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

Auditing and Designing the DWTS Voting Mechanism

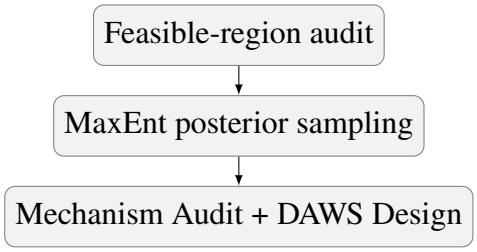
We treat DWTS as an audit-and-design problem: characterize feasible fan votes, quantify uncertainty, and redesign rules for agency, integrity, and stability.

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.

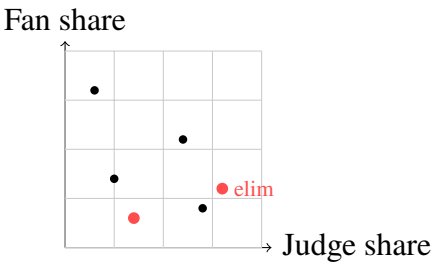
Core Results (selected).

Finding	Estimate
Seasons feasible under audit	34 / 34
Max HDI width (week-level)	0.95
Mean HDI width (week-level)	0.412
Median HDI width (week-level)	0.367
P90 HDI width (week-level)	0.599
Rank vs percent flip rate	25.7%
DAWS stability	0.842
DAWS judge integrity	0.461
Conflict index (Kendall τ)	0.049
DAWS improvement in stability	+-9.7%

Method Flow.



Conflict Map (summary visual).



Recommendation. Adopt DAWS as a cascading protocol (finale override, conflict-triggered judge-save, otherwise percent) 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 as a cascading protocol: a finale override (audience-only), a conflict trigger A_t (judge-save), and a default Percent rule otherwise. The uncertainty signal V_t is used for disclosure and audit budget only, not for intervention. 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.

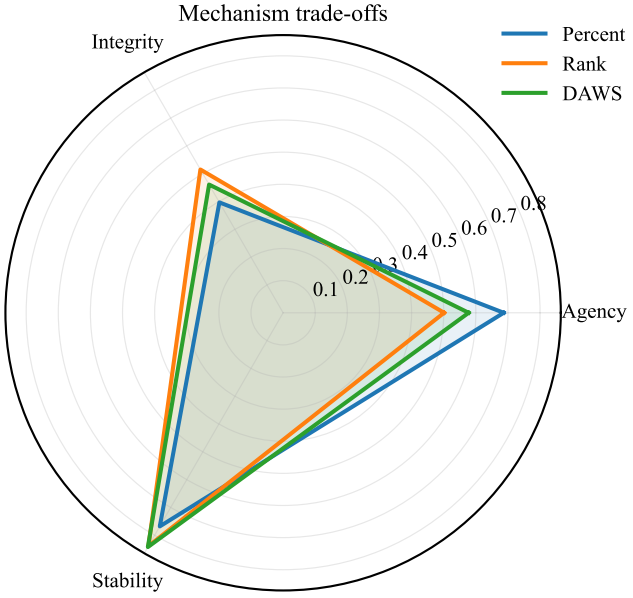


Figure 1: Mechanism trade-offs (all weeks; metrics aggregated across weeks).

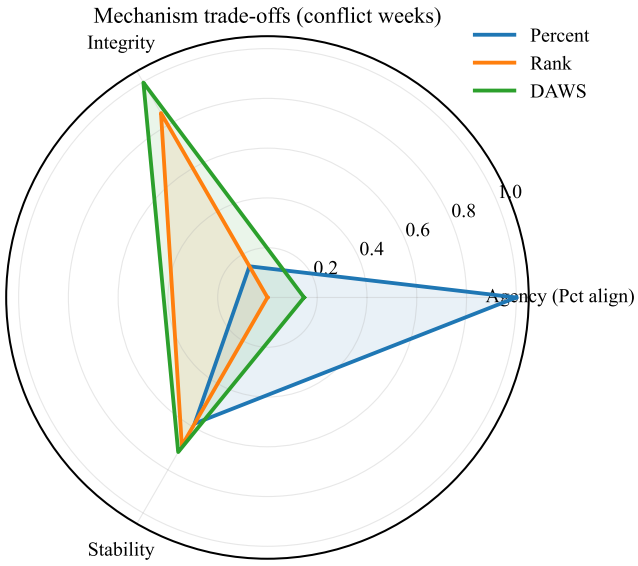


Figure 2: Mechanism trade-offs (conflict weeks only; when Percent/Rank disagree).

Contents

Memo **1**

1 Introduction and Roadmap **4**
1.1 Task-to-Section Mapping 4

2 Data and Rules **4**
2.1 Percent Rule 5
2.2 Rank Rule and Judge Save 5

3 Assumptions and Metrics **5**

4 Model A: Feasible-Region Audit **6**
4.1 Observables and Latents 6
4.2 Percent Rule Feasible-Region Audit 6
4.3 Rank Rule Feasible Orders (Monte Carlo) 7
4.4 Rule-adaptive Weeks 7
4.5 Engineering Approximation and Validation 7
4.6 Audit Specification and Feasibility Criteria 8
4.7 Identifiability and Feasible Mass 9
4.8 Truncated Posterior with Smoothness 11
4.9 Rule-Switch Inference 11

5 Results A: Fan Votes and Uncertainty **12**

6 Model B: Counterfactual Mechanism Evaluation **16**

7 Model C: What Drives Success? (Judges vs Fans) **19**

8 Model D: Mechanism Design (DAWS) **20**
8.1 Judge-save parameter calibration 22

9 Sensitivity and Validation **23**
9.1 Scale Benchmark 25

10 Conclusions and Recommendations **26**

A Sensitivity Analysis **28**

B Double-Elimination Week Feasibility Verification **28**

C LP/MILP Scope Declaration **29**

D Predictive Calibration **29**

E Audit Parameter Specification Table **30**

References **32**

AI Use Report **33**

1 Introduction and Roadmap

Takeaway. We model DWTS as an audit-and-design problem: audit feasible votes, stress-test uncertainty, deploy a conflict-triggered protocol with disclosure tiers, and present 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 cascading conflict-triggered protocol with disclosure tiers, 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 ϵ:

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

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 (weak ordering):

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

Note: A formal Δ min-gap can be introduced for theoretical analysis, but our implementation ***does not enforce any min-gap****; only the bottom-k ordering constraint is used (*$\max(score_E) \leq \min(score_S) + \epsilon$, $\epsilon = 10^{-6}$*). See Appendix C.*

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 β (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 τ) and a democratic deficit indicator.

We assume: (i) fan shares are nonnegative with floor ϵ; (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 τ): 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 is a conflict-triggered protocol: when Percent and Rank agree we follow the 50/50 percent rule; when they disagree we invoke judge-save with a decisive $\beta = 6.0$, and only the finale is audience-only. Quantile lines (P85/P95) are retained for monitoring and visualization.

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 ϵ
2: Filter proposals by elimination constraints (fast/strict)
3: Estimate (L_i, U_i) from accepted samples
4: Output samples and bound summaries

Audit-Weak weeks (disclosure only). When the strict feasibility sampler yields fewer than $N_{\text{strict,min}} = 500$ accepted proposals, we flag the week as *Audit-Weak*. These weeks are excluded from aggregate metrics and appear as marked points only. See Section 4.6 and Appendix E for full specification.

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 π by Monte Carlo
2: **for** each feasible π **do**
3: Draw Dirichlet proposals and retain those consistent with π
4: **end for**
5: Aggregate samples across feasible π

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	0.0238
Top-1 agreement (fast vs strict)	26.6%
Top-2 agreement (fast vs strict)	22.3%
Conflict index shift (Kendall τ)	0.104
Agency shift (percent)	0.081
Flip-rate shift (percent vs rank)	12.13%

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.

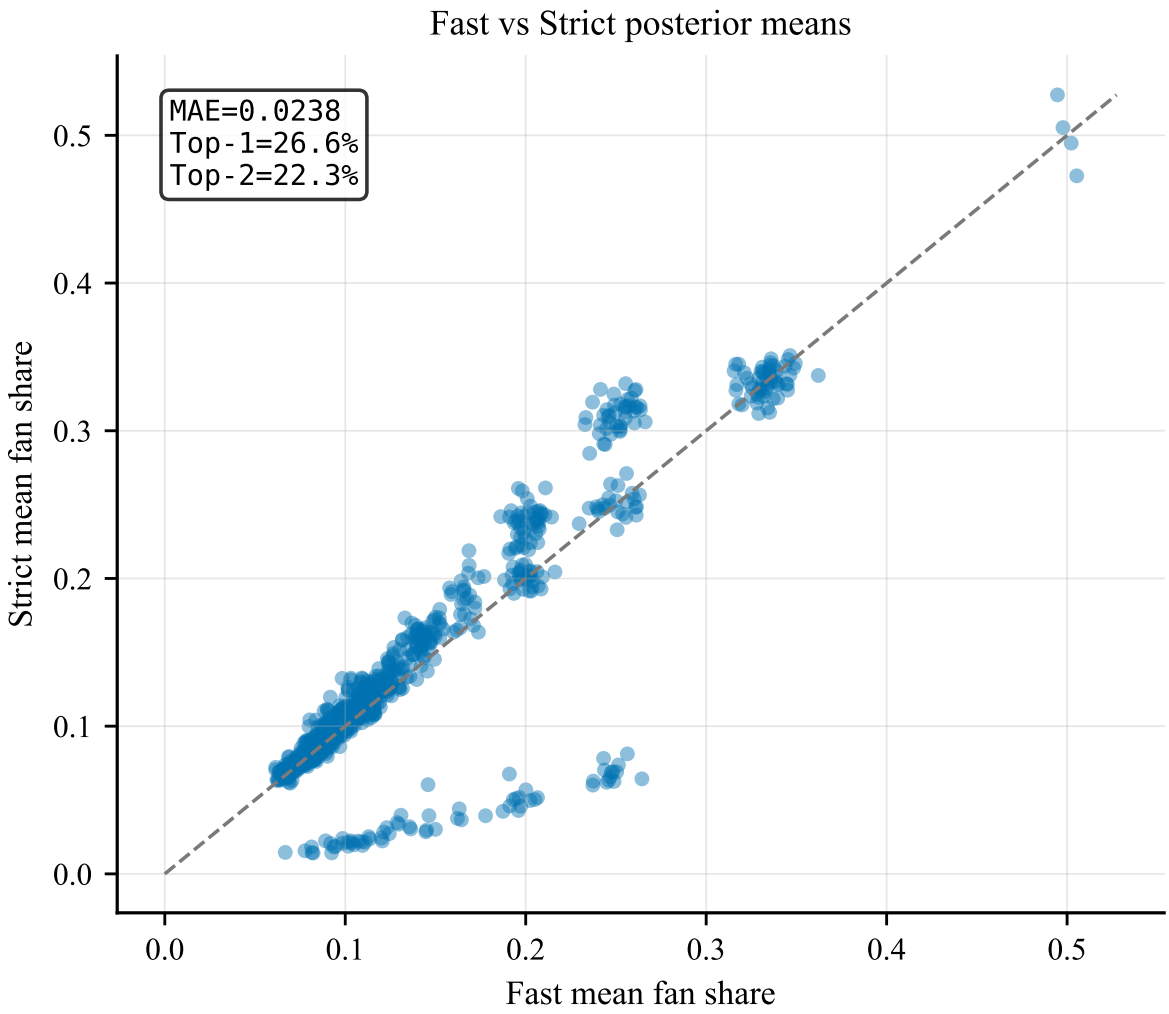


Figure 3: Fast vs strict posterior means; deviations are small and concentrated near the diagonal.

