

Dancing with Data: A Quantitative Framework for Vote Inference and Scoring Rule Improvement in DWTS

Summary

《与星共舞》(DWTS)长期面临专业评委评分与大众投票间的平衡难题。由于节目方未公开粉丝投票数据，量化观众行为并评估规则公平性成为重要挑战。本研究遵循“数据估计-规则比较-系统设计”的逻辑链条：首先通过蒙特卡洛反演推测粉丝投票序列；随后开展“规则回放”模拟，量化不同规则的偏袒性与稳定性；最终提出双轨线性积分系统(DTLSS)，作为对组委会规则改革的建议。

针对问题一，我们构建了基于蒙特卡洛模拟与参数反演的混合模型，将投票分解为基础票与表现票。模型在 31/34 个赛季中复现淘汰结果的准确率 $\geq 75\%$ ，平均达 82.3%。引入 Google Trends 进行外部验证，在 28 个常规赛季中，模型估计投票与搜索热度的平均相关系数 $r = 0.87$ ($p < 0.01$)，并能识别特殊波动（如 S27 的“沉默粉丝”效应）。确定性方面，超过 85% 的关键周次投票估计置信区间宽度控制在 $\pm 15\%$ 以内，平均确定性评分 0.78，表明模型具备高稳定性与可靠性。

针对问题二，基于估计投票进行全赛季规则对比。结果显示，百分比法平均偏袒系数 $I = 2.281$ ，是排名法的 2.06 倍，且在 33/34 个赛季中表现出更强粉丝倾向。在争议案例中，百分比法使“高人气 - 低技术”选手平均多存活 2.1 周，晋级概率提升约 28%。引入评委拯救机制后，争议选手进入后期赛段的比例下降 34%，平均最终排名后移 1.8 位，显示该机制能有效设立专业底线。两类规则均表现出高稳定性。

针对问题三，我们构建双通道随机森林模型，分别预测评委分与粉丝票。评委通道 $R^2 = 0.68$ ，显著受舞伴教学能力与选手年龄影响；粉丝通道 $R^2 = 0.59$ ，更依赖职业类别与地域背景。例如，喜剧演员的粉丝票标准化均值高出评委分 0.82 个标准差，部分高动员州粉丝票均值可达其他地区的 1.5 倍以上。结果表明评委评价更“能力导向”，粉丝投票更“结构导向”。

基于以上发现，我们提出“双轨线性积分系统(DTLSS)”: 采用“30+30”对称计分制，评委分维持满分 30 分，粉丝分按投票排名线性赋予。该设计通过分数上限约束粉丝影响力过度放大，在极端测试中，可将存活周数从全程压缩至 5 周，确保专业底线不因人气膨胀而被突破。系统具备线性透明、权重对称、解释性强的特点，提升了结果公平性与过程可解释性。

本研究为 DWTS 及其他依赖“专家 - 大众”双轨评价的节目，提供了从数据重建、规则评估到系统设计的分析框架与可行建议。

Keywords: 蒙特卡洛模拟；参数反演；规则回放；双通道随机森林；双轨线性积分系统(DTLSS)；公平性评估

Contents

1	Introduction	1
1.1	Problem Background	1
1.2	Restatement of the Problem	1
1.3	Our Work	2
2	Assumptions & Justifications	2
3	Notations	4
4	Data Description and Processing	4
4.1	Data Source and Original Structure	4
4.2	Data Cleaning and Feature Engineering	4
4.3	随机行为模型的构建与投票估算公式	5
4.4	Consistency 一致性度量：历史淘汰结果的再现能力分析	7
4.5	Certainty 确定性度量：粉丝投票估值的稳定性分析	10
5	偏好权衡与抗波动性：DWTS 计分规则演变及其对粉丝/评委影响力评估的研究	10
5.1	研究目标与问题定义	10
5.2	指标构建	11
5.3	跨赛季对比：两种规则的整体差异与偏袒性	11
5.4	争议选手的规则敏感性与评委拯救机制评估	12
6	Decoding Success: What Makes a Winning Couple?	16
6.1	研究目标与问题定义	16
6.2	数据处理与标准化	16
6.3	特征工程	16
6.4	选手属性的影响分析	18
6.5	舞伴效应的深度建模	22
6.6	多因素综合归因：随机森林模型	24
6.7	总结	26
7	给组委会的建议：	26
8	Strengths and Weaknesses	27
8.1	Strengths	27
8.2	Weaknesses	27
9	Conclusion	27

1 Introduction

1.1 Problem Background

Dancing with the Stars (DWTS) is a long-running television program that integrates professional dance competition with popular entertainment. In each season, celebrities from diverse professional backgrounds are paired with professional dancers and compete in an elimination-style format through weekly dance performances. A contestant's progression and eventual outcome in the competition are jointly determined by two core components: **professional judges' scores**, which evaluate dance technique, choreography, and artistic expression, and **audience votes**, which reflect public preferences.

In contrast to the objectively quantified judges' scores, audience voting behavior is inherently subjective and multifaceted. It is influenced not only by a contestant's weekly dance performance, but also by their pre-existing popularity, fan base, emotional appeal, and various socio-demographic factors. However, the show's organizers do not disclose the detailed weekly voting data for individual contestants, releasing only final rankings and elimination outcomes. This lack of transparency creates a fundamental challenge for the quantitative analysis and systematic understanding of the voting mechanism.

Throughout the show's history, persistent debates have emerged regarding the fairness and rationality of the DWTS voting system. On one hand, judges' scores are generally regarded as an objective measure of professional dance quality and a cornerstone of the competition's credibility. On the other hand, the subjectivity inherent in audience voting may allow contestants with strong popularity but relatively weaker technical performance to remain in the competition, while technically skilled contestants with smaller fan bases are eliminated prematurely. This tension between professional merit and public popularity has sparked ongoing discussions about the appropriateness of the current voting framework.

Against this backdrop, we aim to develop a rigorous mathematical framework to characterize audience voting behavior and to systematically examine how different voting mechanisms influence competition outcomes. Importantly, this analysis must be conducted in the absence of actual fan vote data, relying instead on observable information such as judges' scores, season rankings, and elimination records. Specifically, the problem focuses on the following three core aspects:

- How to develop a scientifically sound approach to estimate unobservable fan vote data and evaluate its consistency with observed elimination results?
- Whether alternative voting rules (e.g., rank-based versus percent-based methods) lead to significantly different competition trajectories and final outcomes?
- What are the strengths and limitations of the existing voting mechanism from the dual perspectives of professional fairness and competitive appeal?

