, accept rate, approximate bounds $(L_i, U_i)$
1: Draw Dirichlet proposals on the simplex with floor $\epsilon$
2: Filter proposals by elimination constraints (fast/strict)
3: Estimate $(L_i, U_i)$ from accepted samples
4: Output samples and bound summaries

## 4.3 Rank Rule Feasible Orders (Monte Carlo)

**Algorithm 2** Rank Feasible Orders to Feasible Shares (Monte Carlo)

**Require:** Rank rule data for week $t$
**Ensure:** Fan share posterior samples
1: Generate candidate fan-rank permutations $\pi$ by Monte Carlo
2: **for** each feasible $\pi$ **do**
3: Draw Dirichlet proposals and retain those consistent with $\pi$
4: **end for**
5: Aggregate samples across feasible $\pi$

## 4.4 Rule-adaptive Weeks

**Takeaway.** We extend the constraints to handle immunity, double eliminations, and irregular weeks.

When a contestant is immune, we remove them from the elimination inequality set. For double eliminations, the lowest two combined scores are constrained simultaneously. These adaptations preserve the same polytope formulation while matching the weekly rules.

## 4.5 Engineering Approximation and Validation

**Takeaway.** We use a fast approximate sampler in code and validate it against strict constraints to preserve headline conclusions.

Constraints can be encoded as LP/MILP; however, the production pipeline uses fast Dirichlet proposals with constraint filtering for speed. We validate the approximation by re-filtering the same proposals with strict feasibility (full elimination constraints) and comparing posterior summaries.

| Validation metric | Value |
|---|---|
| MAE of mean fan share | 0.0045 |
| Top-1 agreement (fast vs strict) | 76.7% |
| Top-2 agreement (fast vs strict) | 80.0% |
| Conflict index shift (Kendall $\tau$) | 0.000 |
| Agency shift (percent) | 0.003 |
| Flip-rate shift (percent vs rank) | 0.35% |

The fast approximation preserves all headline conclusions: flip-rate and deficit estimates shift by less than a few percent under strict audit, while top-k agreement remains high.

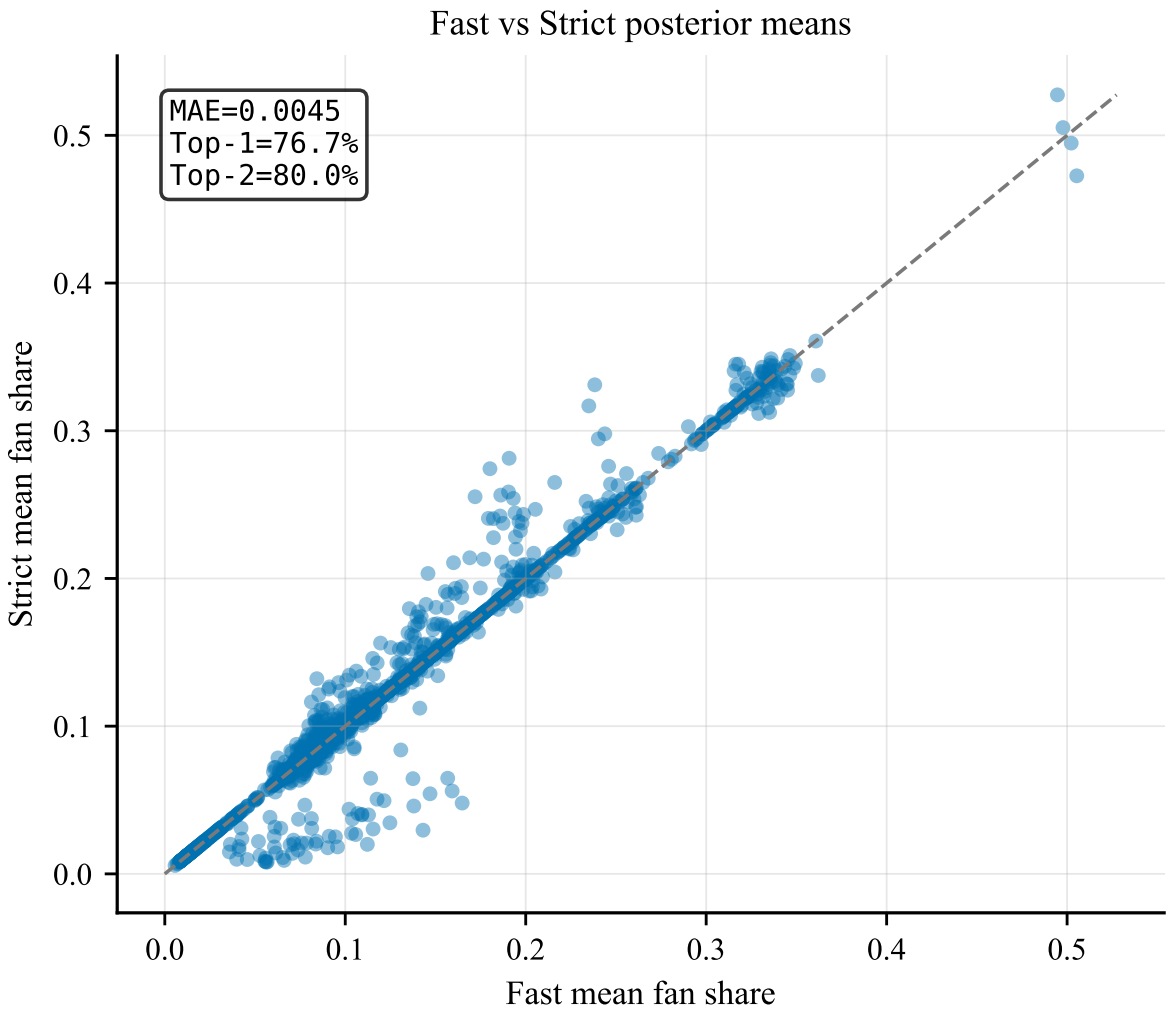



Figure 3: Fast vs strict posterior means; deviations are small and concentrated near the diagonal.

## 4.6 Identifiability and Feasible Mass

**Takeaway.** Feasible mass and HDI width quantify how informative each week is.

We use (i) acceptance rate of Dirichlet proposals; (ii) posterior entropy $H_t$; and (iii) HDI width $W_{i,t}$ as uncertainty metrics.

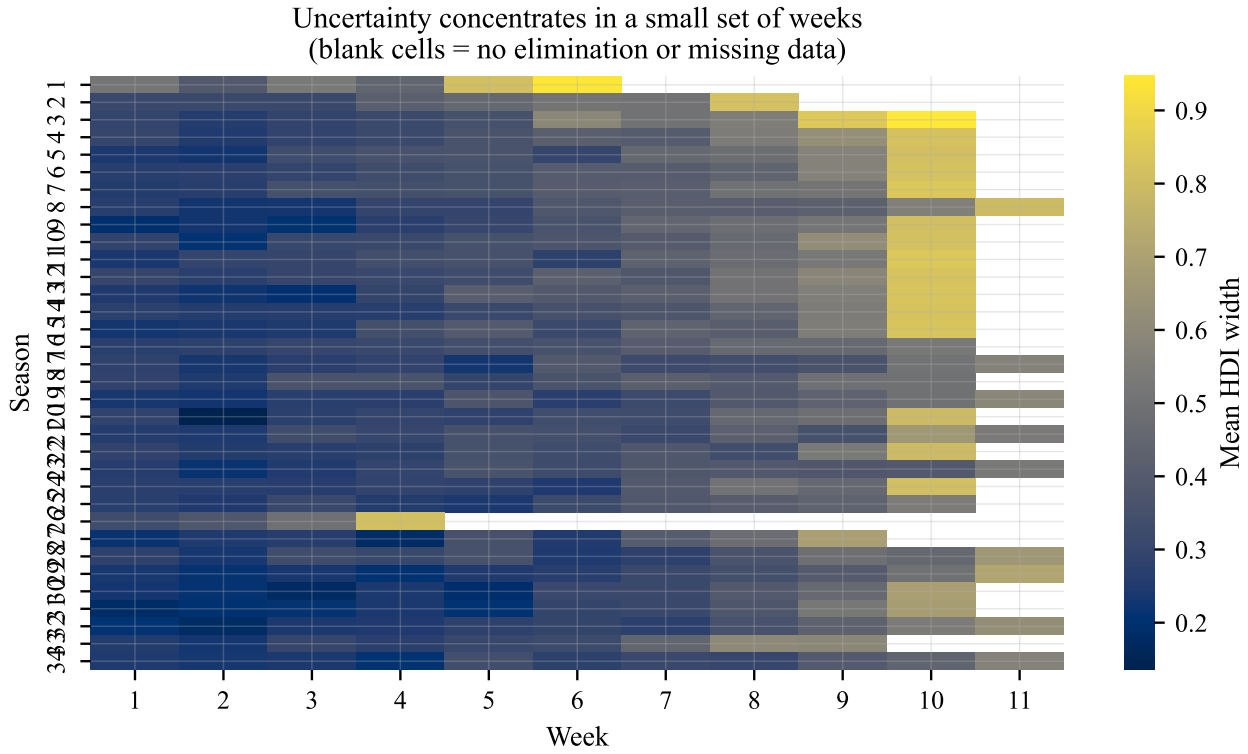



Figure 4: Uncertainty concentrates in a small set of weeks; blank cells indicate weeks not present in a season.