4.6 Audit Specification and Feasibility Criteria

Takeaway. We fix all tolerance parameters and sampling budgets as non-adjustable policy to avoid post-hoc tuning.

Strict feasibility definition. A vote vector $\mathbf{v}$ is *strict feasible* if it satisfies:

1. **Simplex**: $v_i \geq 0$ for all i, and $|\sum_i v_i - 1| \leq \varepsilon_{\text{sum}}$ with $\varepsilon_{\text{sum}} = 10^{-9}$.
2. **Elimination**: For all eliminated contestants $e \in E$ and all survivors $s \in S$, we require $\max_e C_e \leq \min_s C_s + \varepsilon_{\text{ord}}$ with $\varepsilon_{\text{ord}} = 10^{-6}$.

Ties (where eliminated and surviving contestants have equal combined scores) are permitted; the tolerance ε_{ord} handles numerical precision only.

Sampling budget gate. We require at least $N_{\text{strict,min}} = 500$ strict feasible samples per week. The default proposal count is determined by the *10th percentile* of accept rates (not the median), because

exclusion risk is driven by low-acceptance-rate tail weeks:

$$N_{\text{proposals}} \geq \left\lceil \frac{N_{\text{strict,min}}}{q_{0.10}(\text{accept\_rate\_strict})} \right\rceil.$$

The quantile $q = 0.10$ is fixed as non-adjustable policy. In our data, $q_{0.10} \approx 0.07$, yielding a recommended budget ≈ 7159; we set $N_{\text{proposals}} = 8000$ (default).

Excluded weeks and downgrade trigger. Weeks with fewer than $N_{\text{strict,min}}$ strict feasible samples are flagged as *Audit-Weak* and excluded from aggregate metrics (they appear as marked points only). **The goal is not to guarantee all weeks pass**; worst-case weeks (with accept rates near 0.008) may be excluded, and this is expected and disclosed.

We **pre-register** the following reporting rule: if $r_{\text{excl}} \geq 20\%$, all season-level conclusions are downgraded to "exploratory." We do *not* tune parameters to minimize r_{excl}; the 20% threshold and all audit parameters are fixed *a priori* and applied uniformly.

4.7 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.

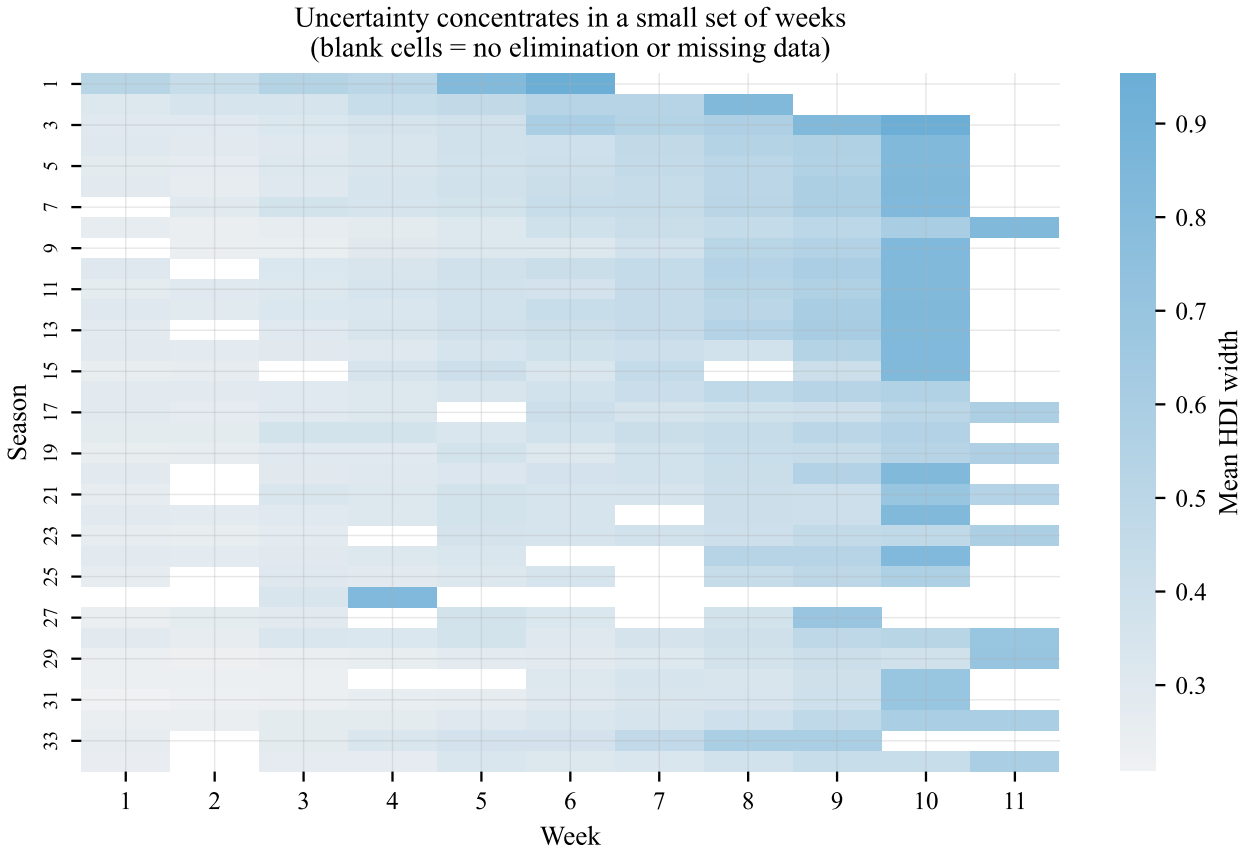


Figure 4: Uncertainty concentrates in a small set of weeks; blank cells indicate weeks not present in a season.

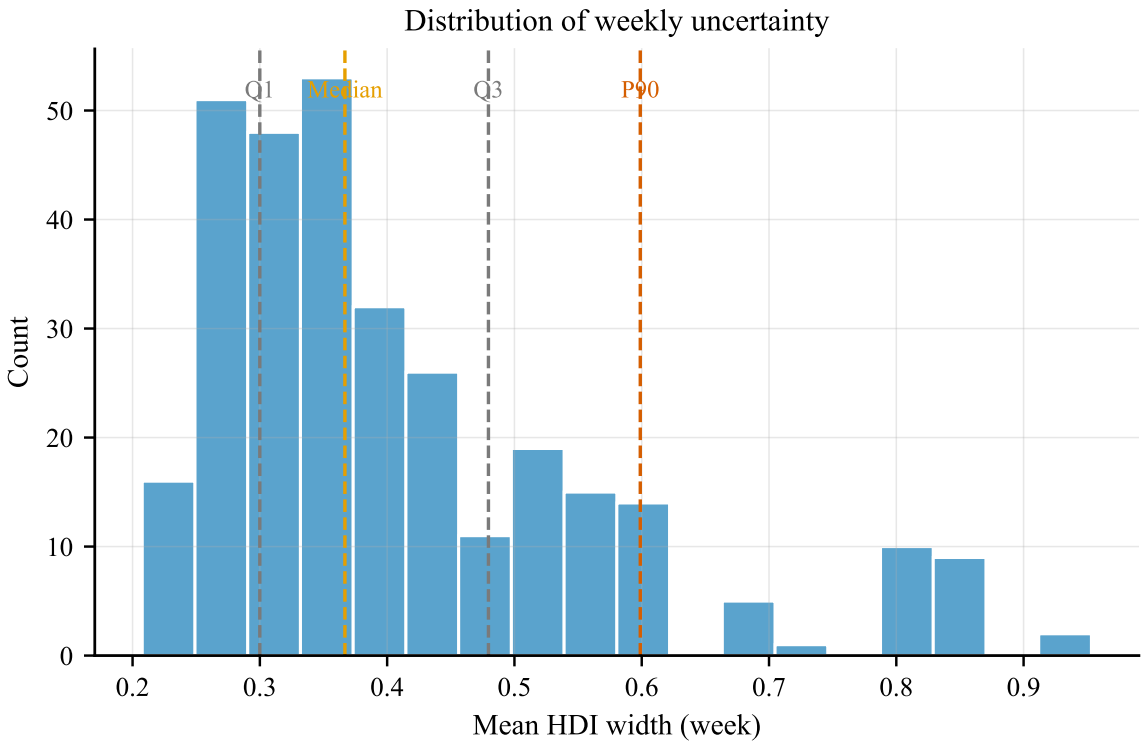


Figure 5: Distribution of weekly HDI widths; extreme weeks are rare.