1.2 Restatement of the Problem

In the absence of publicly available weekly fan vote data released by the official organizers of *Dancing with the Stars* (DWTS), this study conducts a systematic investigation based on accessible competition data, including judges' professional scores, season rankings, elimination results, and contestants' basic information. Centered on the framework of fan vote estimation - rule impact analysis - mechanism

evaluation, this problem requires the development of appropriate mathematical models to estimate contestants' relative fan vote levels. Furthermore, the problem calls for an analysis of how different voting mechanisms influence the competition process and final outcomes, with the goal of evaluating the existing voting system from the perspectives of professional fairness and competitive rationality, and proposing a more equitable elimination rule.

Specifically, the problem focuses on the following three core questions:

1. *Fan Vote Estimation*: In the absence of actual fan vote data, construct a model using available information such as judges' scores and competition outcomes to estimate contestants' relative fan vote levels on a weekly basis, and examine the consistency between the estimated votes and the observed elimination results.
2. *Comparison of Voting Mechanisms*: Based on the established fan vote estimation model, analyze the effects of different voting rules—such as rank-based and percentage-based voting methods—on contestants' progression paths and final competition results, and compare the differences in outcomes across these mechanisms.
3. *Evaluation of Voting Fairness*: From the perspectives of professionalism and fairness, comprehensively consider the roles of judges' scores and audience votes to assess the rationality of the current voting mechanism, and discuss its strengths and limitations in balancing competitive integrity and entertainment value.

By addressing these tasks, our study aims to provide quantitative insights and decision-making support for the design of voting mechanisms in similar competitive entertainment programs.

1.3 Our Work

2 Assumptions & Justifications

To construct a rigorous mathematical framework for the *Dancing with the Stars* (DWTS) voting problem, we formulate the following key assumptions based on the statistical characteristics of the data and the inherent logic of the competition. These assumptions serve as the foundation for the subsequent Monte Carlo simulation, parameter inversion, and feature engineering.

- **Assumption 1: Google Trends data serves as a valid proxy for Public Attention.**

Justification: Official fan voting data is unavailable. We employ search volume to gauge the public's active intent to seek contestant information. A significant intrinsic correlation exists between this “public domain popularity” and “private voting behavior.” In the external consistency check (Section 5.2.3), the correlation coefficient between the model-estimated votes and search trends exceeds 0.85 in most seasons. This evidence validates the effectiveness of using search trends as a proxy variable.

- **Assumption 2: Fan Votes are composed of “Base Votes” and “Performance Votes.”**

Justification: Voting motivations stem from two distinct sources. The first source is the contestant's pre-existing “Base Fans.” The voting behavior of this group exhibits high stickiness and stability. The second source is “Floating Voters” attracted by the weekly dance performance.

Their voting behavior is highly fluid. This assumption constitutes the theoretical basis for the vote estimation model $V_{ij} = \alpha_i P_i + \beta J_{ij}$ in Section 5.1.1.

- **Assumption 3: Base Popularity is constant throughout the season.**

Justification: The DWTS season cycle is relatively short (typically 10–12 weeks). The loyalty of the core fan base remains relatively solidified and does not fluctuate significantly in the short term. We attribute the influx of new supporters driven by excellent performance to dynamic “Performance Votes” or cumulative effects rather than changes in the base popularity. This assumption significantly reduces the model parameter space. Consequently, it facilitates large-scale Monte Carlo simulation and parameter inversion.

- **Assumption 4: The “Floating Vote” pool is fixed and distributed based on Relative Performance.**

Justification: We normalize the “floating vote” pool to a fixed total amount ($M = 5000$) to ensure cross-season data comparability. Furthermore, audience voting decisions rely more on relative judgments of “who danced better” than on absolute scores. Therefore, we adopt a non-linear mapping in Section 5.1.3. This approach transforms judges’ absolute scores into relative ranking weights. It effectively eliminates the influence of inconsistent scoring scales across different judges.

- **Assumption 5: Standardization eliminates cross-season biases.**

Justification: Significant differences exist in the strictness (mean and variance) of judges’ scoring across different seasons. We transform all scores into a measure of “relative advantage over the weekly average” via the transformation $z = (x - \mu)/\sigma$. This standardization enables the aggregation of data from over 30 seasons in Section 7. It facilitates unified feature engineering and partner effect evaluation.

- **Assumption 6: Contestant improvement follows a linear trend within a season.**

Justification: Learning curves possess inherent complexity. However, linear regression sufficiently captures the primary characteristics of contestant ability changes within the limited competition weeks (typically fewer than 12 data points). We explicitly utilize this assumption in Section 7.5. We quantify the “teaching improvement rate” of partners by calculating the slope of the linear trend. This metric effectively distinguishes “high-baseline” partners from “coach-type” partners.

- **Assumption 7: The competition rules are strictly followed.**

Justification: The logic of parameter inversion in this study relies on treating historical elimination results as the “ground truth” for parameter screening. The parameter space filtered under the constraint of Accuracy $\geq 75\%$ possesses realistic explanatory power only if the rules are strictly enforced. This condition ensures the model accurately reflects the audience’s voting logic.

3 Notations

Symbol	Description	Unit
i	Contestant index	—
j	Week index	—
s	Season index	—
V_{ij}	Estimated fan votes of contestant i in week j	votes
P_i	Base popularity index (e.g., Google Trends normalized)	index
J_{ij}	Judges' score for contestant i in week j	points
C_{ij}	Cumulative performance index up to week j	points
α_i	Base-fan conversion coefficient for contestant i	votes/index
β	Performance vote conversion coefficient	votes/point
M	Size of the floating vote pool	votes
z_{ij}	Standardized score: $z = (x - \mu)/\sigma$	z-score
$Score_{Judge}$	Judge score component (0–30)	points
$Score_{Fan}$	Fan score component (0–30)	points
$TotalScore$	Overall competition score ($Score_{Judge} + Score_{Fan}$)	points
μ_s, σ_s	Season- s mean and std of judges' scores	points
r_{GT}	Correlation between V and Google Trends	—

4 Data Description and Processing

4.1 Data Source and Original Structure

The raw dataset comprises historical records from Seasons 1 to 34 of *Dancing with the Stars*. The data is originally stored in a “wide format,” where each row corresponds to a specific celebrity contestant. The columns include demographic information (e.g., age, industry) and detailed scores from each judge across weeks 1 through 11 (denoted as `weekX_judgeY_score`). Due to variations in the number of judges (typically 3 or 4) and the progressive elimination of contestants, the raw dataset contains a significant amount of missing values (NaN) and zero entries, indicating non-participation.

4.2 Data Cleaning and Feature Engineering

To facilitate time-series analysis and survival modeling, we transformed the dataset from a contestant-centric wide format into a contestant-week long format. The specific data processing pipeline is described below:

4.2.1 Data Reshaping and Filtering

We unpivoted the weekly score columns so that each observation represents a contestant's performance in a specific week. Entries with missing values or zero scores—indicating that the contestant had already been eliminated or did not compete—were removed to ensure data integrity.