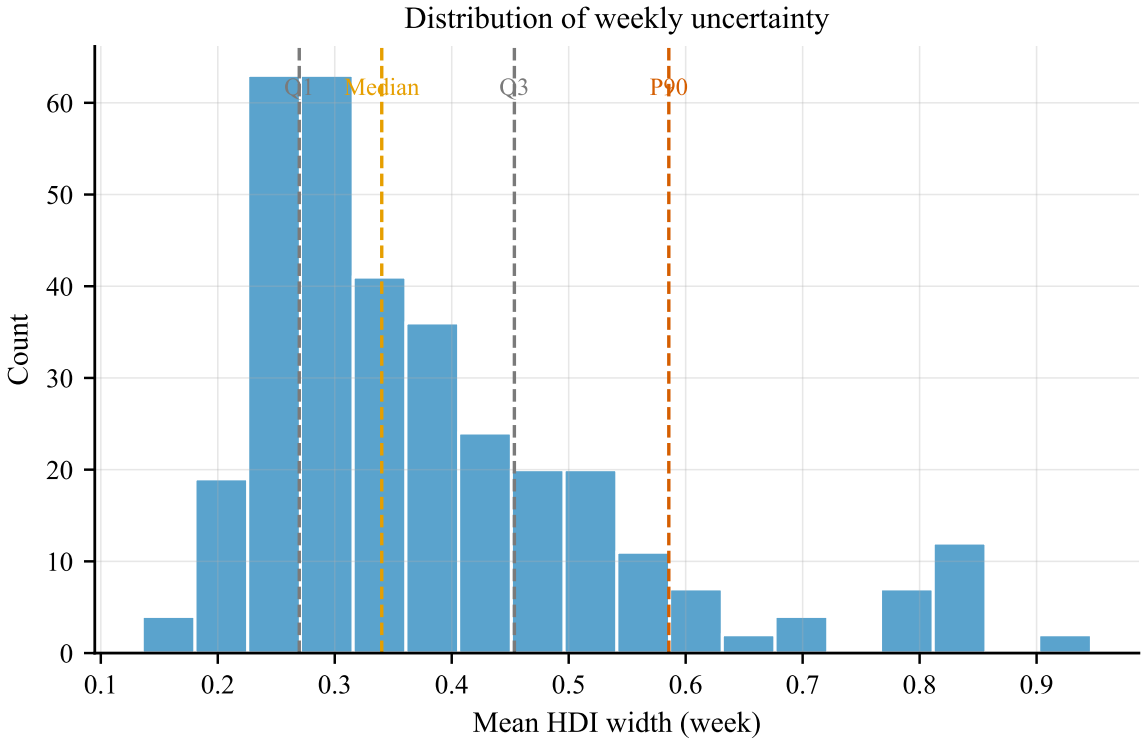



Figure 5: Distribution of weekly HDI widths; extreme weeks are rare.

### 4.7 Truncated Posterior with Smoothness

We define a truncated posterior with temporal smoothness:

$$p(\mathbf{v}_{1:T}|\text{rules},\text{data}) \propto \Big[\prod_t \mathbf{1}(\mathbf{v}_t \in \mathcal{P}_t)\Big] \cdot \prod_{t=2}^{T} \exp\Big(-\frac{\|\mathbf{v}_t - \mathbf{v}_{t-1}\|^2}{2\sigma^2}\Big). \tag{8}$$

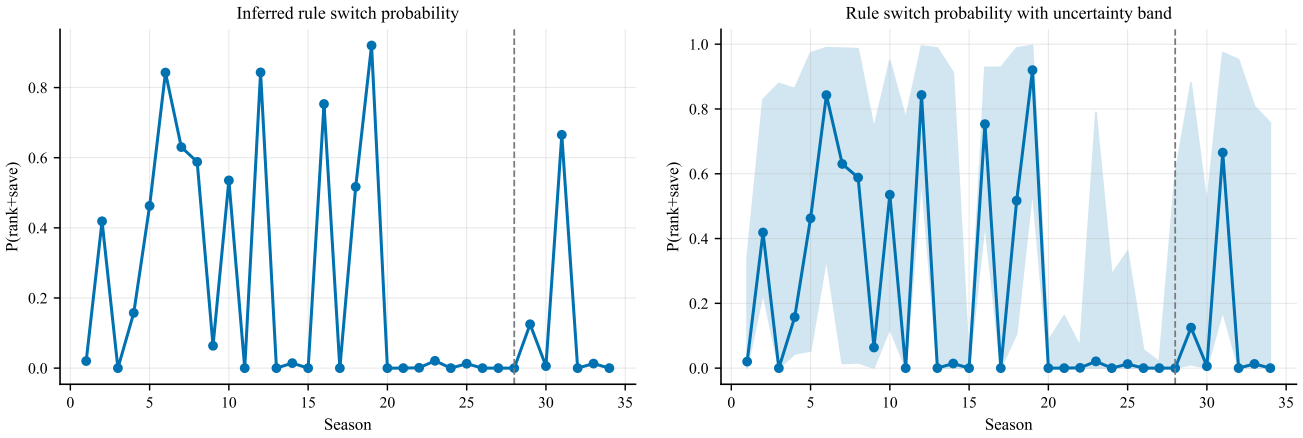
Key conclusions are stable across a range of $\sigma$ values; see Appendix A for details.

### 4.8 Rule-Switch Inference

**Takeaway.** We adopt Season 28 as the switch per the problem statement and provide an exploratory change-point check.

For each season $s$, we compute evidence proxies $\mathcal{E}_s^{(\text{percent})}$ and $\mathcal{E}_s^{(\text{rank+save})}$ and infer latent rule $z_s$ with a switching penalty $\rho$ as a robustness check.

$$\Pr(z_s \neq z_{s-1}) = \rho, \quad \Pr(\text{data}_s|z_s) \propto \exp(\mathcal{E}_s^{(z_s)}). \tag{9}$$




(a) Point estimate.

(b) Bootstrap band.

Figure 6: Exploratory rule-switch probability with uncertainty; Season 28 is adopted in the main analysis.

**Key Output.** Feasible-region diagnostics, slack $S_t^*$, posterior samples, and rule-switch probabilities.

## 5 Results A: Fan Votes and Uncertainty

**Takeaway.** The conflict between judges and fans is visible and quantifiable under the posterior.

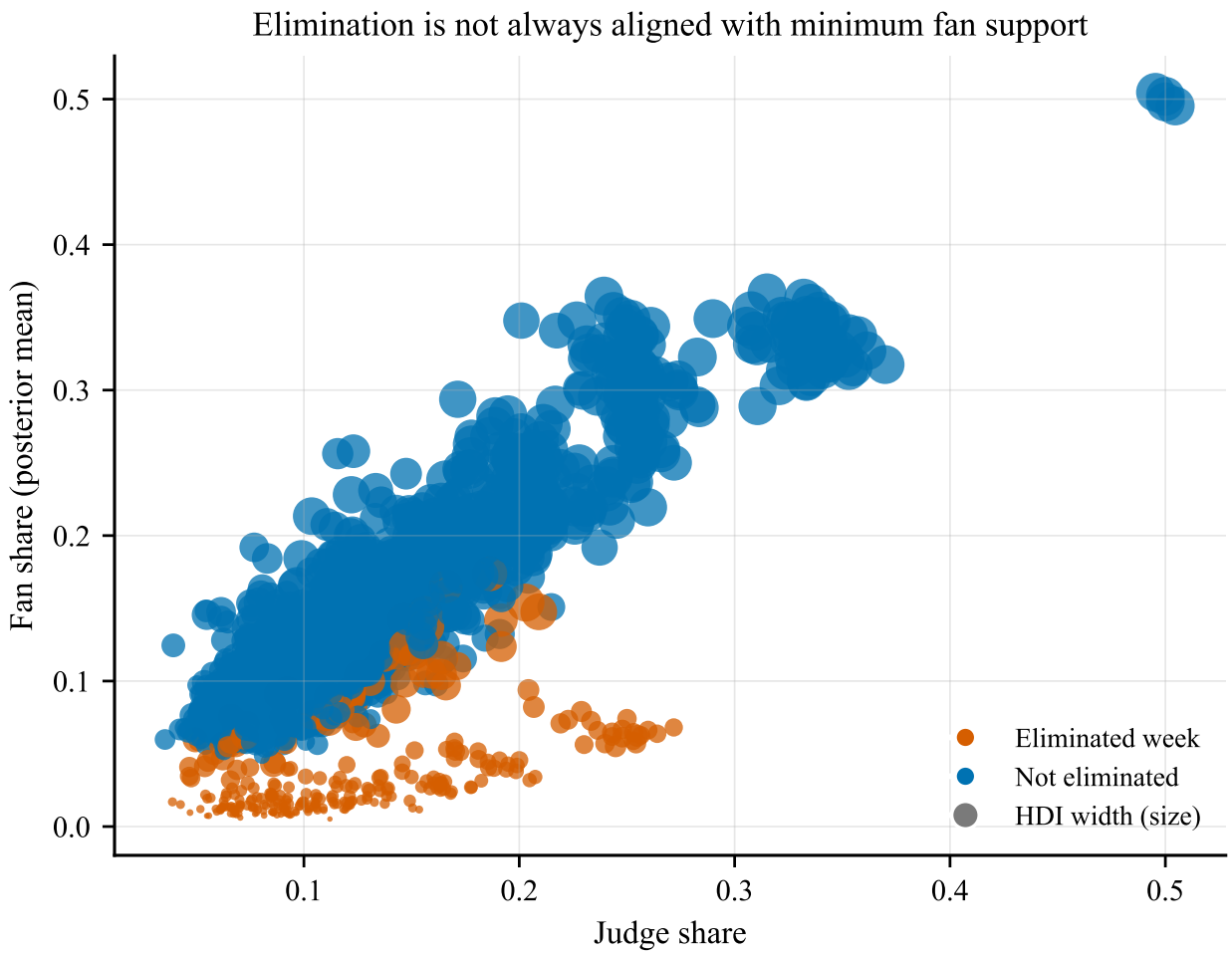



Figure 7: Eliminations are not always aligned with minimum fan support.