4.8 Truncated Posterior with Smoothness

We define a truncated posterior with temporal smoothness:

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

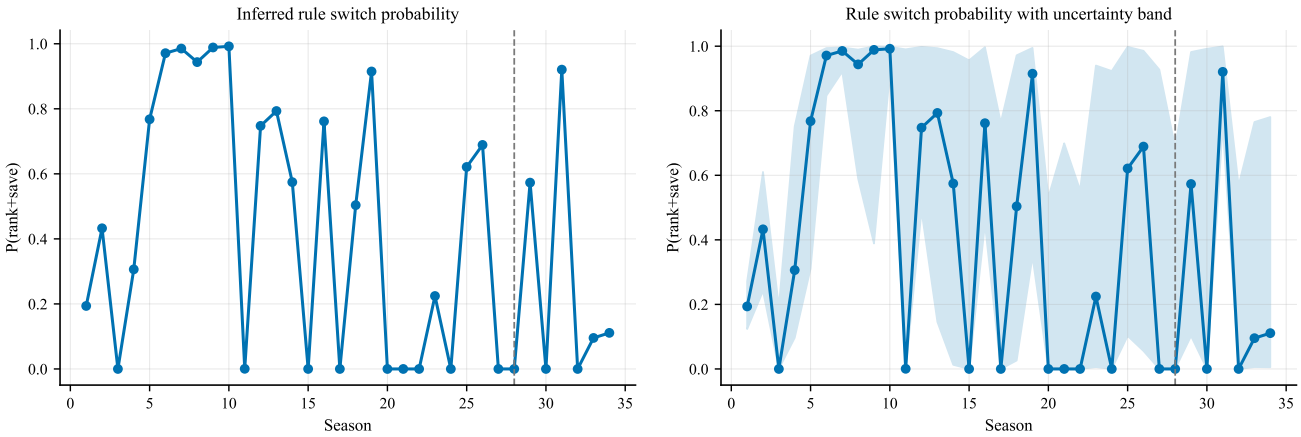
Key conclusions are stable across a range of σ values; see Appendix A for details.

4.9 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 ρ 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.

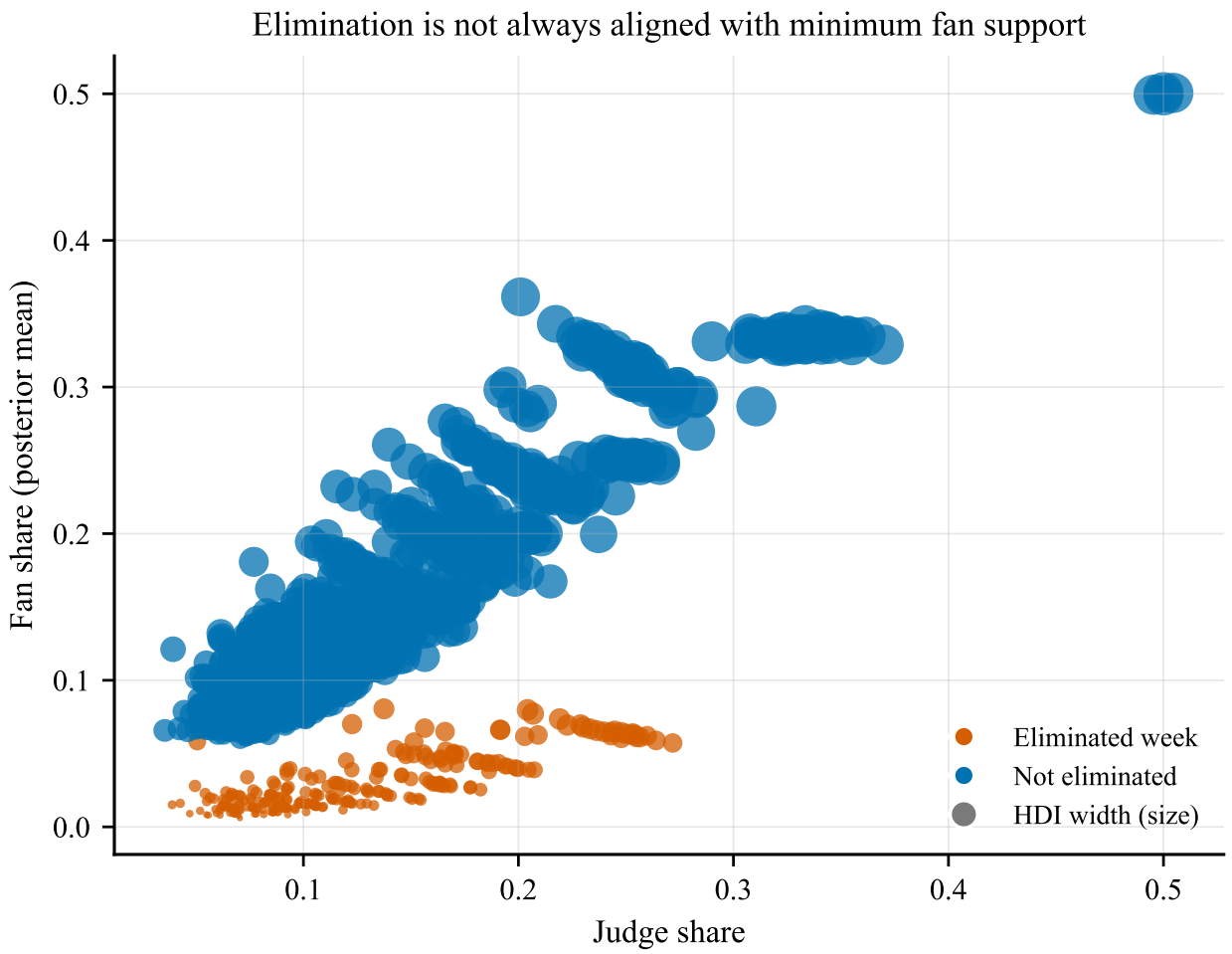


Figure 7: Eliminations are not always aligned with minimum fan support.

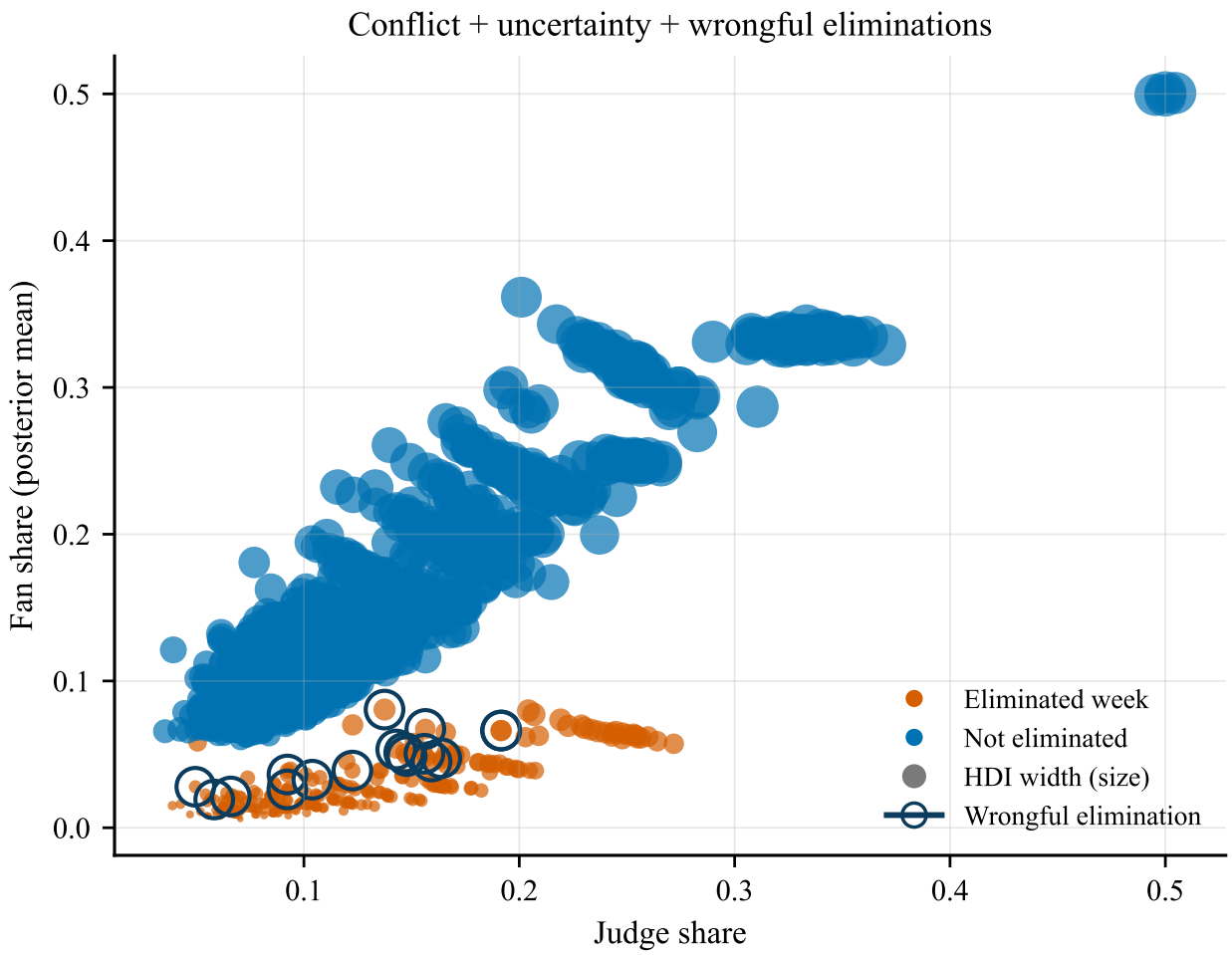


Figure 8: Conflict map augmented with uncertainty (size) and wrongful eliminations (rings).