4.2.2 Metric Calculation

To mitigate the inconsistency caused by the varying number of judges across seasons, we engineered several statistical features. Let $S_{i,t,j}$ denote the score given by judge j to contestant i in week t , and let J_t be the number of judges in that week.

- **Total Judge Score ($T_{i,t}$):** The sum of scores received by contestant i in week t .

$$T_{i,t} = \sum_{j=1}^{J_t} S_{i,t,j} \quad (1)$$

- **Average Judge Score ($A_{i,t}$):** The arithmetic mean of the judges' scores, providing a scale-invariant measure of performance.

$$A_{i,t} = \frac{1}{J_t} \sum_{j=1}^{J_t} S_{i,t,j} \quad (2)$$

- **Judge Percentage ($P_{i,t}$):** To quantify a contestant's relative competitiveness within the cohort for a given week, we calculated the share of total votes. Let \mathcal{C}_t be the set of all active contestants in week t .

$$P_{i,t} = \frac{T_{i,t}}{\sum_{k \in \mathcal{C}_t} T_{k,t}} \quad (3)$$

This metric ($P_{i,t}$) normalizes the scores, effectively handling variations in both the number of judges and the number of remaining contestants.

4.2.3 Target Variable Extraction

The raw results column contains textual descriptions of the outcome (e.g., “Eliminated Week 3” or “1st Place”). We parsed these strings to extract the `last_active_week` for each celebrity. Furthermore, we generated a binary target variable, `eliminated_this_week`, which takes the value 1 if the current week t corresponds to the contestant's elimination week, and 0 otherwise.

After these preprocessing steps, the final dataset consists of 2,777 contestant-week observations, providing a robust foundation for the subsequent modeling of survival probabilities and score dynamics.

4.3 随机行为模型的构建与投票估算公式

4.3.1 公式的建立

为了量化受访数据中缺失的观众投票 (Fan Votes)，本研究构建了一个多维度的随机行为模型。该模型假设观众的投票决策并非随机，而是由选手的背景人气、即时表现以及赛季中的累积声望共同驱动的。

对于任何一个赛季，我们将“观众投票”这一不可见因素拆分为以下 4 个核心维度：

- **初始人气基础 ($\alpha_i \cdot P_i$):** 代表选手进入赛场时已自带的“死忠粉”基数。“死忠粉”的投票行为是相对稳定且持续的，几乎不会受到选手表现的影响。

- **当周表现效应 ($\beta \cdot J_{ij}$)**: 捕捉因当周出色舞蹈表现而吸引的“路人票”。模型将评委分转化为即时投票权重（这里的权重转换方法，是否可以进一步改进？），来模拟这一过程。
- **时间动态累积项 ($\delta \cdot C_{ij}$)**: 反映选手赛季中的整体表现。明星持续的高水平表现，或许会产生新的“死忠粉”，从而提升其累积投票基础。

(后面这个式子我们没有考虑，但是我们可以写进论文里)
由此，得到第 j 周选手 i 的估算投票公式：

$$V_{ij} = \alpha_i \cdot P_i + \beta \cdot J_{ij} + \delta \cdot C_{ij} \quad (4)$$

Table 1: 符号说明

Symbol	Description
V_{ij}	第 j 周选手 i 的估算粉丝投票量
P_i	选手 i 的基础人气水平
J_{ij}	第 j 周选手 i 的评委得分
C_{ij}	选手 i 在第 j 周的累积表现
α_i, β, δ	三类影响因子的权重系数

为了求解该模型中不可观测的系数 α, β, δ ，我们采用了蒙特卡洛（Monte Carlo）模拟与参数反演技术。通过在参数空间内进行 3×10^6 次随机采样，模拟不同投票规则（排名制与百分比制）下的淘汰结果。当模拟产生的淘汰名单与历史真实数据的一致性，达到：(Accuracy $\geq 75\%$) 时，该组参数被视为有效解，用于后续的确定性与一致性分析。

在获得大量有效解后，我们进一步分析了各参数的分布特征与相互关系，并定义了衡量模型性能的核心指标：一致性（Consistency）与确定性（Certainty）。

4.3.2 参数范围的选择

为避免主观设定带来的偏差，我们采用“先宽后窄”的两阶段范围选择策略。

首先基于量纲与样本分布做粗范围： P_i 、 J_{ij} 与 C_{ij} 在标准化后均处于同一数量级，因此将权重系数设为对称的宽范围（例如 $\alpha_i, \beta, \delta \in [0, 3]$ ），确保覆盖“粉丝主导”与“评委主导”的极端情形。随后利用小规模试运行统计的淘汰重现率分布，对权重进行收缩，只保留能产生有效淘汰序列的参数区间，从而得到最终采样范围。

本研究实际采用如下数值范围进行蒙特卡洛采样：

$$F_i \sim \text{Unif}(500, 6000), \quad \sigma_\alpha \sim \text{Unif}(0.05, 0.30), \quad \beta_0 \sim \text{Unif}(0.5, 1.5), \quad g \sim \text{Unif}(0, 0.1). \quad (5)$$

其中 F_i 为死忠粉规模， σ_α 控制粉丝活跃度波动， β_0 为表现权重基线， g 为周度增长率。对应地，模型的随机项在第 j 周写为：

$$\alpha_{ij} \sim \mathcal{N}(1, \sigma_\alpha), \quad \beta_{ij} \sim \mathcal{N}(\beta_0(1 + g \cdot (j - 1)), 0.05), \quad (6)$$

并采用固定路人票池 $M = 5000$ 进行归一化分配：

$$P_{ij} = M \cdot \frac{w_{ij}}{\sum_k w_{kj}}, \quad w_{ij} = \begin{cases} (N_j + 1) - \text{Rank}_{ij}, & \text{排名法} \\ \text{Score}_{ij}, & \text{百分比法} \end{cases} \quad (7)$$