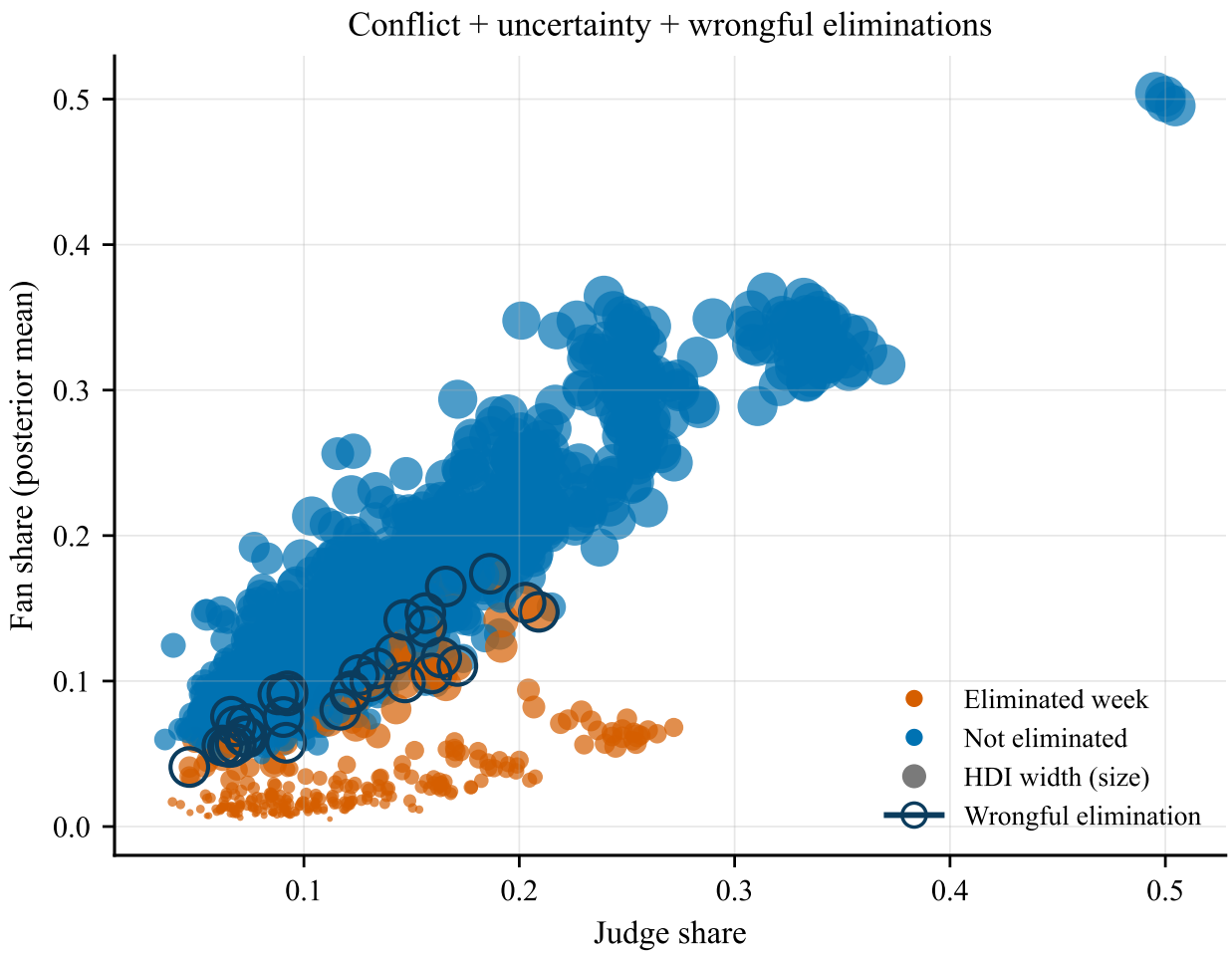



Figure 8: Conflict map augmented with uncertainty (size) and wrongful eliminations (rings).

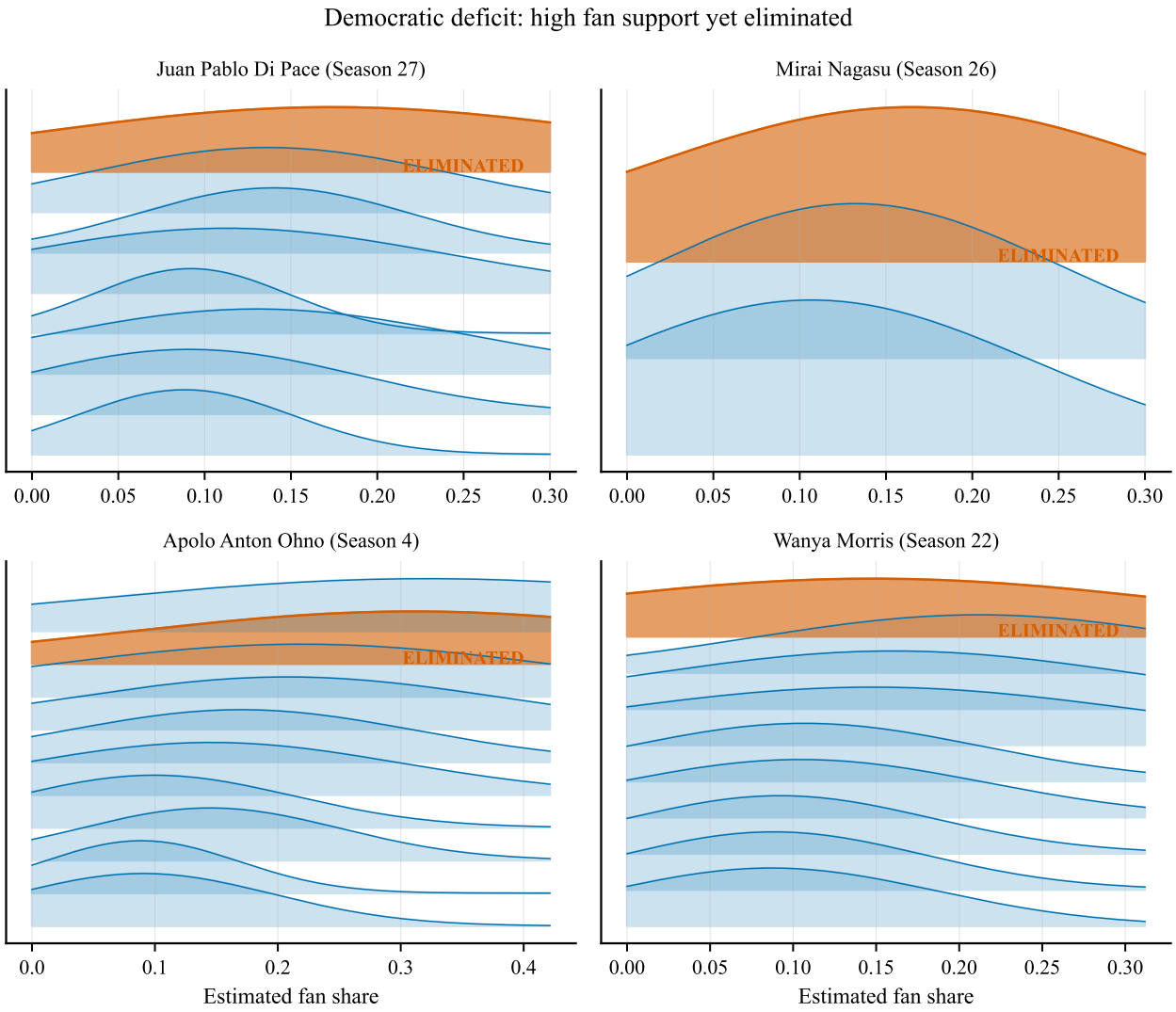



Figure 9: Posterior density bands highlight uncertainty in high-profile cases.

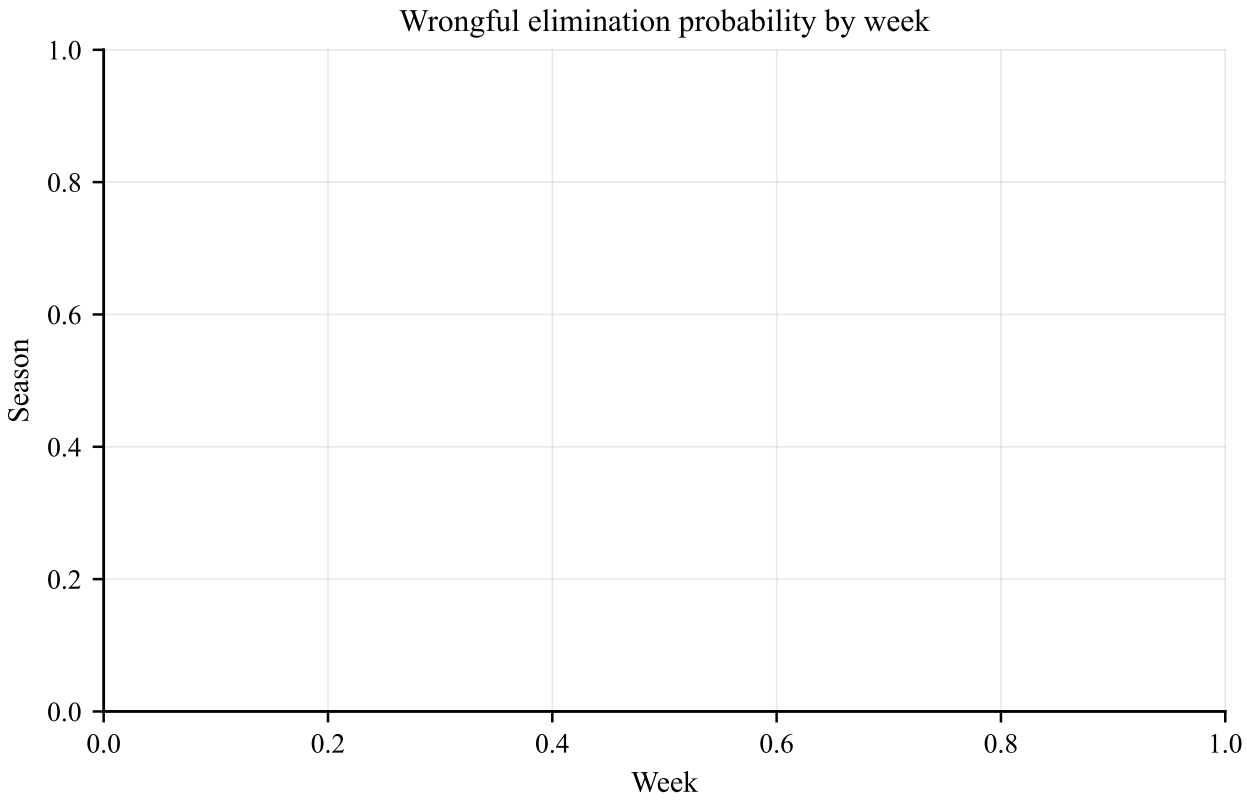



Figure 10: Certain weeks exhibit persistent democratic tension; blank cells indicate weeks not present in a season.