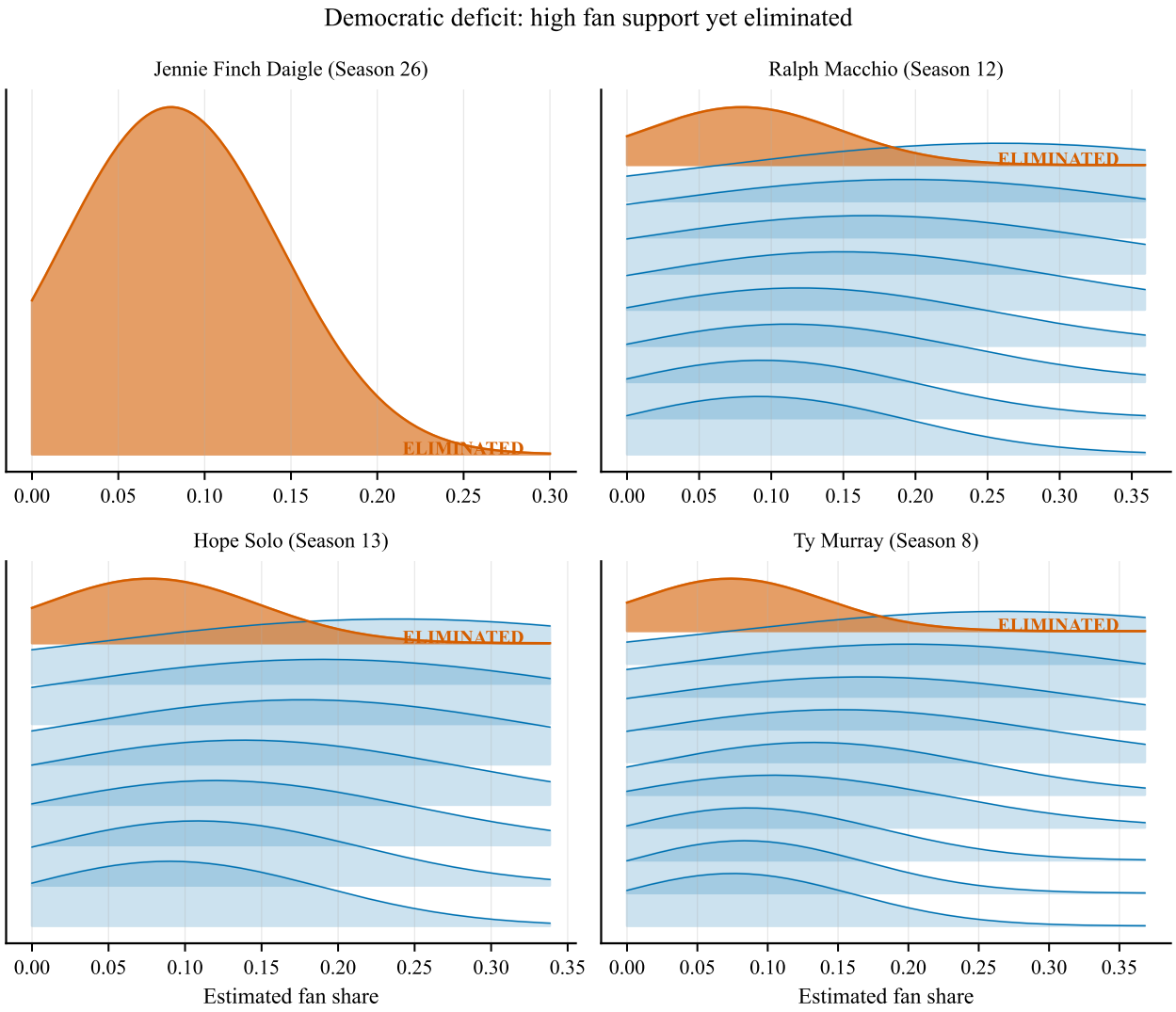


Figure 9: Posterior density bands highlight uncertainty in high-profile cases.

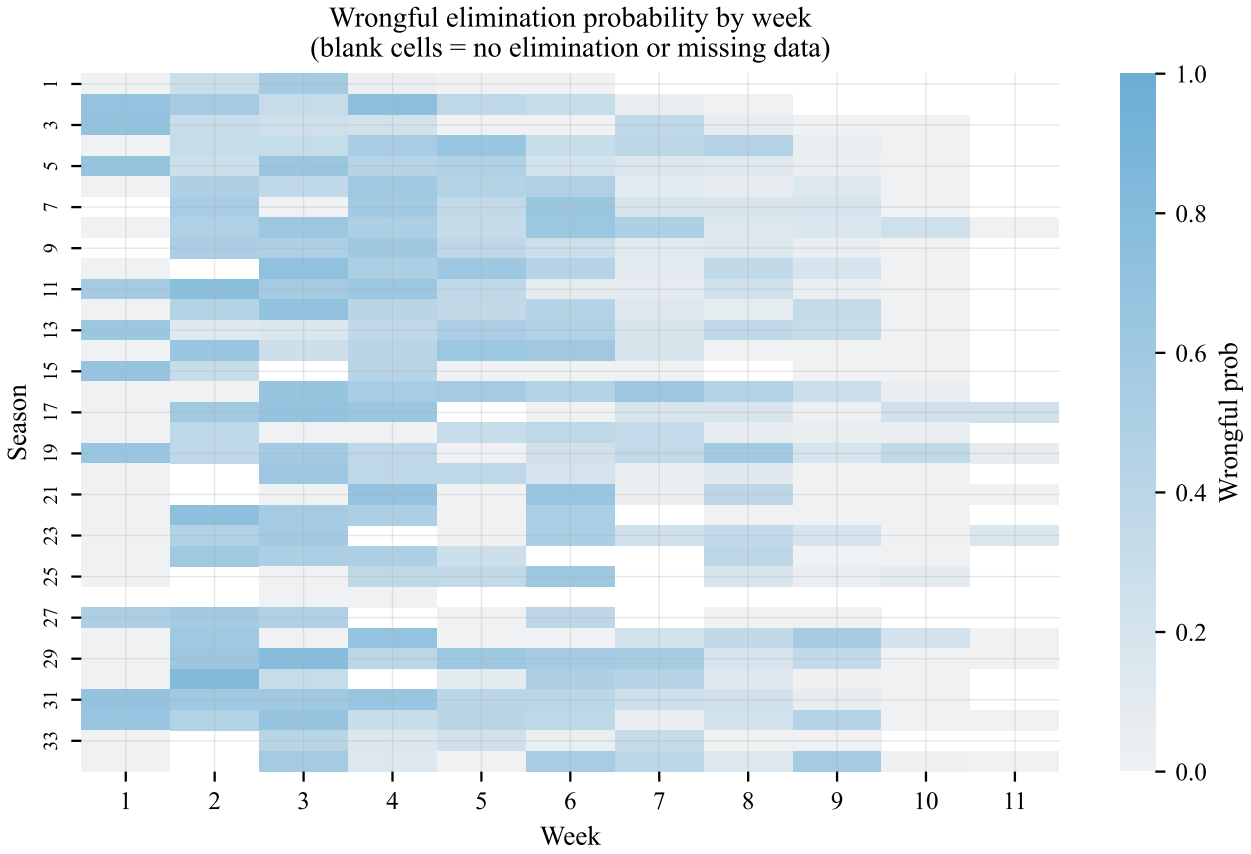


Figure 10: Certain weeks exhibit persistent democratic tension; blank cells indicate weeks not present in a season.

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 τ), 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.

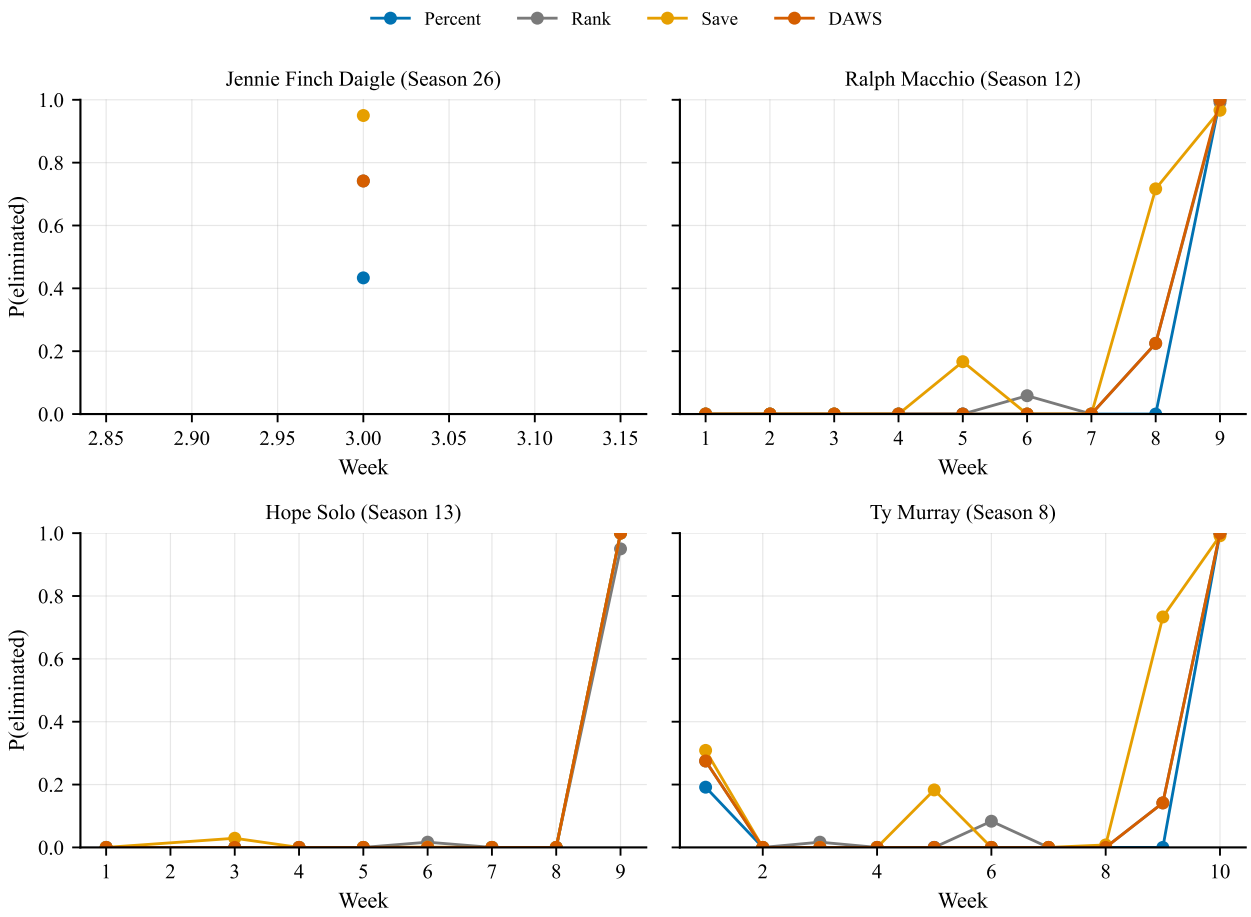


Figure 11: Counterfactual elimination risk over weeks for high-profile cases (percent, rank, judge-save, and DAWS).