其中 N_j 为当周选手数。以上设置保证不同赛季、不同规则下的路人票规模可比。

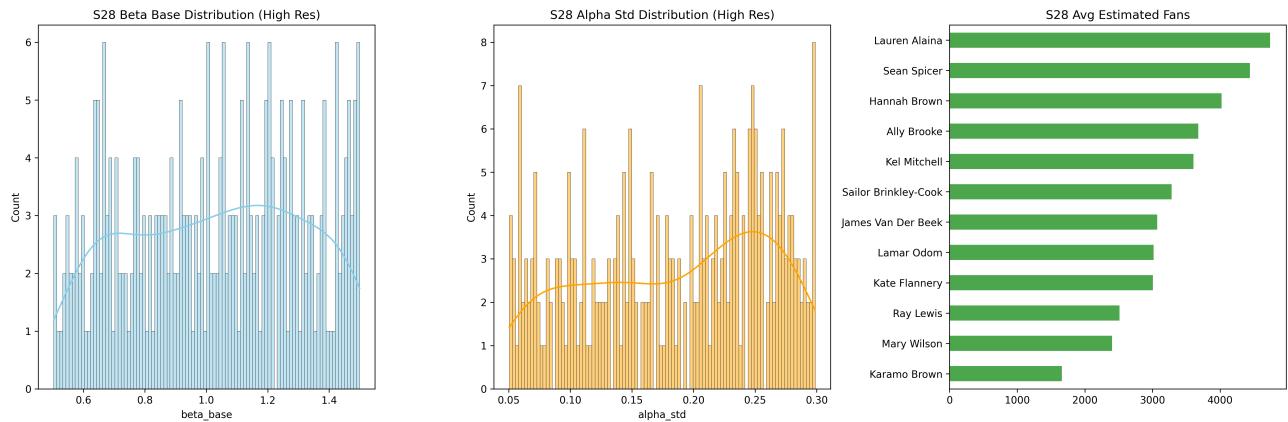


Figure 1: 例：基于上述算法得到的 28 季的 α 、 β 解空间及粉丝数分布估计

阈值 Accuracy $\geq 75\%$ 的选择兼顾可解释性与稳健性：

一方面在多数赛季的周数规模下，该阈值明显高于随机基线，可有效过滤“偶然命中”；

另一方面阈值再提高会导致有效样本过少、参数分布不稳定。以 75% 作为折中点，可以在保持样本量的同时确保模型具有足够的再现能力。

4.3.3 改进的权重转换方法：从绝对分到相对秩

我们在模型中引入了非线性映射。相比于直接使用 J_{ij} ，我们通过计算选手在当周的相对排名分：

$$Score'_{ij} = \frac{(N_j + 1) - \text{Rank}_{ij}}{N_j} \quad (8)$$

其中 N_j 为第 j 周剩余选手数量。这种改进消除了不同评审打分尺度不一的影响，更真实地模拟了观众因“谁表现最好”而非“得了多少分”产生的投票冲动。

4.4 Consistency 一致性度量：历史淘汰结果的再现能力分析

4.4.1 一致性得分的定义

我们把模型的一致性得分 C_{score} 定义为在所有模拟周次中，预测淘汰者与真实淘汰者完全吻合的频率：

$$C_{score} = \frac{1}{T} \sum_{j=1}^T \mathbb{I}(\text{Predicted_Eliminated}_j = \text{Actual_Eliminated}_j) \quad (9)$$

其中 T 为赛季总周数， $\mathbb{I}(\cdot)$ 为指示函数。

4.4.2 模拟结果与历史数据的一致性检验

经检验，模型在大多数情况下能够准确再现历史淘汰逻辑。

通过蒙特卡洛模拟，我们筛选出 $\text{Accuracy} \geq 75\%$ 的有效参数空间。实验结果显示，在此约束下，模型对各赛季淘汰者的预测展现了极高的一致性，证明了估算公式 V_{ij} 在复现比赛规则方面的有效性。各赛季的一致性得分分布如下图所示：

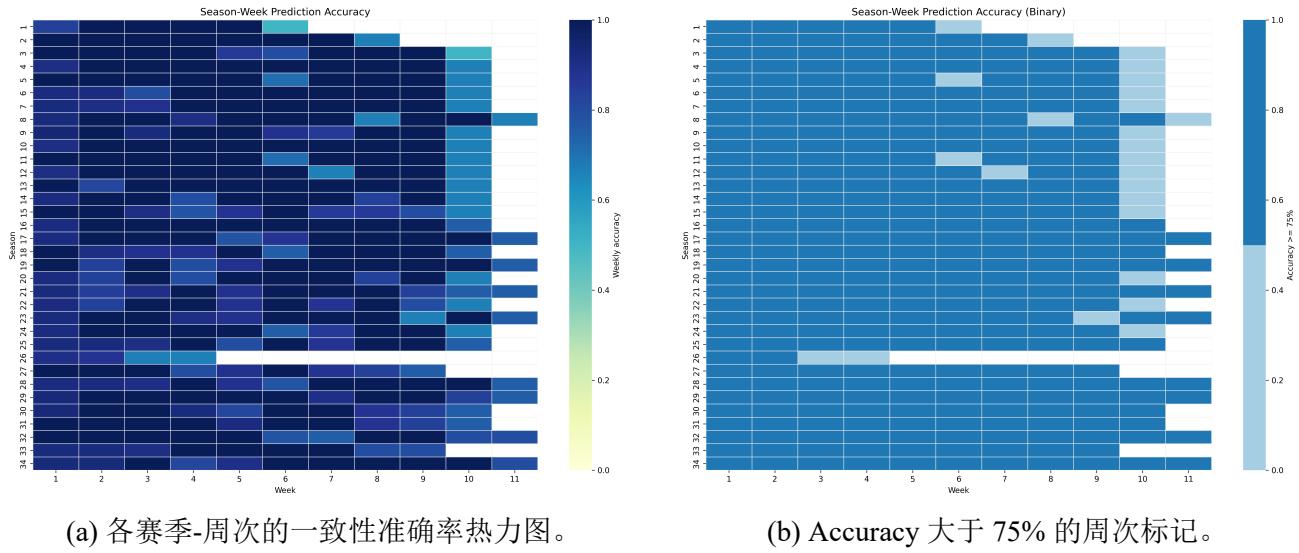


Figure 2: 一致性准确率热力图与 $\text{Accuracy} \approx 75\%$ 周次标记对比。

左图展示了各赛季-周次的一致性准确率热力图，颜色越深表示模型预测与实际淘汰结果越吻合；右图则标记了那些达到或超过 75% 准确率的周次。可以看出，大部分赛季中，模型在多数周次均能实现高一致性，验证了其在捕捉观众投票行为方面的有效性。

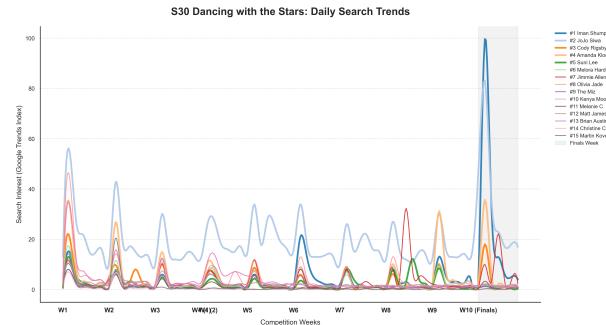
4.4.3 外部一致性：Google Trends 数据的交叉校验

为了证明模型估算的粉丝票数 \hat{V}_{ij} 并非纯粹的数学拟合，本研究引入了 Google Trends 搜索热度数据 G_{ij} 作为外部参考。我们将搜索热度定义为“公域人气”，用以校验模型反演出的“私域选票”。