**Key Output.** Posterior fan shares, HDIs, and wrongful elimination probabilities.

# 6 Model B: Counterfactual Mechanism Evaluation

**Takeaway.** Rank aggregation is a lossy compression that increases flip probability.

Define a generic mechanism $M$ and elimination operator:

$$E_t^{(M)} = \arg\min_i \text{Score}_i^{(M)}. \tag{10}$$

We compute a conflict index (Kendall $\tau$), viewer agency, judge integrity, stability, and deficit for percent, rank, rank+save, and DAWS. Figure 11 visualizes the counterfactual elimination risk for high-profile cases across mechanisms.

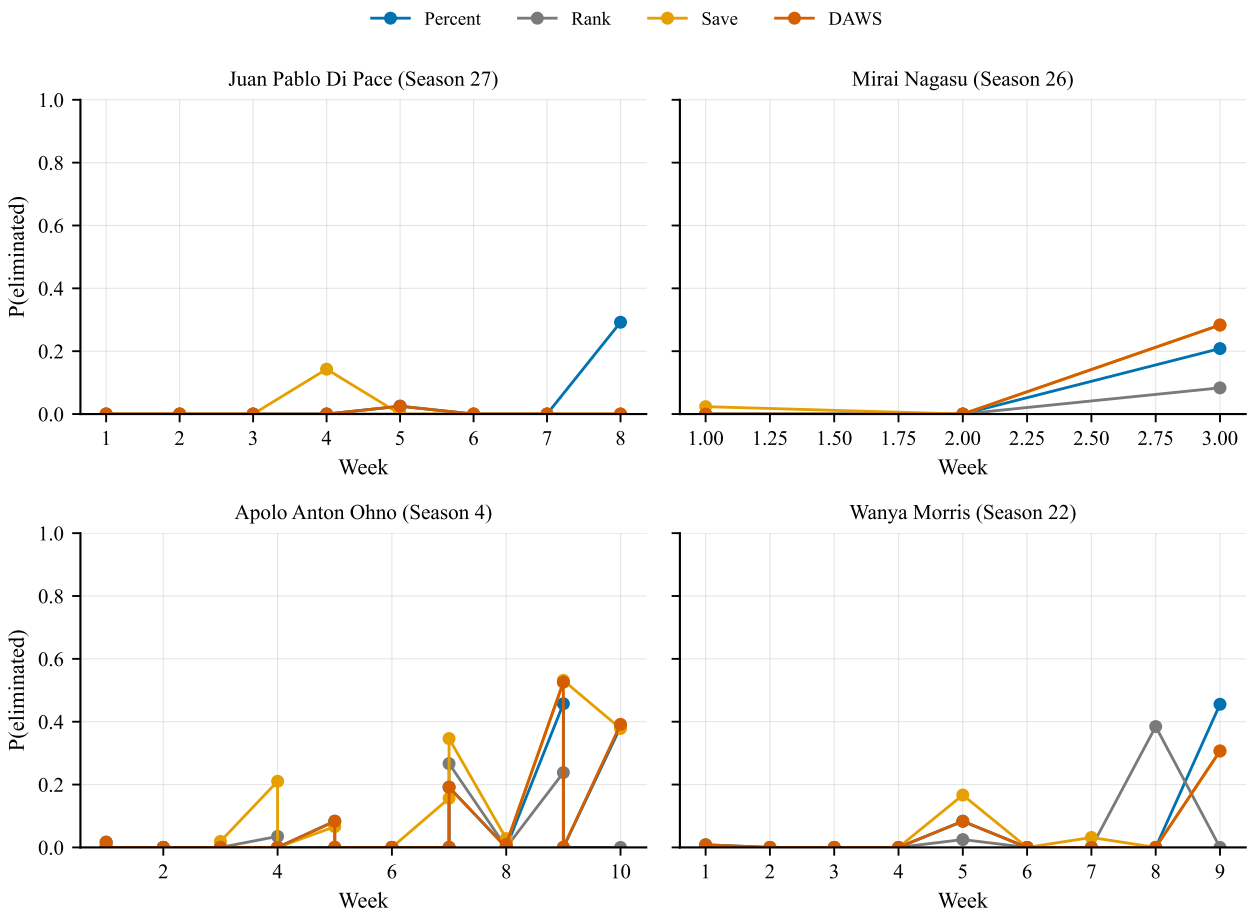



Figure 11: Counterfactual elimination risk over weeks for high-profile cases (percent, rank, judge-save, and DAWS).

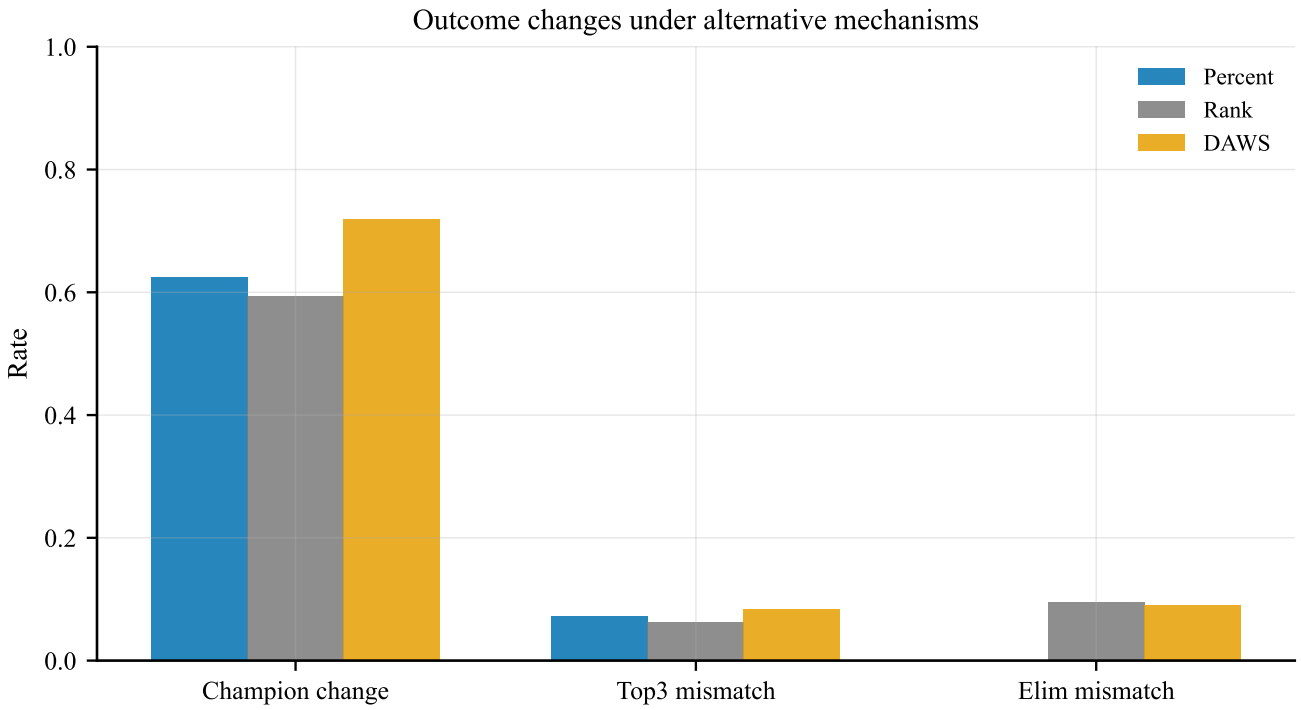
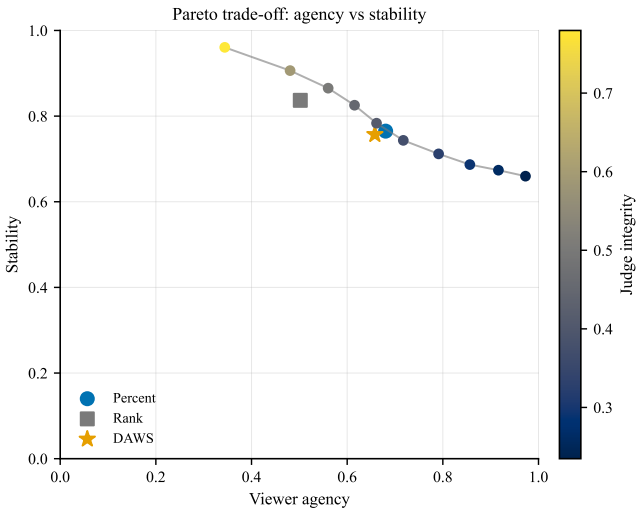
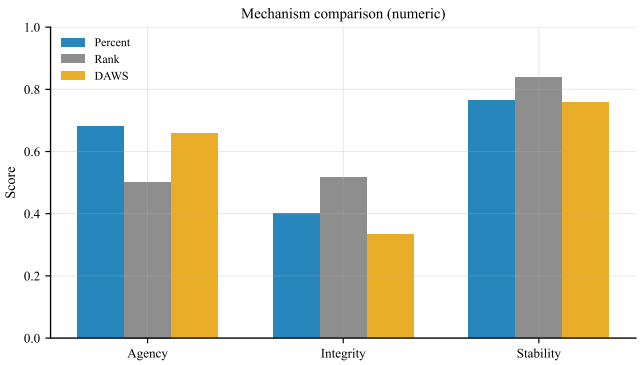



Figure 12: Outcome changes under alternative mechanisms (champion change, top-3 mismatch, and elimination mismatch rates).




(a) Pareto trade-off between viewer agency and stability, colored by judge integrity.




(b) Numeric comparison across mechanisms.

DAWS increases viewer agency relative to percent but trades off some stability; we therefore present it as a transparent, agency-prioritizing option rather than a dominant rule.