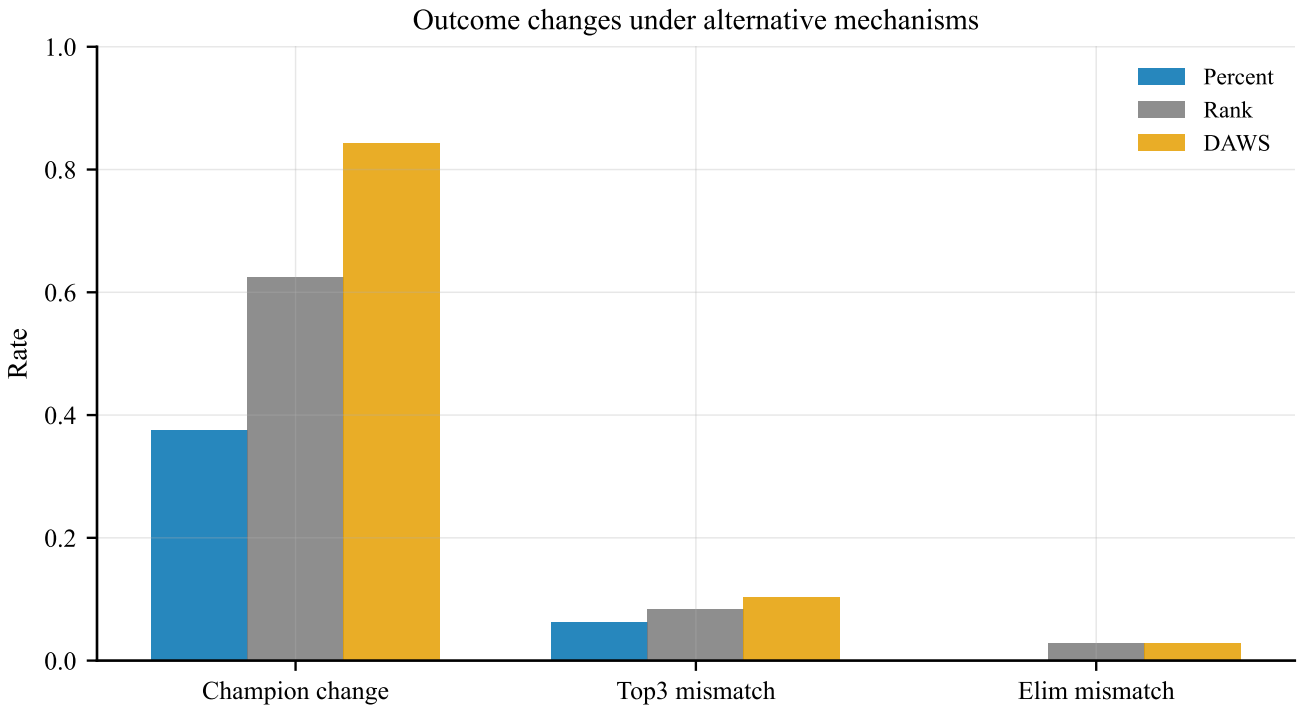
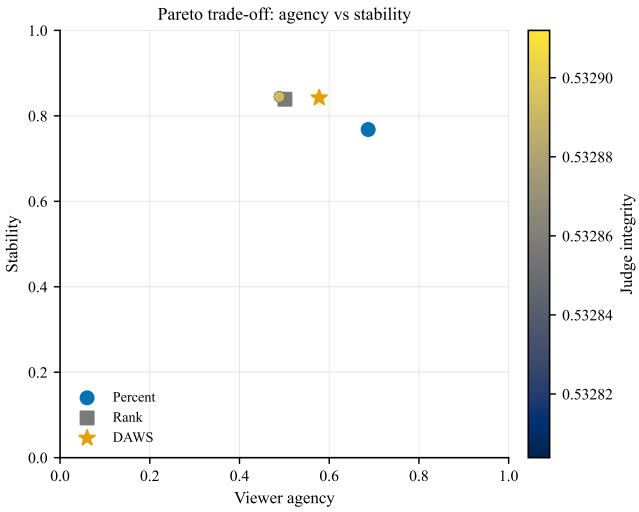
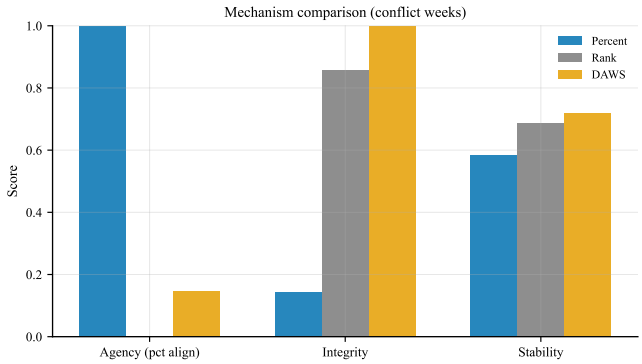


Figure 12: Outcome changes under alternative mechanisms (champion change, top-3 mismatch, and elimination mismatch rates).



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



(b) Numeric comparison across mechanisms (conflict weeks only).

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:

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

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

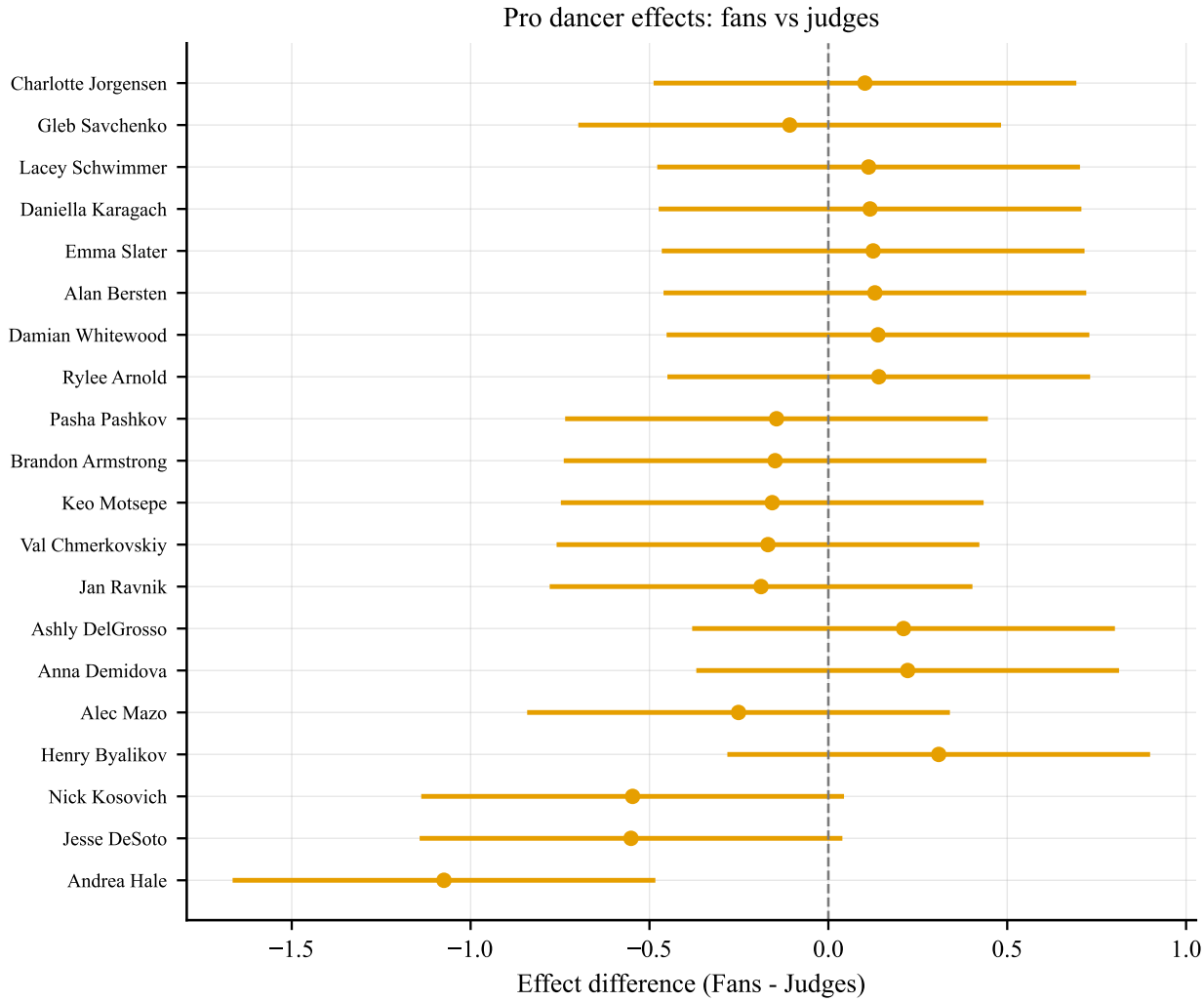


Figure 13: Pro dancer effects (fans minus judges).

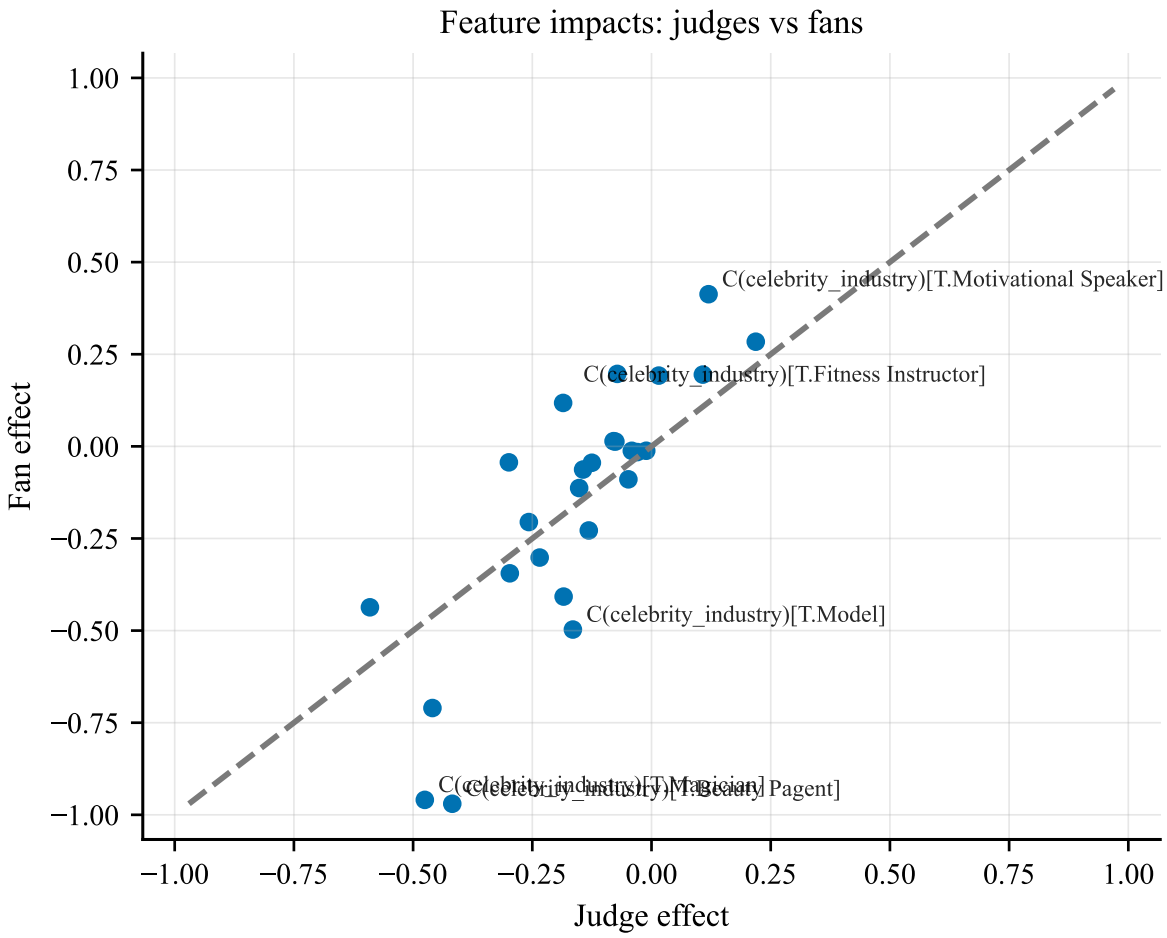


Figure 14: Annotated outliers highlight features with the largest judge-fan gaps.