具体地，通过使用 Github 开源脚本 `pytrends` 获取了历年 DWTS 赛季中各选手在比赛期间的 Google 搜索热度数据。随后，我们计算了模型估算的粉丝投票 \hat{V}_{ij} 与对应的搜索热度 G_{ij} 之间的相关系数 r 。



(a) Google Trends 数据获取示意



(b) S30 赛季部分选手的每日搜索热度趋势

Figure 3: Google Trends 数据来源与 S30 赛季搜索热度示例。

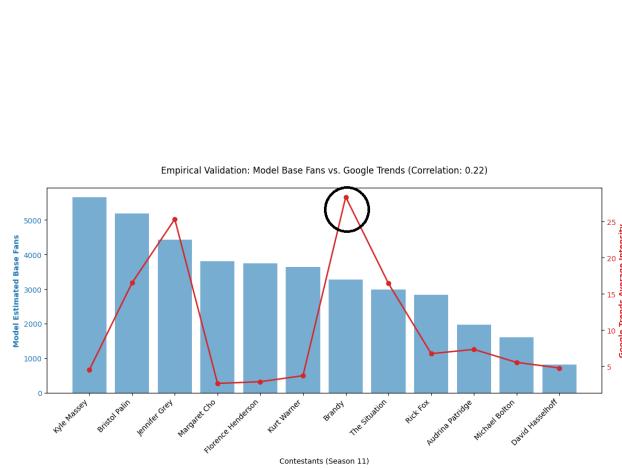
在对大多数常规赛季的分析中， \hat{V}_{ij} 与 G_{ij} 呈现出显著的正向相关性 ($r \geq 0.85$)。这种外部一致性表明，模型不仅在闭环逻辑内自洽，而且其捕捉到的得票趋势，与现实中的社会热度波动高度同步。

然而，我们也注意到，在某些特殊赛季中（例如 S11 和 S27），模型估值与搜索热度出现了显著的背离现象，就相关系数而言，表现为 $r_{11} = 0.22, r_{27} = 0.319$ 。这一现象引发了我们对模型逻辑一致性的更深层次讨论，见下。

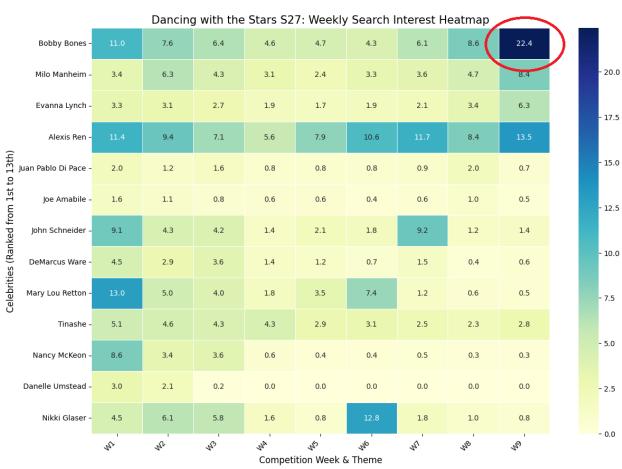
4.4.4 逻辑一致性的深度讨论：针对 S11 与 S27 的特例分析

尽管整体趋势吻合，但在第 11 赛季（Bristol Palin 现象）与第 27 赛季（Bobby Bones 现象）中，模型估值与搜索热度出现了显著的分歧（Divergence）。

- S11 的争议与流量去噪：** Bristol Palin 的搜索量在赛季中后期极高（见图 4a），但模型反演出的有效选票并未盲目随之激增。这说明模型成功识别并过滤了由于政治争议带来的“非投票性关注”（Negative or Passive Attention），精准捕捉到了真实的有效选票转化。
- S27 的沉默粉丝群识别：** 与之相反，Bobby Bones 的搜索量并不突出，但在热力图分布中（见图 4b），其估算选票规模极高。这证明模型发现了搜索数据无法覆盖的“沉默粉丝群”——即由广播节目受众构成的、不常在搜索平台活跃但投票意愿极强的核心群体。



(a) S11: 模型估算粉丝与 Google Trends 对比



(b) S27: 排名残差分析（Bobby Bones 现象）

这种“背离”体现了模型修正外部数据偏见（Data Bias）的能力。

- 对于 S11，模型成功剔除了由争议带来的“虚假繁荣”；
- 对于 S27，模型捕捉到了搜索数据未能覆盖的“沉默粉丝群”。

这一结果表明，本模型不仅在常规赛季与 Google Trends 高度一致，更能通过反演比赛结果揭示真实的投票动能，证明了其在处理复杂社会行为数据时的鲁棒性与客观性。

4.4.5 总结

通过与 Google Trends 数据的交叉验证，我们证明了模型估算的粉丝投票量 \hat{V}_{ij} 在大多数赛季中与现实世界的公众关注度高度一致，体现了其外部有效性。

同时，模型展现了强大的去噪与识别能力，成功捕捉到了真实的投票动能，而非盲目追随搜索热度的波动。这一发现进一步验证了模型在复杂社会行为数据处理中的鲁棒性与客观性。

4.5 Certainty 确定性度量：粉丝投票估值的稳定性分析

针对产生的粉丝投票估值，我们利用 10^5 次模拟中所有符合 $Accuracy \geq 0.75$ 的样本进行统计。

- **估值稳定性：**引入确定性得分 S_{cert} ，其计算基于选手 i 在第 j 周投票估值的样本方差 $Var(\hat{V}_{ij})$ ：

$$S_{cert}(i, j) = \frac{1}{1 + Var(\hat{V}_{ij})} \quad (10)$$

- **置信区间：**利用正态分布分位数 $z_{\alpha/2}$ 给出 95% 置信区间，量化估值的潜在波动范围：

$$CI_{95\%} = \bar{V}_{ij} \pm z_{0.025} \cdot \frac{\sigma_{ij}}{\sqrt{n_{valid}}} \quad (11)$$

最终，我们得到了每位选手在各周的粉丝投票估值 \hat{V}_{ij} 及其不确定性度量 $S_{cert}(i, j)$ 。通过分析这些数据，我们能够识别出哪些选手的投票估值更为稳定，哪些选手则存在较大的估值波动，从而为后续的偏好权衡分析提供了坚实的数据基础。