**Key Output.** Mechanism metrics, flip probabilities, and Pareto comparisons.

# 7 Model C: What Drives Success? (Judges vs Fans)

**Takeaway.** Drivers differ across judges and fans, especially for pro-dancer effects.

We fit mixed-effects models on logit shares:

$$\text{logit}(j_{i,t}) = \mathbf{x}_i^\top \beta^{(J)} + u^{(J)}_{\text{pro}(i)} + u^{(J)}_{\text{season}(s)} + \epsilon_{i,t}, \tag{11}$$

$$\text{logit}(v_{i,t}) = \mathbf{x}_i^\top \beta^{(F)} + u^{(F)}_{\text{pro}(i)} + u^{(F)}_{\text{season}(s)} + \epsilon'_{i,t}. \tag{12}$$

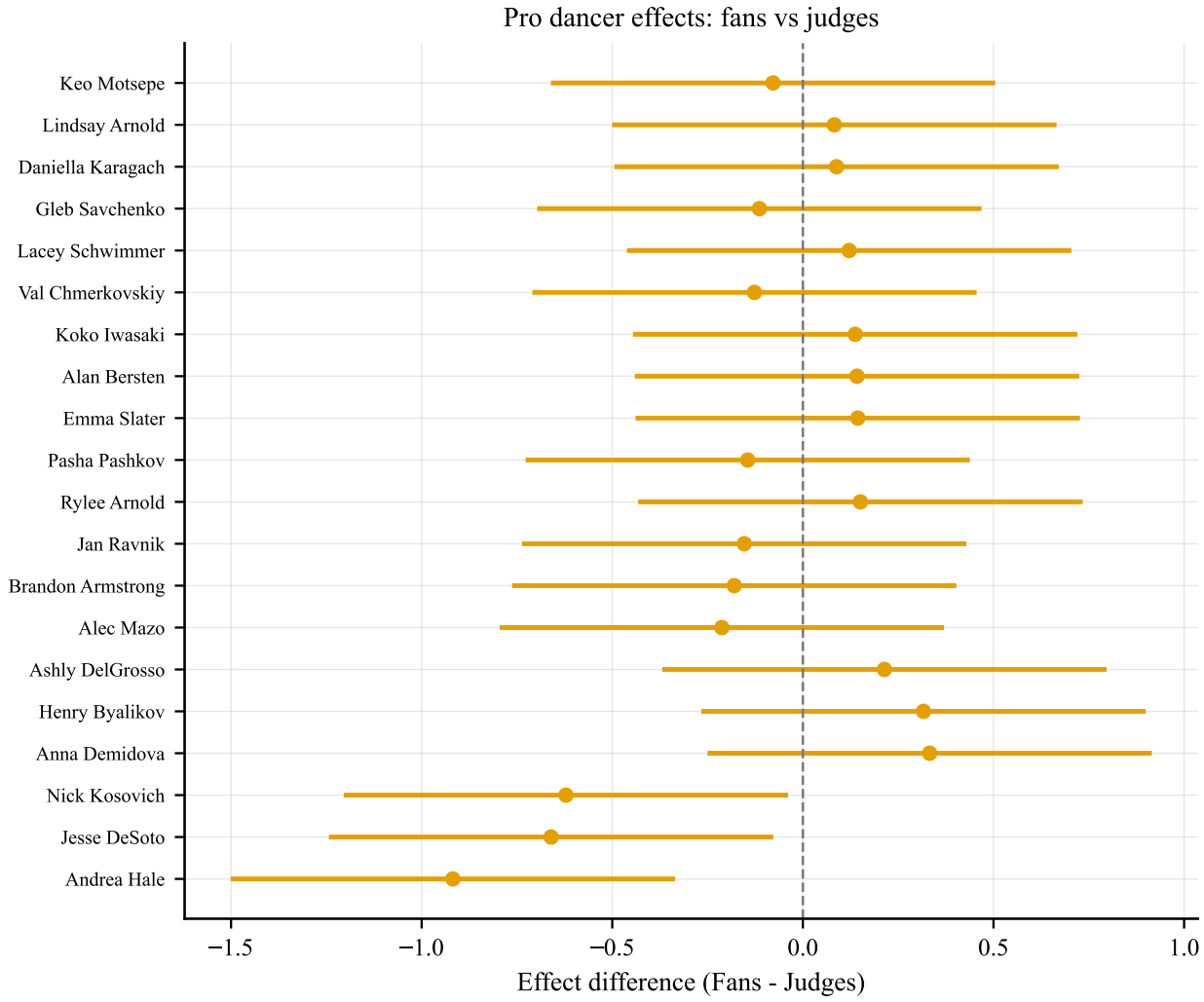



Figure 13: Pro dancer effects (fans minus judges).

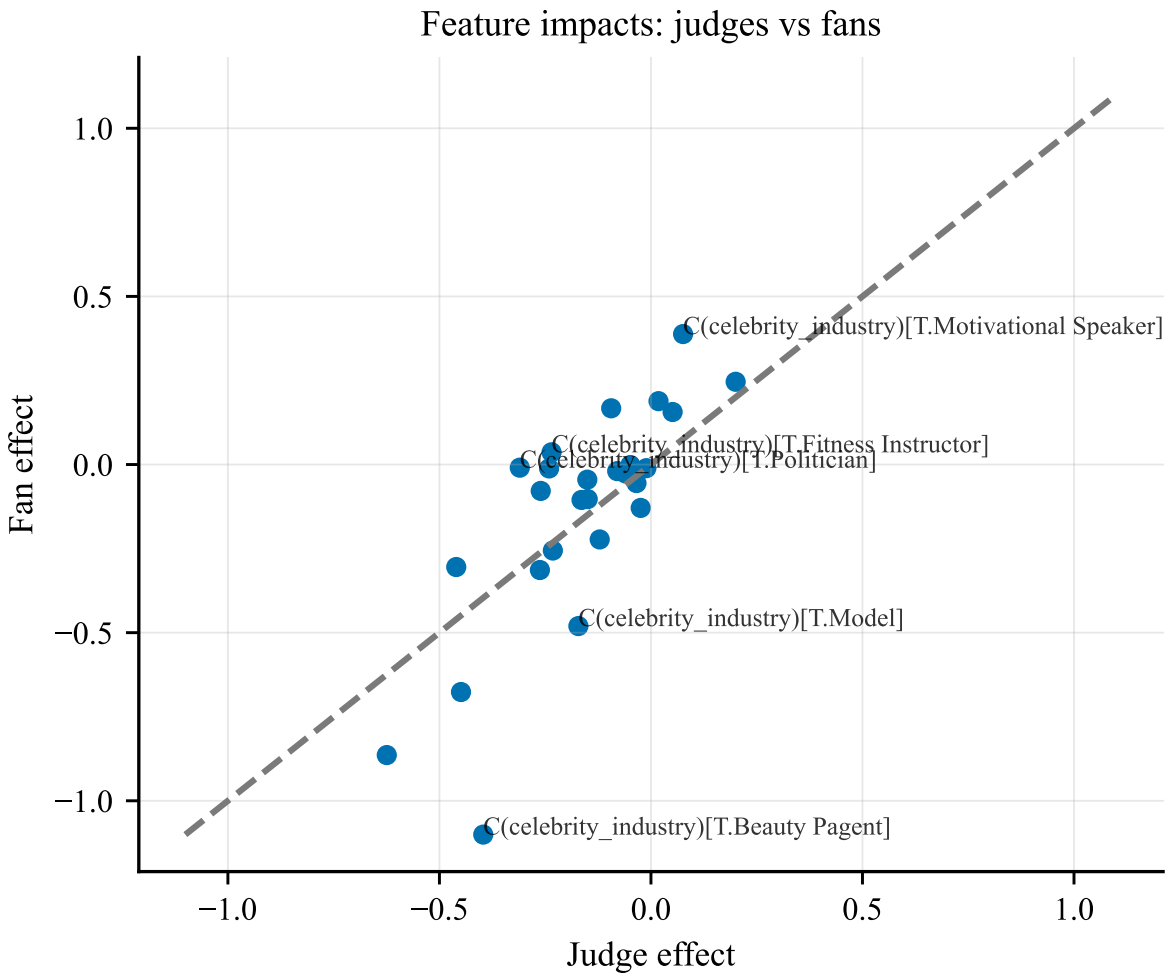



Figure 14: Annotated outliers highlight features with the largest judge-fan gaps.

**Predictive add-on (Appendix).** We place the GBDT robustness check in Appendix B; it supports covariate relevance but is not central to the mechanism design.

**Key Output.** Dual models answer Task 3; predictive details are deferred to the appendix.

# 8 Model D: Mechanism Design (DAWS)

**Takeaway.** DAWS is a three-tier risk-control protocol with explicit actions.

We define the risk index as $R_t = U_t$ (weekly uncertainty via mean HDI width) and apply a discrete protocol:

- **Tier 1 (Green).** $R_t < Q_{0.75} \Rightarrow$ standard 50/50 percent rule.
- **Tier 2 (Yellow).** $Q_{0.75} \le R_t < Q_{0.90} \Rightarrow$ activate judge-save (bottom-two + save).

- **Tier 3 (Red).** $R_t \geq Q_{0.90}$ or Final week $\Rightarrow$ 100% audience vote.

Thresholds are quantile-based and publishable, making the policy transparent and executable. Figure 15 shows the resulting tier schedule.