Predictive add-on (Appendix). We place the GBDT robustness check in Appendix D; 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 conflict-triggered protocol that patches rule disagreement.

We define the democratic deficit as $D = \Pr(E_t^{(\text{rank})} \neq E_t^{(\text{percent})})$ and use this conflict as the trigger. DAWS runs in two operational modes plus a finale override:

- **Consensus (A=0).** If Percent and Rank agree, follow Percent (50/50) to preserve viewer agency.
- **Conflict (A=1).** If they disagree, activate judge-save between the two candidates to restore integrity.

- **Finale (Red).** Audience-only voting.

Intervention is triggered solely by A_t (rule conflict); V_t governs disclosure/audit budget only. We retain U_t as a monitoring signal for the dashboard; Fig. 15 shows U_t with P85/P95 bands for transparency.

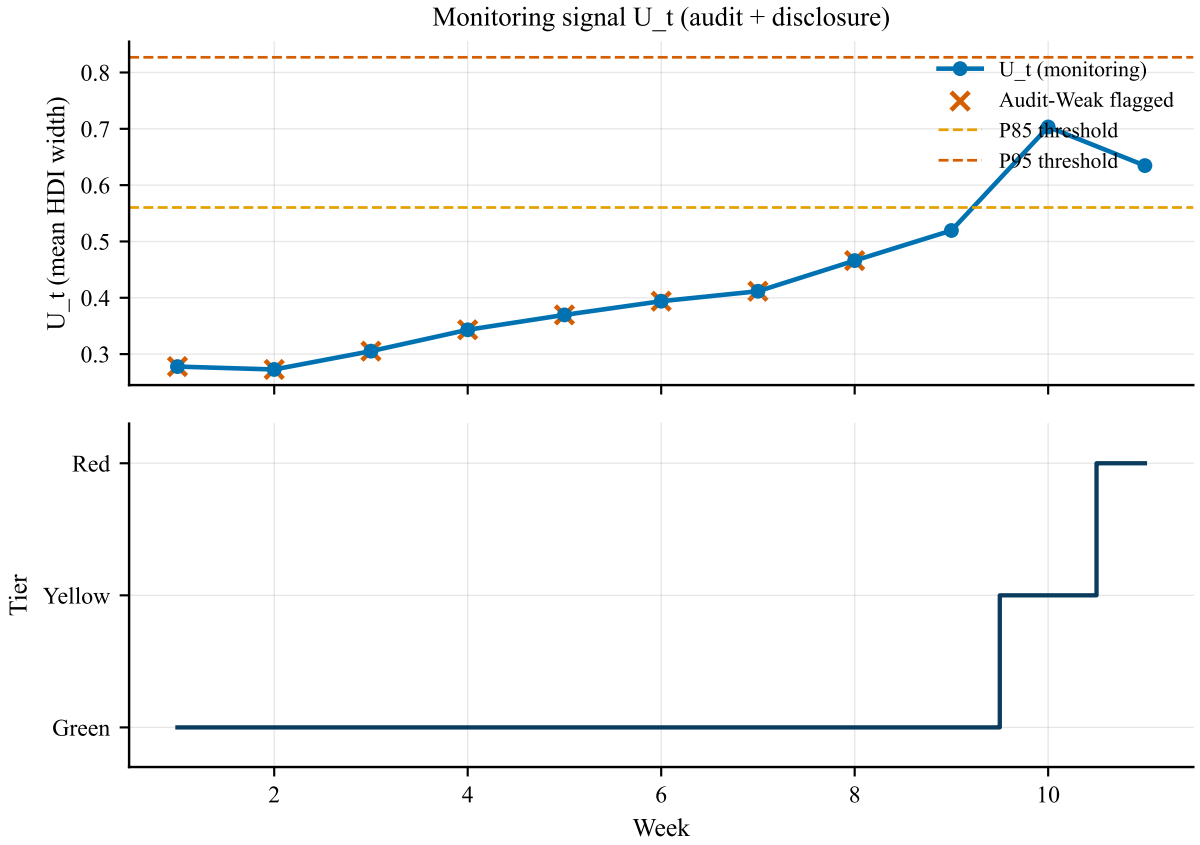


Figure 15: DAWS monitoring panel: weekly uncertainty U_t with dashed P85/P95 bands for transparency; activation is conflict-triggered.

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 β in

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

In conflict weeks, we treat judges as decisive gatekeepers and set $\beta = 6.0$ to reflect a strong corrective response against popularity bias.

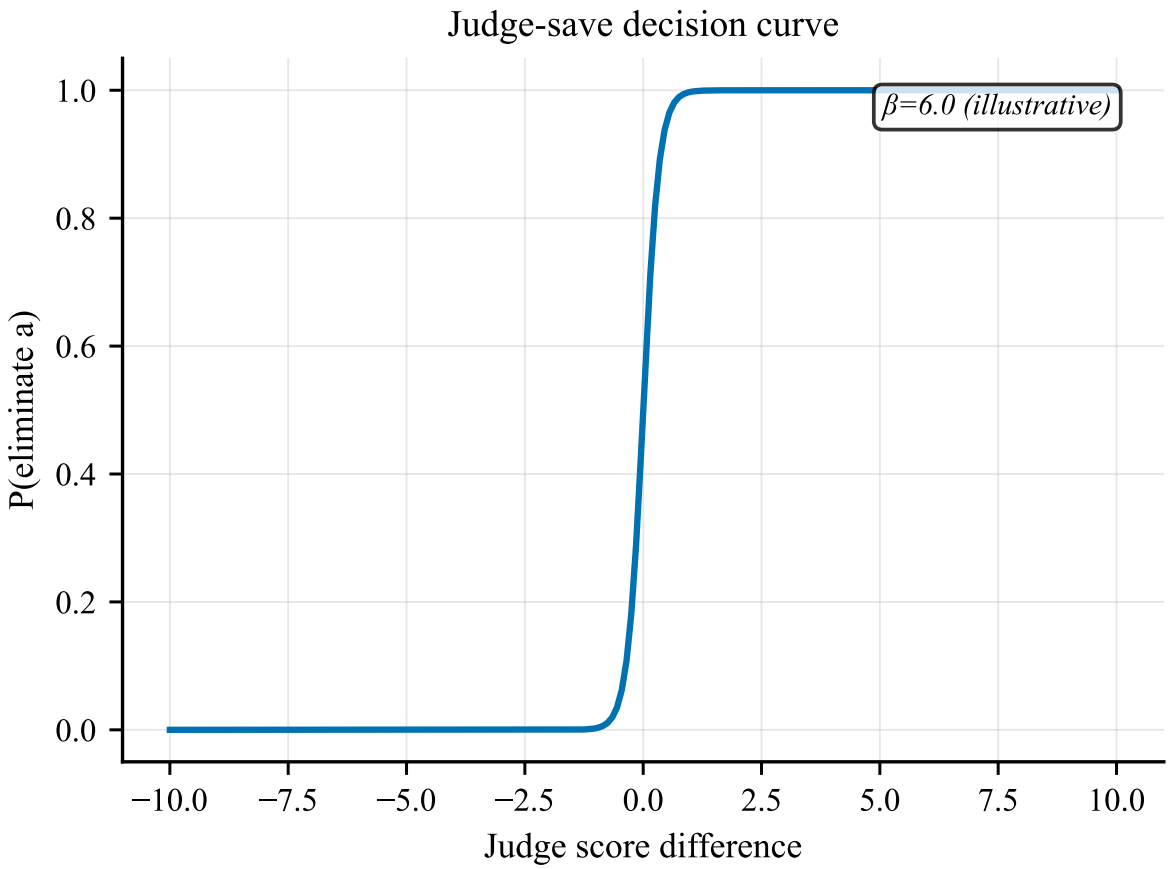


Figure 17: Judges prefer higher score within the bottom two; the curve uses calibrated $\beta = 6.0$ to illustrate conflict-week decision sensitivity.

Key Output. Conflict-triggered DAWS protocol and calibrated judge-save behavior.

9 Sensitivity and Validation

Takeaway. Key claims are stable to σ, ϵ, and rule-switch priors.

We vary σ (smoothness), ϵ (vote floor), and ρ (switch probability). Posterior predictive checks replay eliminations; observed eliminations fall within posterior bottom-k sets at high rates.