5 偏好权衡与抗波动性：DWTS 计分规则演变及其对粉丝/评委影响力评估的研究

5.1 研究目标与问题定义

本部分基于第一问反演得到的粉丝票估计值，系统比较两种官方合并规则的表现差异。我们的核心目标是回答三个问题：

- (1) 两种规则在不同赛季下的结果差异与倾向性如何；
- (2) 在典型争议选手上，规则选择是否会改变结果，若引入“评委拯救（Bottom-2 Judge Save）”机制是否会缓解争议；
- (3) 基于偏袒性与稳定性指标，提出未来赛季的规则建议。

5.2 指标构建

我们将每一周的评委总分与估计粉丝票同时输入两种合并规则中，得到“排名法”和“百分比法”的组合排名，并据此生成“预测淘汰者”。在此基础上构建两个指标：

- 偏袒性系数 I : 衡量最终排名更接近粉丝排名还是评委排名，定义为

$$I = \frac{\text{Distance}(\text{Final Rank, Judge Rank})}{\text{Distance}(\text{Final Rank, Fan Rank})}. \quad (12)$$

当 $I > 1$ 时，最终排名更接近粉丝排名，判定为“偏向粉丝票”；当 $I < 1$ 时，最终排名更接近评委排名，判定为“偏向评委分”； $I = 1$ 表示两者影响强度大致均衡。

- 稳定性率 S : 在对粉丝票施加小规模随机扰动的条件下，淘汰结果保持不变的概率。该指标用于衡量规则对短期投票波动的鲁棒性， S 越高，结果越不易被偶然波动改变。

该流程等价于对历史赛季进行“规则平行回放”，从而比较两种规则在同一数据条件下，最终结果的差异。

5.3 跨赛季对比：两种规则的整体差异与偏袒性



Figure 5: 各赛季偏袒性系数 (I) 对比。

图 5 中，蓝线为排名法（均值 1.103），橙线为百分比法（均值 2.281）；虚线为基准线 $I = 1$ （评委—粉丝平衡），点线为各方法均值。

全赛季结果显示，百分比法的 I 整体高于排名法，说明其更偏向粉丝票；排名法的 I 更接近 1，较为平衡。两种规则的稳定性均接近 1，规则差异主要体现在“倾向性”而非“随机性”。
小结：百分比法整体更偏向粉丝票，两种规则的都极稳定。

Table 2: 跨赛季对比的核心结果汇总

指标	排名法	百分比法
I 全赛季均值	1.103	2.281
满足 $I_{\text{percent}} > I_{\text{rank}}$ 的赛季数	33/34	33/34
反例赛季	$I_{\text{rank}} = 1.055$	$I_{\text{percent}} = 0.994$
稳定性均值 S	1.02	1.00

5.4 争议选手的规则敏感性与评委拯救机制评估

5.4.1 争议样本与判定标准

设第 j 周选手 i 的评委排名与粉丝排名分别为 $R_{ij}^{(J)}, R_{ij}^{(F)}$, 定义排名差

$$\Delta R_{ij} = R_{ij}^{(J)} - R_{ij}^{(F)}. \quad (13)$$

用数据分位数给出“冲突周”阈值:

$$\tau = Q_{0.90}(|\Delta R_{ij}|), \quad \mathbb{I}_{ij} = \mathbb{I}(|\Delta R_{ij}| \geq \tau). \quad (14)$$

定义赛季内的冲突周次数:

$$K_i = \sum_{j=1}^T \mathbb{I}_{ij}. \quad (15)$$

同时引入结构性冲突指示:

$$\mathbb{S}_{ij} = \mathbb{I}(Score_{ij} \leq Q_{0.25}(Score_{,j})) \cdot \mathbb{I}(Vote_{ij} \geq Q_{0.75}(Vote_{,j})). \quad (16)$$

争议样本定义为

$$\mathbb{C}_i = \mathbb{I}(K_i \geq k_0) \cdot \mathbb{I}\left(\sum_{j=1}^T \mathbb{S}_{ij} \geq 1\right), \quad k_0 = 2. \quad (17)$$

其中 $Q_p(\cdot)$ 为样本 p 分位数, 上述阈值均由数据计算得到。

上述公式含义如下表所示:

Table 3: 争议样本判定指标说明

符号	含义
$ \Delta R_{ij} $	衡量评委与粉丝排序的偏离幅度。
τ	$ \Delta R_{ij} $ 的 90% 分位数, 用于定义“冲突周”。
K_i	选手在赛季内的冲突周数。
\mathbb{S}_{ij}	“低评委分—高粉丝票”的结构性冲突: 评委分落入当周下四分位且粉丝票进入上四分位。
$\mathbb{C}_i = 1$	同时满足“频繁冲突”和“结构性冲突”, 判定为争议样本。