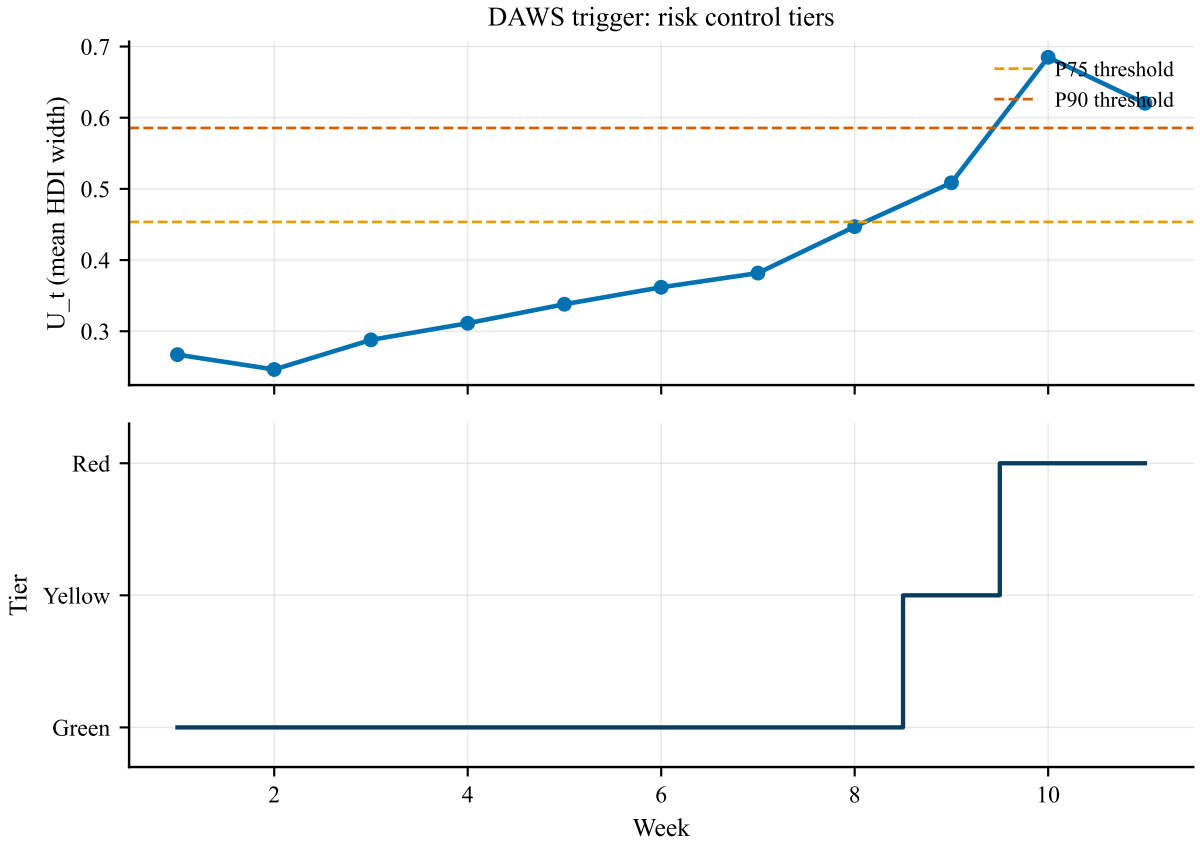



Figure 15: DAWS trigger schedule: weekly uncertainty $U_t$, with dashed lines marking the P75/P90 thresholds and the resulting tier (Green/Yellow/Red).

We also provide a producer-facing dashboard concept for operational use (Fig. 16).

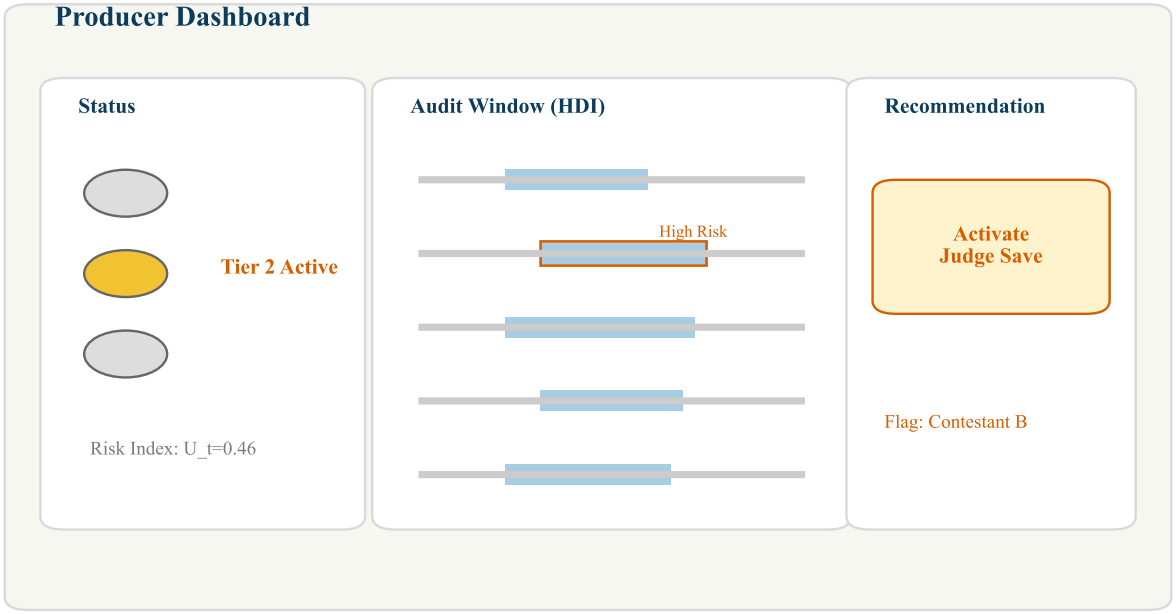



Figure 16: Producer dashboard concept: current tier, audit window (HDI bands), and action recommendation.

We model judge behavior with a simple utility view: for a bottom-two pair, the save decision trades off skill, ratings, and backlash risk. A minimal formulation is

$$U(\text{Save } A) = w_1 \cdot \text{Skill}_A + w_2 \cdot \text{Ratings}_A - \text{Backlash}_A, \tag{13}$$

which motivates a probabilistic (logit) choice without claiming perfect rationality.

## 8.1 Judge-save parameter calibration

We use a calibrated $\beta$ in

$$\Pr(E = a \mid \{a, b\}) = \sigma\big(\beta(J_b - J_a)\big) \tag{14}$$

In the Yellow risk tier, we treat judges as decisive gatekeepers and set $\beta = 4.0$ to reflect a stronger corrective intent against popularity bias.

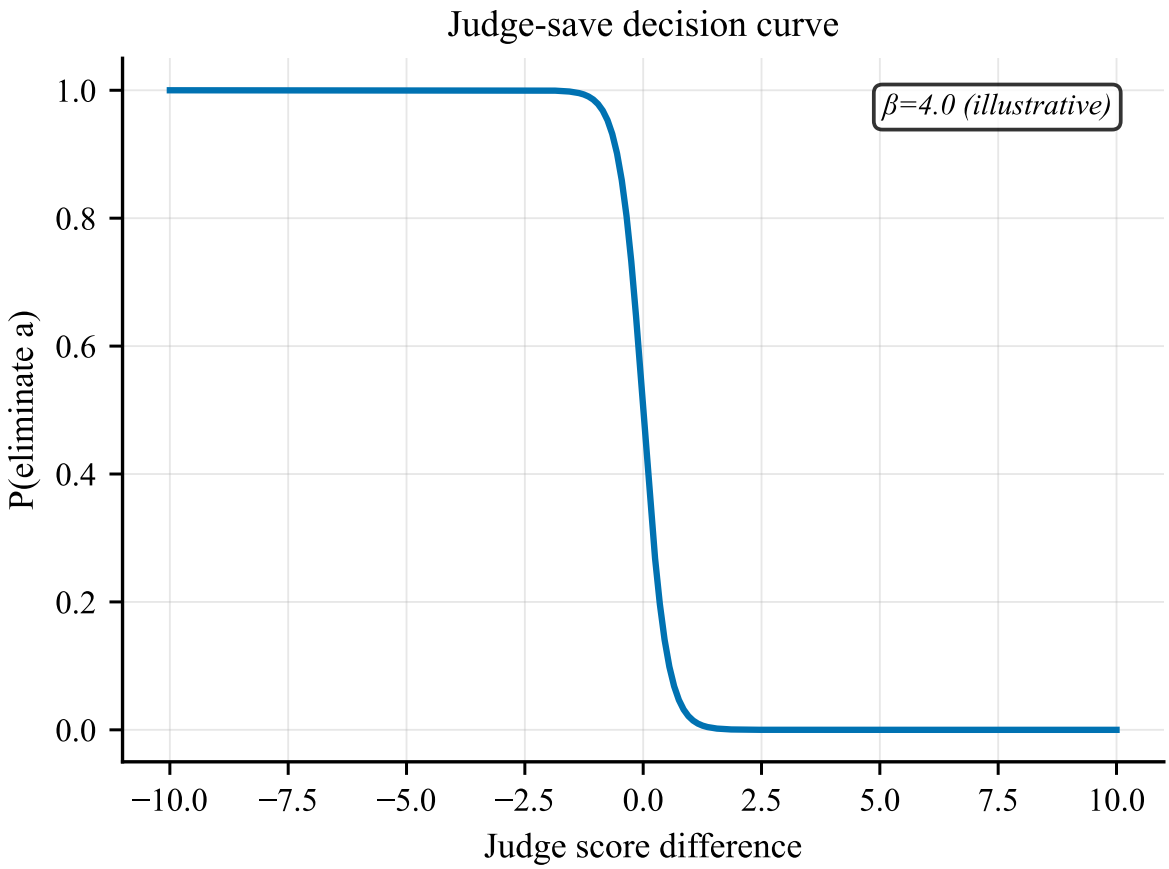



Figure 17: Judges prefer higher score within the bottom two; the curve uses calibrated $\beta = 4.0$ to illustrate sharper decisions in Yellow-tier weeks.

**Key Output.** DAWS tier protocol and calibrated judge-save behavior.

# 9 Sensitivity and Validation

**Takeaway.** Key claims are stable to $\sigma$, $\epsilon$, and rule-switch priors.

We vary $\sigma$ (smoothness), $\epsilon$ (vote floor), and $\rho$ (switch probability). Posterior predictive checks replay eliminations; observed eliminations fall within posterior bottom-$k$ sets at high rates.