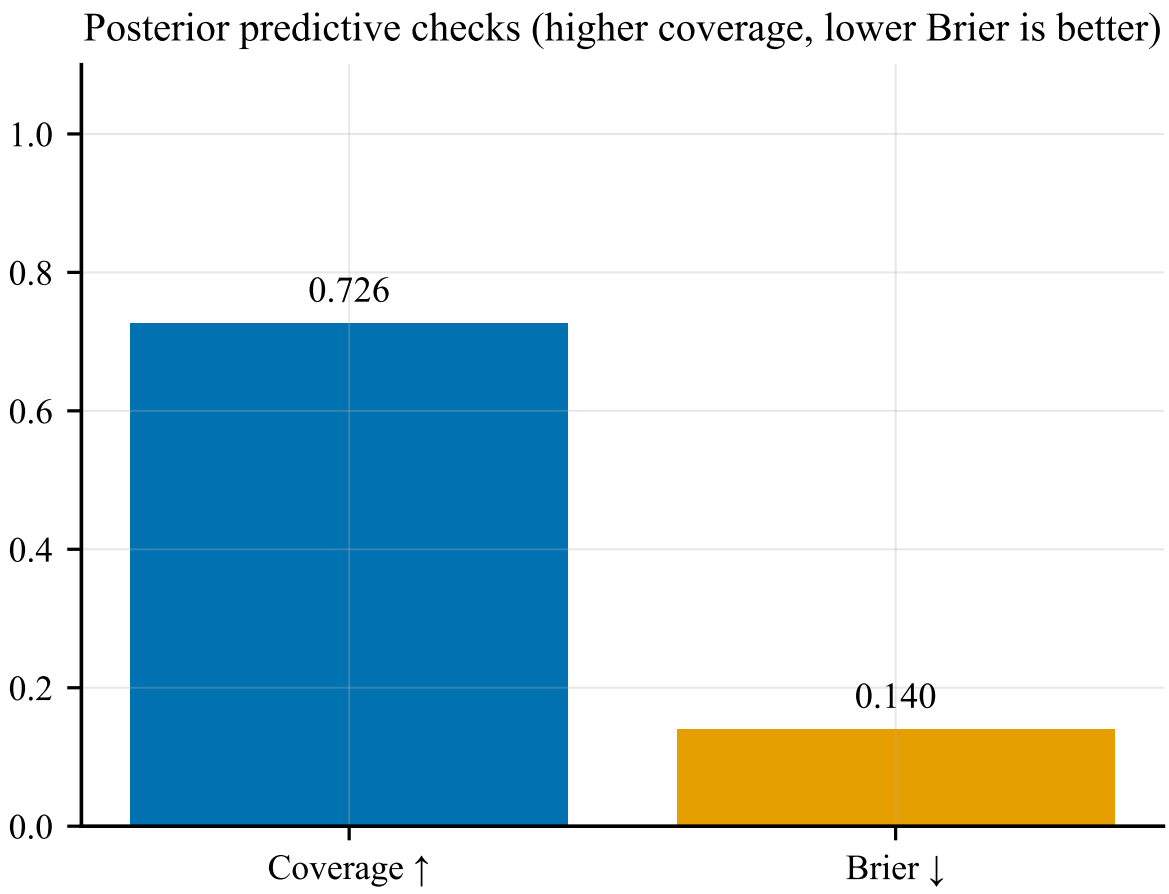


Figure 18: Model reproduces eliminations while preserving uncertainty.

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).

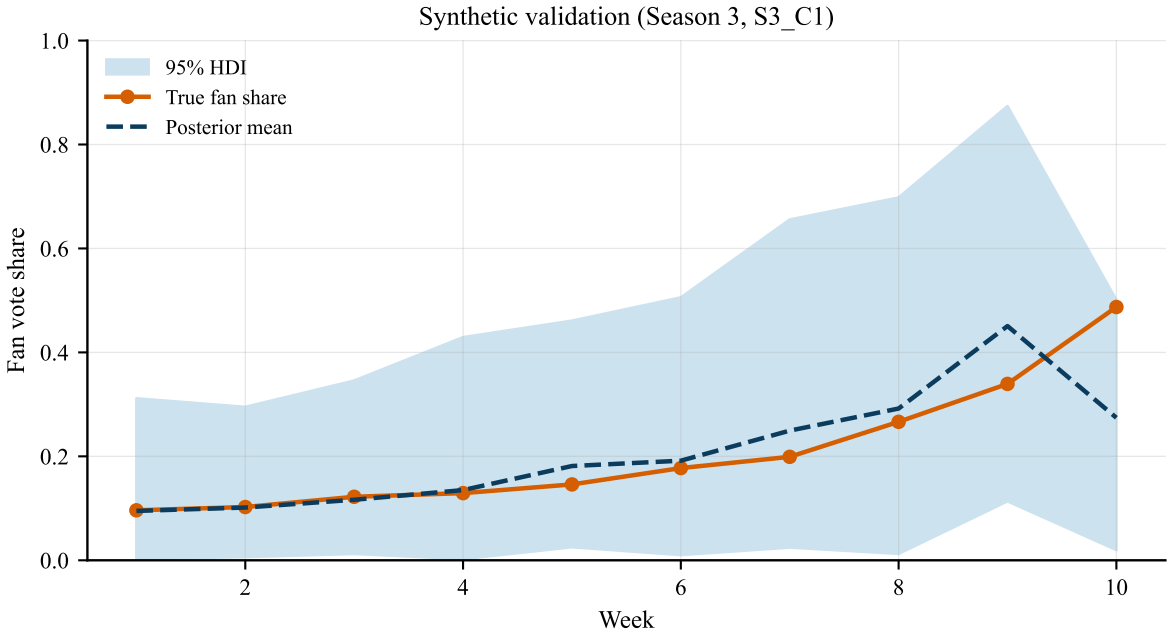


Figure 19: Synthetic validation: true fan share (red) lies within the 95% HDI band (blue) under a high-noise stress test.

Judge-save intensity sensitivity. We evaluate β on conflict weeks only. Fig. 20 reports the decision curve and the integrity–agency trade-off; $\beta = 6.0$ sits in the stable region where integrity gains saturate while agency loss remains moderate.

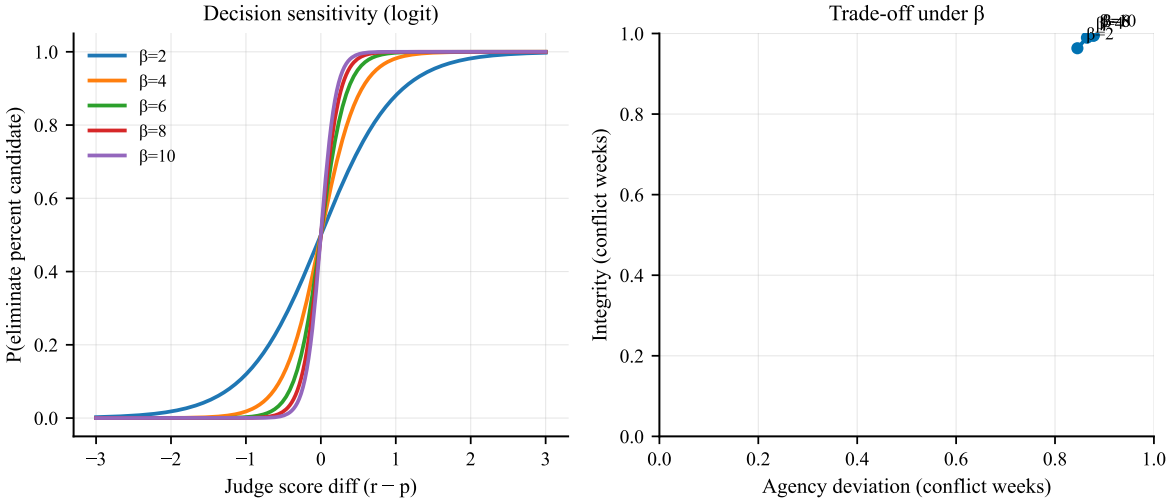


Figure 20: Sensitivity of judge-save intensity β (conflict weeks only): logit decision curves and the integrity–agency trade-off.

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 τ). The curves show diminishing returns in uncertainty

reduction beyond mid-scale settings; the elbow (dashed line) marks our final scale choice.

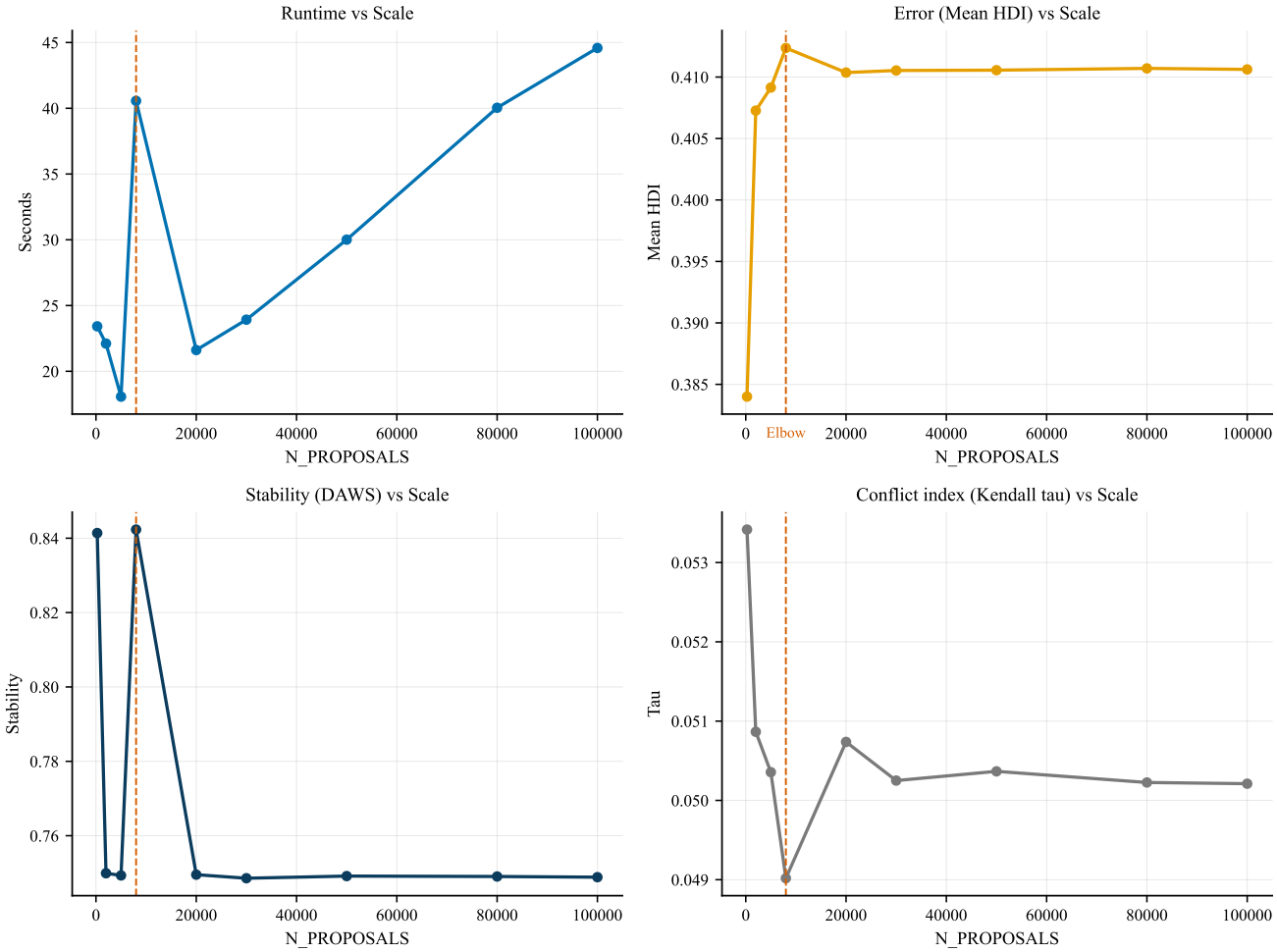


Figure 21: Scale benchmark across $N_{\text{proposals}}$ with runtime, error, stability, and theory-fit.