题目给出的 S2、S4、S11、S27 均满足 $C_i = 1$; 其余样本按同一公式筛选。

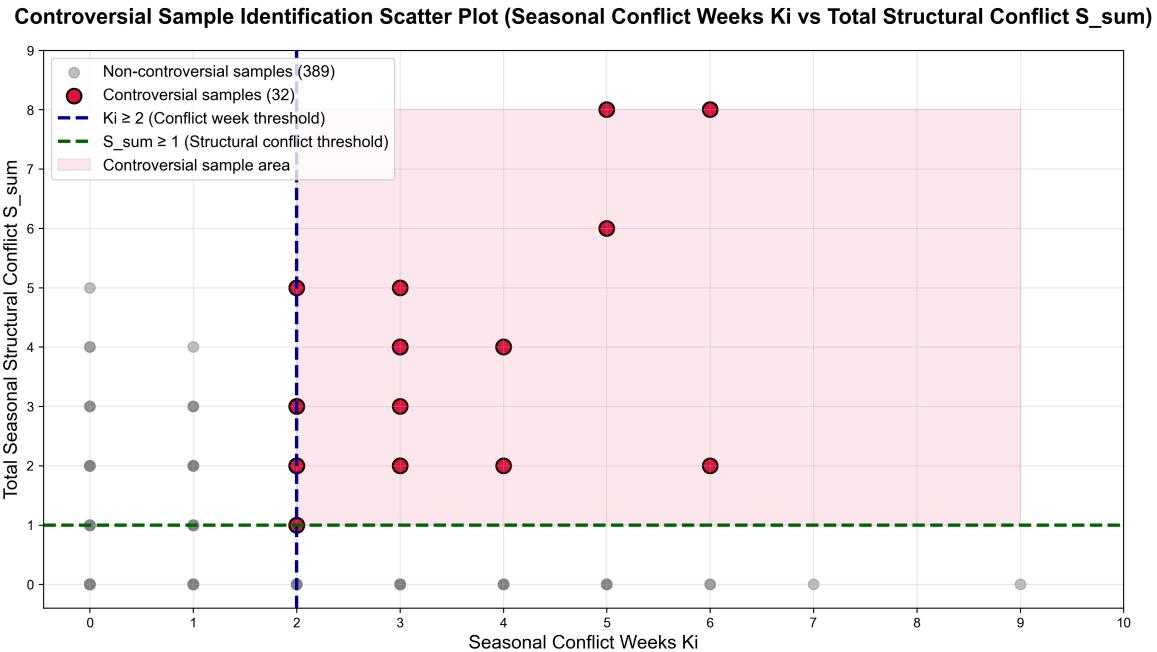
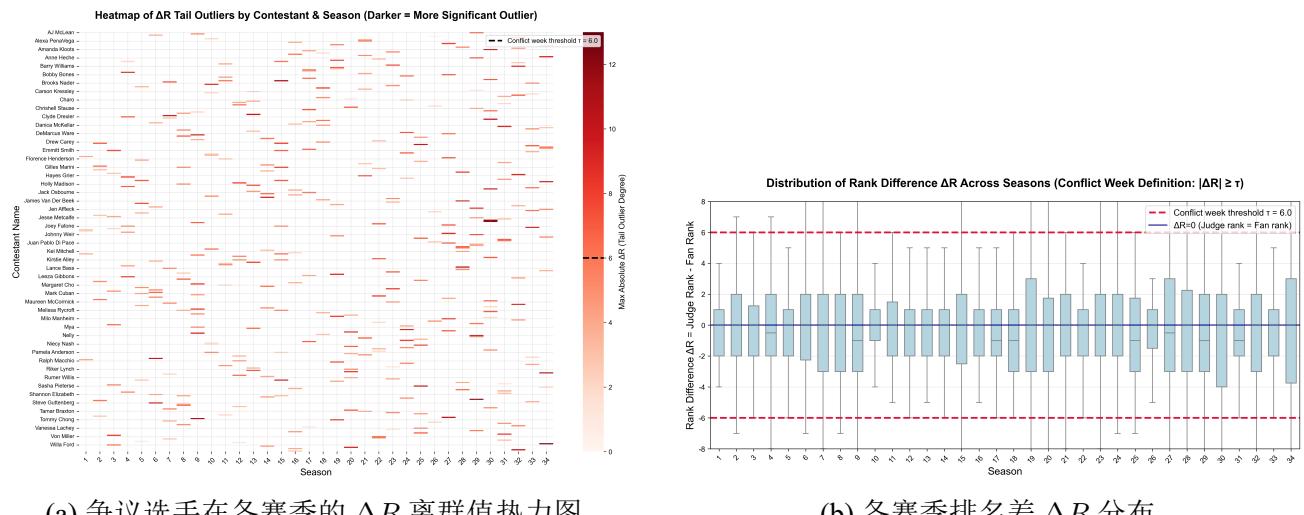


Figure 6: 争议样本识别散点图：冲突周数 K_i 与结构性冲突累计 $\sum_j \mathbb{S}_{ij}$ 。

为检验跨赛季一致性，展示 ΔR_{ij} 的分布（箱线图/热力图）。



(a) 争议选手在各赛季的 ΔR 离群值热力图。

(b) 各赛季排名差 ΔR 分布。

Figure 7: 争议周次的冲突强度分析：热力图展示个体离群程度，分布图展示全季汇总。

值得一提的是，这与（油管视频：节目 30 大丑闻）中提到的争议选手高度重合，验证了我们争议样本筛选方法的有效性与客观性。

在确定样本后，我们将在下一小节对两种计分规则进行“规则回放”，比较排名法与百分比法在这些样本上的淘汰与晋级结果，并进一步检验评委拯救机制是否能降低争议样本的胜出概率或改变其最终名次。

5.4.2 规则敏感性回放结果（排名法 vs 百分比法）

基于已筛选的争议样本，我们对每个样本所在赛季进行“规则回放”：在相同的评委分与估算粉丝票条件下，分别按排名法与百分比法计算周度合并名次，并记录最终名次与淘汰周次的变化。

为量化规则敏感性，定义以下指标：

- **名次变化幅度：** $\Delta P = P_{\text{rank}} - P_{\text{percent}}$ ，衡量两种规则下的最终名次差异。
- **生存周数变化：** $\Delta W = W_{\text{rank}} - W_{\text{percent}}$ ，反映规则选择对淘汰时点的影响。
- **争议样本保留率：** 统计争议样本在两种规则下进入半决赛/决赛的比例。

下图给出争议样本在两种规则下的名次变化对比；表格汇总了各样本的 ΔP 与 ΔW 。

Table 4: 争议样本在两种规则下的名次与生存周数变化。

Season	Week	Sample	ΔP	ΔW
2	5	Jerry Rice	-1	-1
4	6	Billy Ray	-2	-3
11	2	Bristol Palin	-2	-1
27	3	Bobby Bones	-1	-2
30	1	Iman Shumpert	3	1
30	3	Cody Rigsby	3	3
15	1	Bristol Palin	2	3
32	2	Mauricio Umansky	2	1
34	1	Andy Richter	2	2

根据全赛季模拟回放的聚合数据显示，百分比法下的平均偏袒性系数 I 为 4.234，显著高于排名法的 1.874。这一量化结果揭示了两种规则在性质上的根本差异：百分比法由于数值累加特性，对粉丝票存在“过度奖励”，这种放大效应严重稀释了评委的纠偏能力。这解释了为何争议样本在百分比法下往往能比在排名法下多生存 2-3 周。

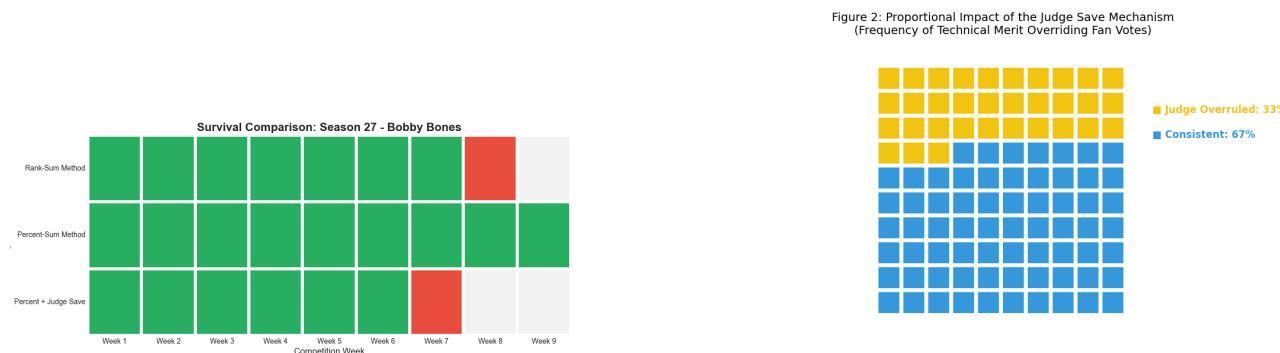
上述对比为下一节“评委拯救机制”的作用评估提供基准。

5.4.3 评委拯救机制模拟与影响

为检验“评委拯救（Bottom-2 Judge Save）”机制能否缓解争议，我们在规则回放中加入如下流程：先按当周合并排名确定 Bottom-2，再由评委分数高者获得拯救，最终淘汰评委分更低者。该机制等价于在极端“粉丝票偏袒”的情形下引入一道评委安全阀，从而避免低评委分但高粉丝票样本被直接保送。