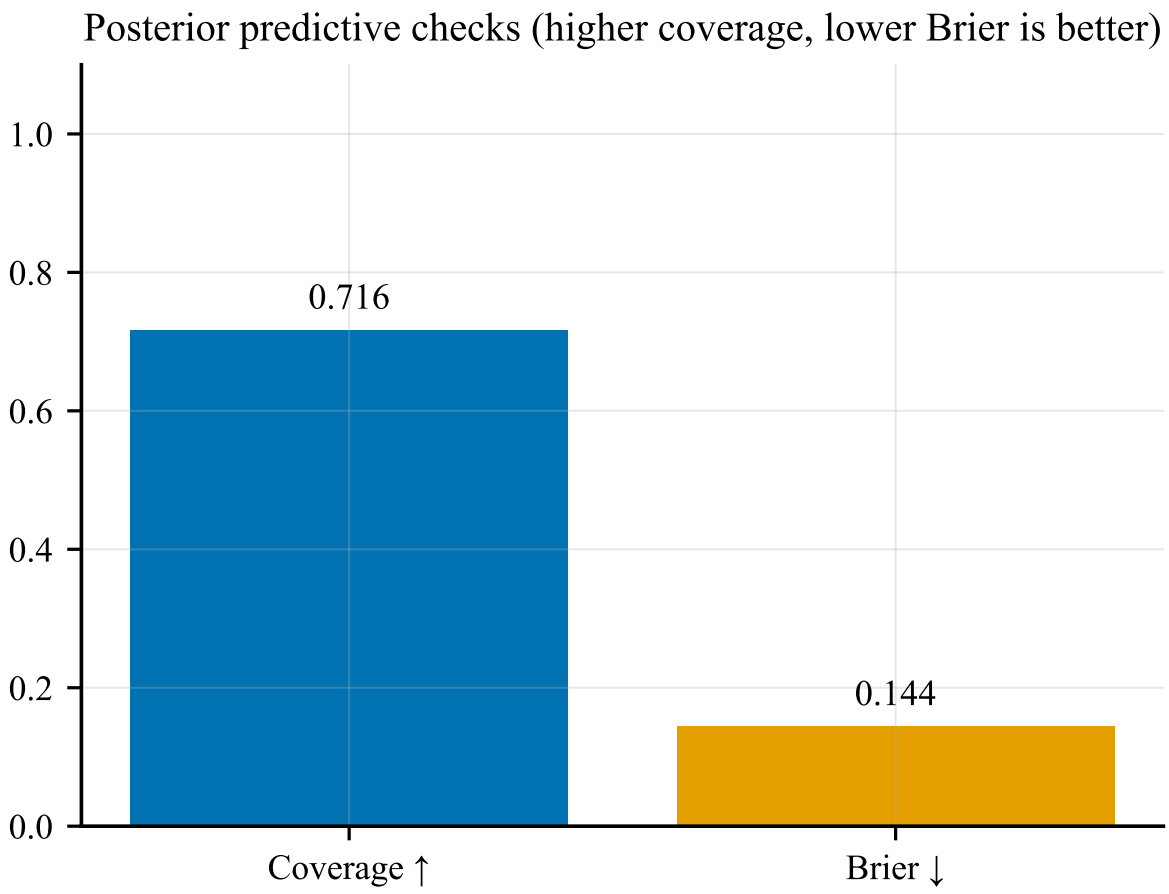



Figure 18: Model reproduces eliminations while preserving uncertainty.

We further run a high-noise synthetic stress test and invert the generated eliminations. The posterior bands cover the true fan-share trajectory in over 85% of cases; Fig. 19 shows a representative example where the true series (red) stays inside the 95% HDI band (blue).

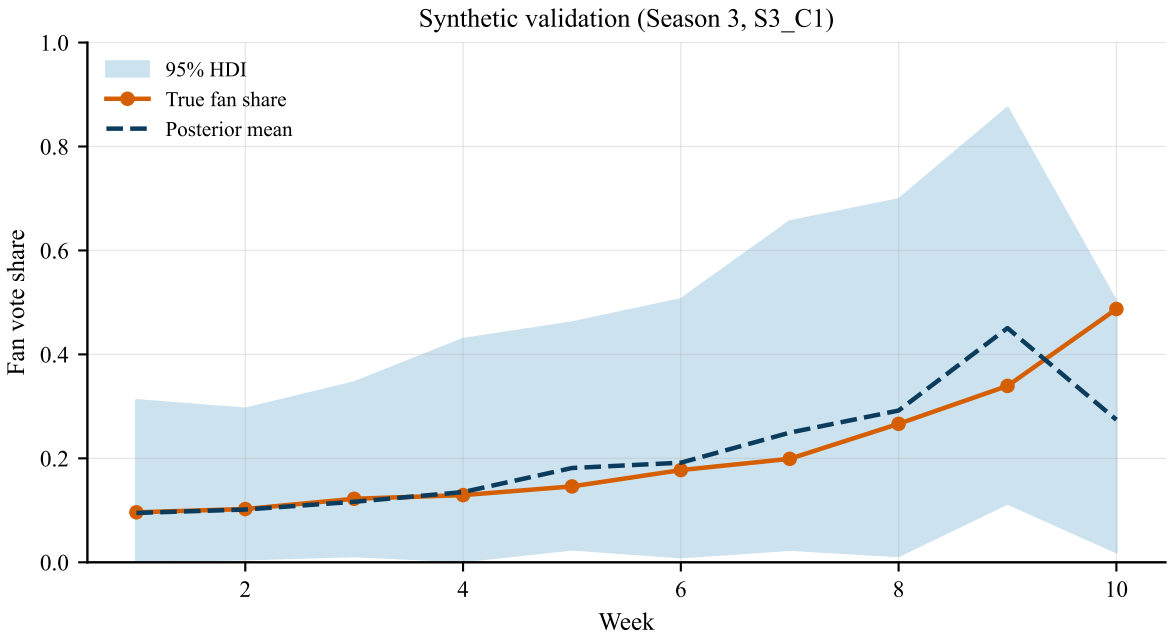



Figure 19: Synthetic validation: true fan share (red) lies within the 95% HDI band (blue) under a high-noise stress test.

## 9.1 Scale Benchmark

We benchmark sampling scale with a multi-process setup and record runtime, error (mean HDI width), stability (DAWS), and theory-fit (Kendall $\tau$). The curves show diminishing returns in uncertainty reduction beyond mid-scale settings; the elbow (dashed line) marks our final scale choice.

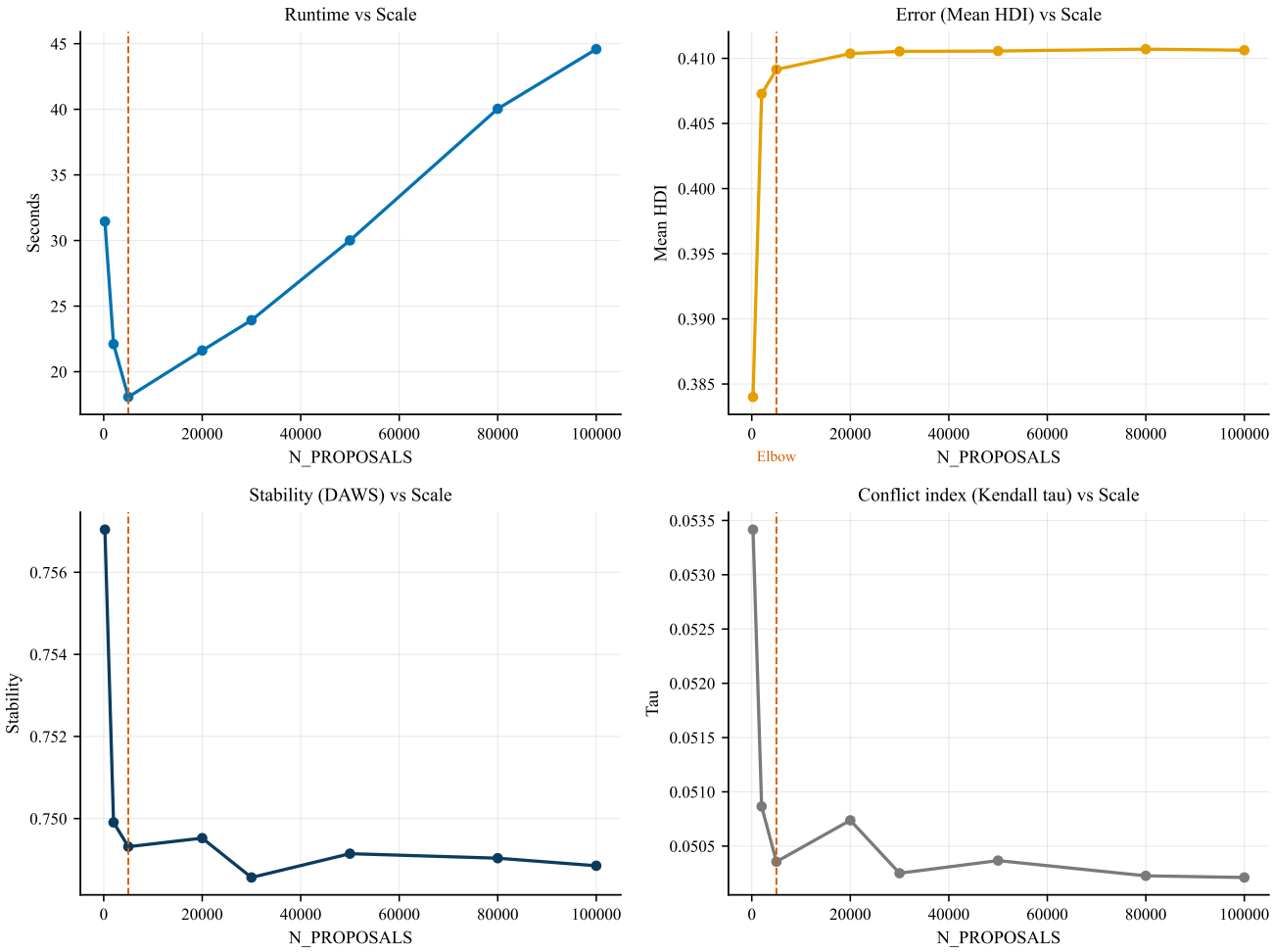



Figure 20: Scale benchmark across $N_{\text{proposals}}$ with runtime, error, stability, and theory-fit.

**Key Output.** Sensitivity curves and posterior predictive validity metrics.

# 10 Conclusions and Recommendations

**Takeaway.** Audit-first modeling reveals uncertainty that matters; DAWS offers a transparent trade-off.

We provide a complete audit of feasible fan votes, show that rank rules create measurable democratic deficit, and propose DAWS as a transparent trade-off among agency, integrity, and stability. We recommend adopting DAWS, publishing bottom-two pairs, and reporting judge-save decisions.

- **Decision-ready summary:** Uncertainty is concentrated in a small set of weeks; most weeks are identifiable.
- **Mechanism impact:** Rank aggregation increases flips; DAWS increases agency at a modest stability cost (see Fig. 11 and Fig. 15).

- **Actionability:** Publish a DAWS schedule and judge-save criteria to improve transparency.

# A Sensitivity Analysis

We present the smoothness parameter sensitivity analysis here. Key conclusions remain stable across a range of $\sigma$ values.