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.
- **Methodological limitations:** Double-elimination weeks have smaller feasible regions (acceptance rate $R \approx 30\%$ of single-elimination), consistent with combinatorial expectation (Appendix B); LP/MILP is used for diagnostics only (Appendix C).

A Sensitivity Analysis

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

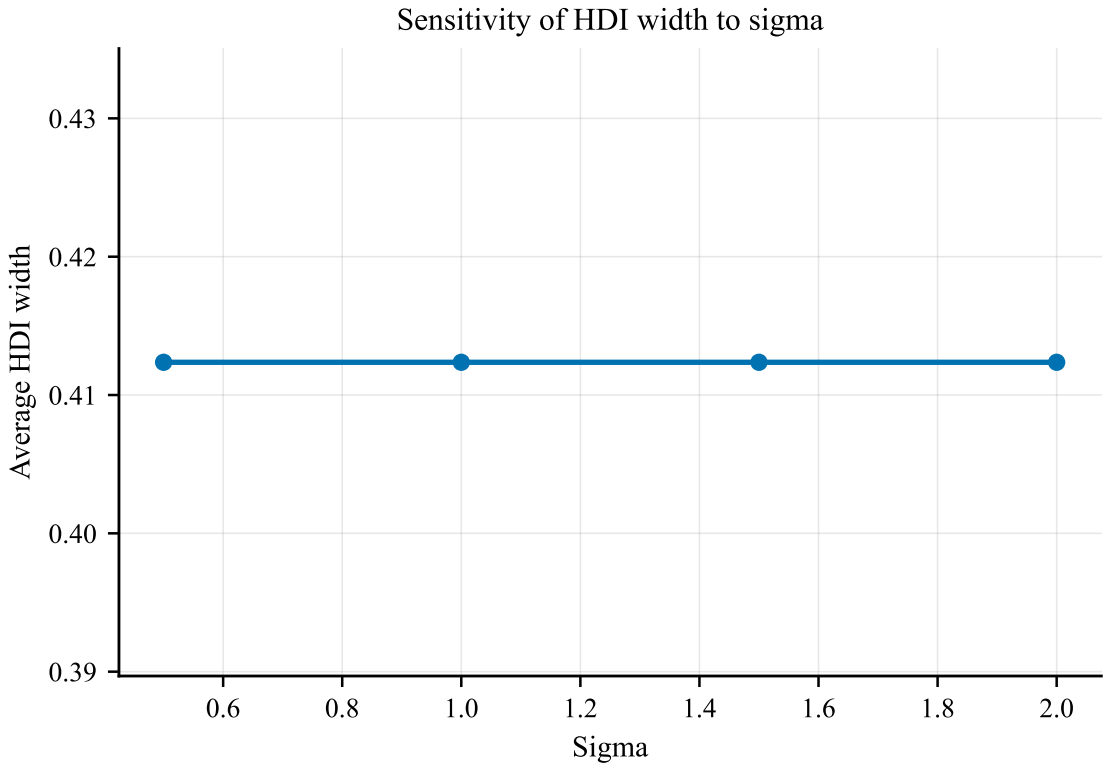


Figure 22: Sensitivity of key metrics to smoothness parameter σ. Conclusions are robust across the tested range.

B Double-Elimination Week Feasibility Verification

This appendix reports feasibility verification results for double-elimination weeks ($k = 2$) versus single-elimination weeks ($k = 1$).

Pre-registered Smoke Test (R-ratio) We pre-registered $R = \text{median}(\text{accept\_rate}_{k=2})/\text{median}(\text{accept\_rate}_{k=1})$ with expected range $[0.5, 2.0]$. Observed $R = 0.28$, **FAIL**. This result is retained (not adjusted) because $R < 0.5$ is mathematically expected: double-elimination constraints are stricter, yielding smaller feasible regions.

Theoretical Baseline Alignment From combinatorics: $R_{\text{baseline}} = 2/(n-1)$ where n is the number of active contestants. For median $n = 8$, $R_{\text{baseline}} = 0.286$. Observed $R/R_{\text{baseline}} \approx 1$, confirming the ratio matches theoretical expectation rather than an implementation bug.

Structural Recomputation (Primary Evidence) To avoid circular self-verification, we explicitly recompute combined scores and residuals (residual = $\max(\text{score}_E) - \min(\text{score}_S)$) on accepted samples, **without calling** `strict_feasible_mask`. Results:

- Double-elimination weeks: total=31, valid=15, checked=15, min_check_rate=1.0 ✓
- Single-elimination weeks: total=232, valid=225, checked=225, min_check_rate=1.0 ✓

All accepted samples pass independent recomputation, verifying constraint encoding correctness.

C LP/MILP Scope Declaration

This appendix clarifies the role of LP/MILP in this paper: **diagnostic tool only**, not part of the main inference pipeline.

Main Model Implementation MaxEnt Dirichlet sampling + `strict_feasible_mask()` constraint filtering. Posterior is a sample-based approximation, not an LP optimization solution.

LP/MILP Usage (Diagnostic)

- `slack_cache`: audit metadata (constraint tension indicator)
- Bounds visualization: reference only, not primary conclusions

LP/MILP NOT Used For Sampling filtering, feasibility checking, posterior estimation, or mechanism metric computation. Current implementation is a stub (returns fixed slack=0.001).

See output file `audit_lp_milp_scope.json` for machine-verifiable declaration.

D Predictive Calibration

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

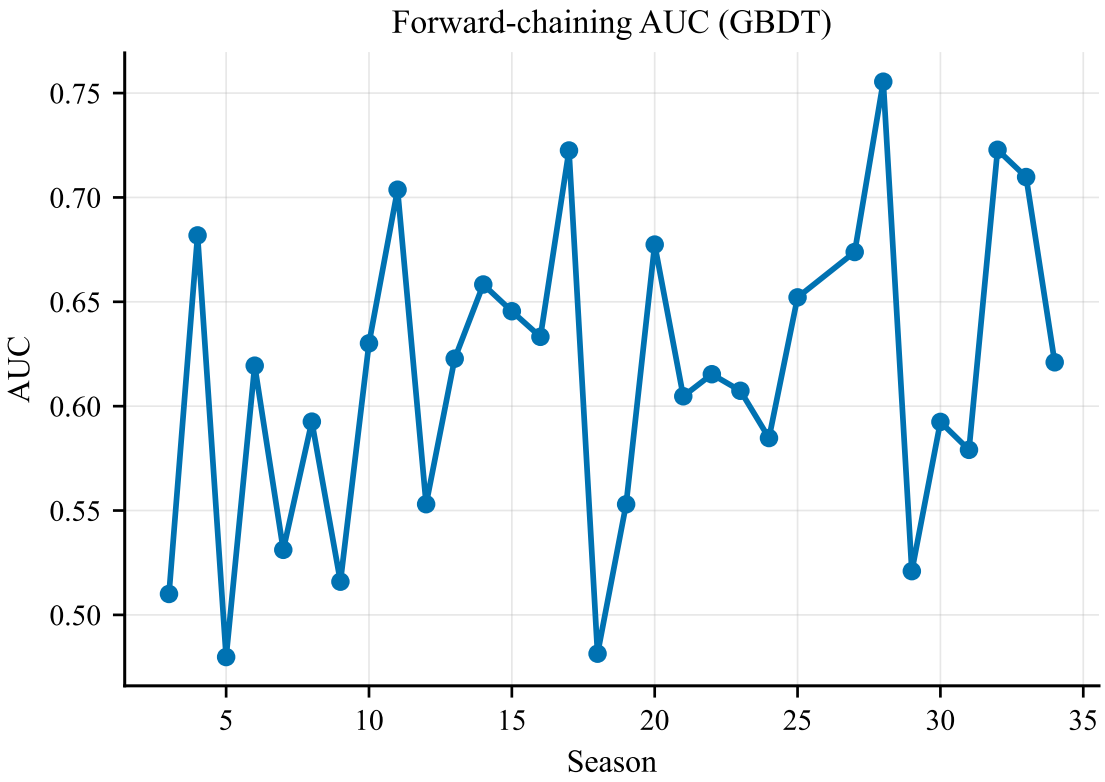


Figure 23: Forward-chaining AUC curve. Predictive performance is stable and supports the selected covariates.

E Audit Parameter Specification Table

All parameters below are fixed *a priori*; they are not tuned based on results.

Table 1: Fixed audit parameters (Block 5 specification).

Parameter	Value	Type	Description
ε_{sum}	10^{-9}	Audit rule	Simplex sum tolerance
ε_{ord}	10^{-6}	Audit rule	Elimination ordering tolerance (continuous)
$\varepsilon_{\text{rank}}$	0	Audit rule	Rank ordering tolerance (discrete; ties allowed)
$N_{\text{strict,min}}$	500	Audit rule	Minimum strict feasible samples
q_{gate}	0.10	Audit rule	Quantile for budget calculation
$r_{\text{excl,max}}$	20%	Reporting rule	Downgrade trigger threshold
$N_{\text{proposals}}$	8000	Compute budget	Default proposal count (main)
$N_{\text{proposals,fast}}$	2000	Compute budget	Fast mode (smoke tests only)

Interpretation.

- **Audit rules** define feasibility and output eligibility; they are not optimization targets.
- **Compute budget** affects Monte Carlo precision but is not a model assumption.

- **Reporting rule**: The 20% threshold triggers downgrade, not parameter re-tuning.
- **Ties allowed**: Eliminated contestants may tie with survivors ($\max E = \min S$).
- **Excluded weeks**: Flagged as Audit-Weak, shown as marked points, not in aggregates.
- **Worst-case disclosure**: Weeks with accept rate ≈ 0.008 may be excluded; this is expected.
- **Diagnostic score source**: Percent-rule diagnostics use `judge_share` as an audit proxy for the combined percent score used in our elimination model.
- **Rank threshold**: $\varepsilon_{\text{rank}} = 0$ means eliminated contestants' rank must be $\leq$ survivors' rank (ties allowed).

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.