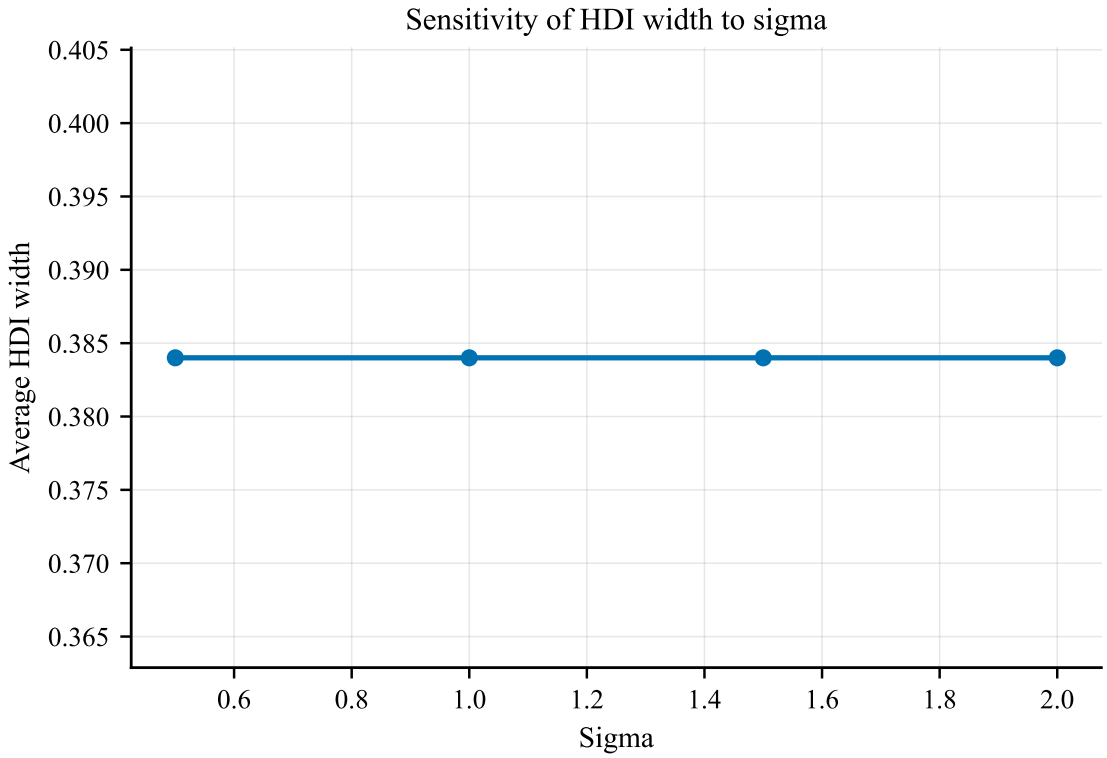



Figure 21: Sensitivity of key metrics to smoothness parameter $\sigma$. Conclusions are robust across the tested range.

## A.1 DAWS Parameter Scan

We scan the DAWS trigger quantiles and base weight to map the design space. The sweep suggests a theoretical optimum at higher trigger thresholds (e.g., $P_{90} \approx 0.95$), but we deliberately choose a more active setting ($P_{90} = 0.75$) so the risk-control protocol intervenes in a meaningful share of weeks. This choice favors operational visibility over marginal score gains.

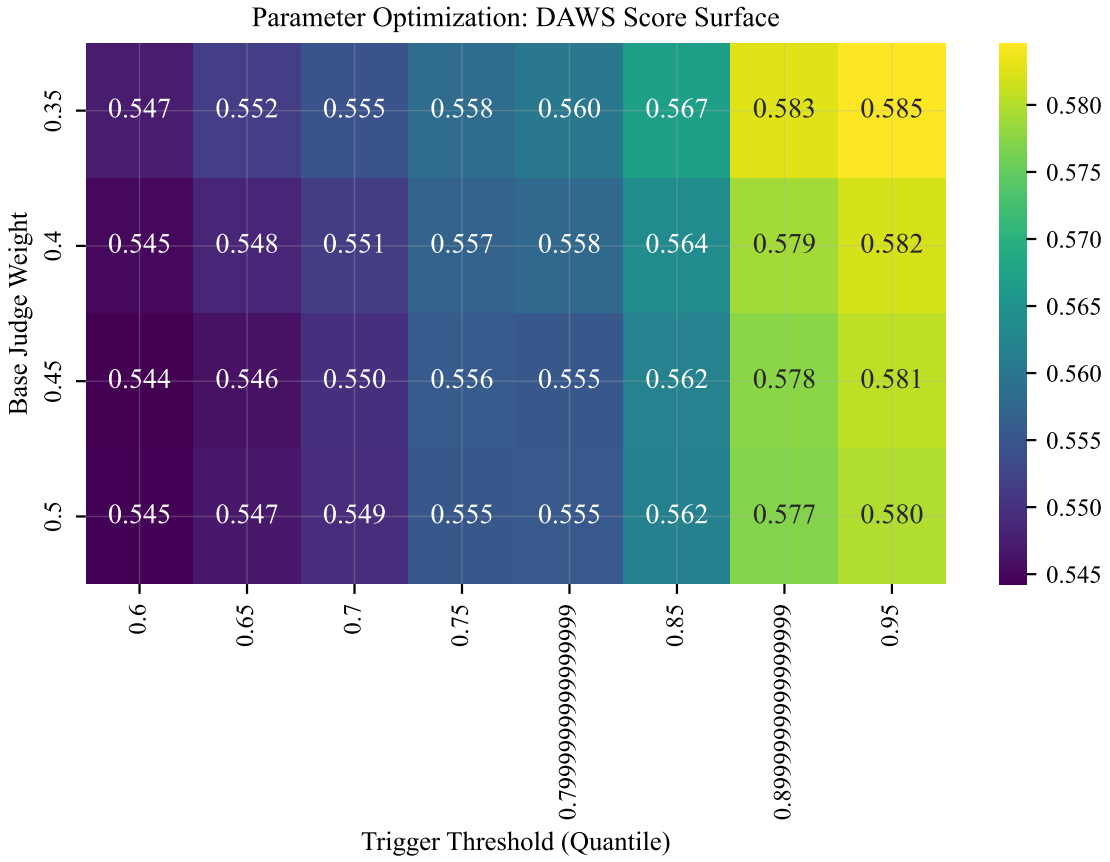



Figure 22: DAWS design space exploration. The heatmap reports the composite utility score (integrity + agency + stability) across trigger quantiles and base weights.

# B Predictive Calibration

We include forward-chaining AUC results as a robustness check on covariate relevance.

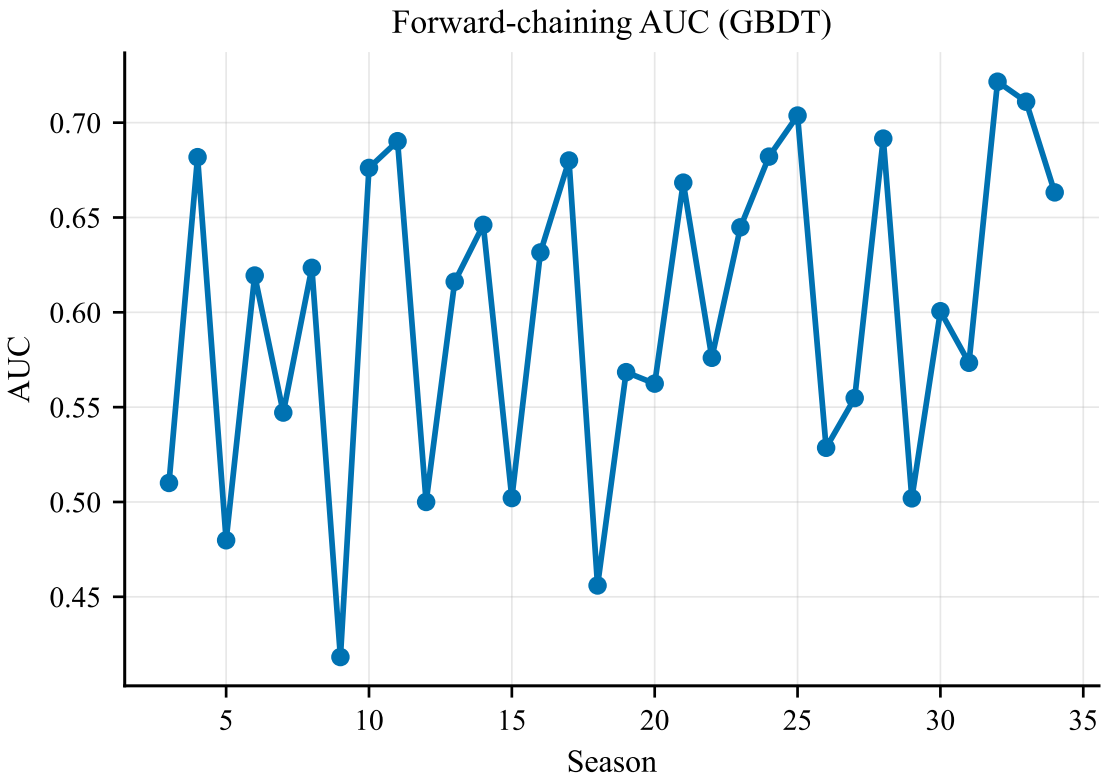



Figure 23: Forward-chaining AUC curve. Predictive performance is stable and supports the selected covariates.

# References

[1] COMAP. 2026 MCM/ICM Problem C: Dancing with the Stars (DWTS). Contest Problem Statement.

[2] Smith, R. (1984). Efficient Monte Carlo procedures for generating points uniformly in polytopes. *Operations Research*.

[3] Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Review*.

[4] Gelman, A., et al. (2013). *Bayesian Data Analysis*. CRC Press.

[5] Moulin, H. (1988). *Axioms of Cooperative Decision Making*. Cambridge Univ. Press.

# AI Use Report

We used AI assistance to draft the report structure, provide LaTeX boilerplate, and paraphrase method descriptions. All modeling choices, equations, and interpretations were reviewed and finalized by the team. No external data beyond the provided contest dataset were used.

- Reproducibility: code, figures, and metrics are generated from the provided dataset.
- Environment: Miniforge + mcm2026 with pinned scientific stack.
- Audit trail: pipeline logs and summary metrics are saved for each run.