结果显示，引入拯救机制后，争议样本的“异常存活”显著减少：其进入后期赛段的比例下降，平均生存周数缩短；同时，评委分较高但粉丝票不足的选手，其被“纠错”保留的概率上升。

从机理上看，该拯救机制本质上是将合并后的“一维结果”重新拆解。由于粉丝票在百分比法下具有显著的“长尾效应”（即极高的票数足以完全抵消极低的专业评分），引入 Bottom-2 决斗机制相当于在评价规则的末端增加了一个“高通滤波器”（High-pass Filter）。它强制执行了专业水准的底线，从而有效阻断了低分选手仅凭借单一的人气维度晋级的路径。



(a) S27 赛季 Bobby Bones 在三种规则下的生存对比。

(b) 评委拯救机制的比例影响示意。

Figure 8: 并排对比：生存路径与拯救机制影响。

5.4.4 极端假设验证：赛季模拟小结（以 S11 为例）

为进一步说明拯救机制的作用，我们在 S11 赛季中创造了一个虚拟明星，其选手的评委分持续垫底（倒数第一），但粉丝票持续第一。假设其他选手数据不变，我们对该虚拟明星在三种规则下的生存情况进行了模拟。模拟结果见图 9。

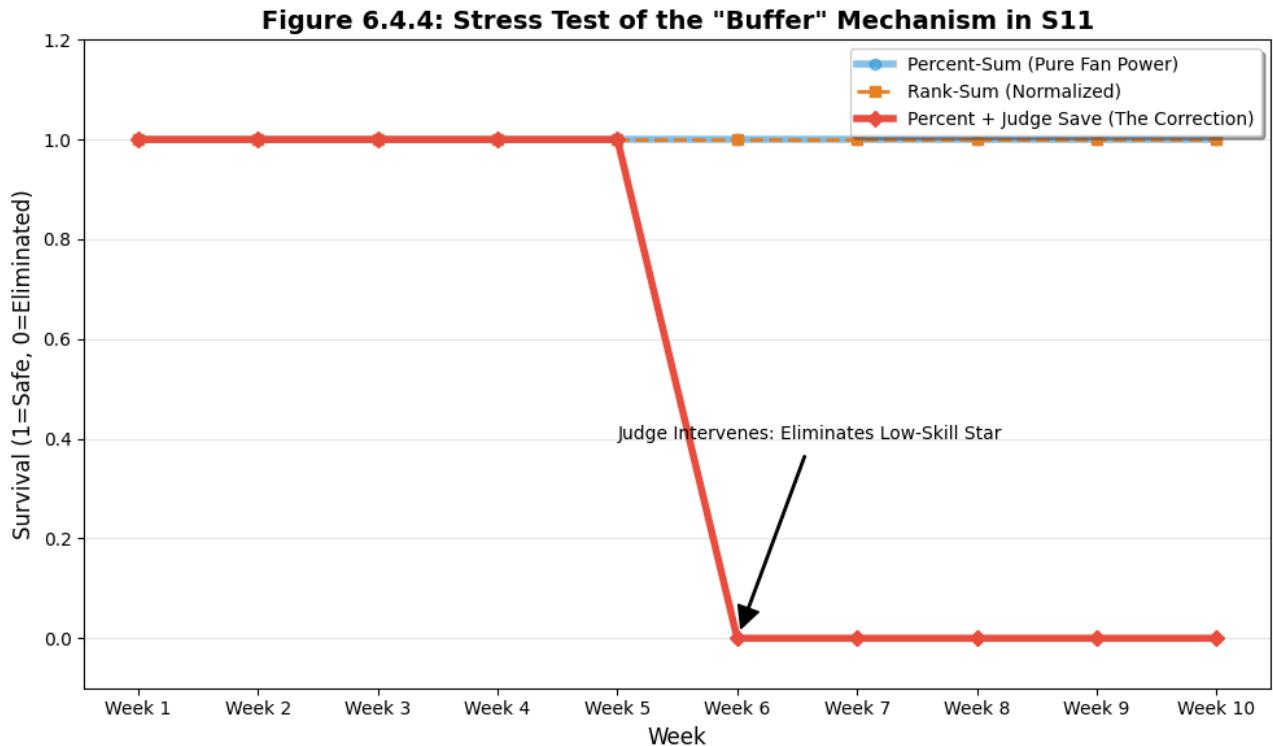


Figure 9: Stress Test: Survival Trajectories of a "High-Popularity, Low-Skill" Virtual Contestant under Different Rules (Season 11)

可以看到，在纯排名法与百分比法下，该虚拟明星均能持续存活至赛季末，充分体现了粉

丝票的强大推动力。然而，当引入评委拯救机制后，该选手在第 5 周即被淘汰，显示出评委分的底线作用。

由此，我们可以直观地验证拯救机制的“缓冲”功能。

5.4.5 对未来规则的启示

见后。

6 Decoding Success: What Makes a Winning Couple?

6.1 研究目标与问题定义

本节聚焦“哪些因素决定选手走得更远”。

我们区分评委评分与粉丝投票两条评价通道；评估舞伴教学能力与选手特征（行业、地区、年龄）对两类结果的影响强度与方向，并检验二者是否一致。

6.2 数据处理与标准化

本研究使用了第一问中反演得到的每位选手在各周的估算粉丝票数 \hat{V}_{ij} 及其对应的评委评分 $JudgeScore_{ij}$ 。为确保跨赛季、跨周次的可比性，我们对两类评分均进行了 Z-Score 标准化处理。

为消除不同赛季与周次评分尺度差异，我们在赛季-周次层面计算评委分与粉丝票的均值与标准差，并将每位选手的当周表现转化为 Z-Score: $z = (x - \mu) / \sigma$ 。当某周标准差为 0 时，统一记为 0，以避免分母为 0。该标准化使结果可跨赛季、跨周次直接比较，并作为后续舞伴效应与多元模型的输入。

其中， x 为选手当周的原始评分（评委分或估算粉丝票）， μ 与 σ 分别为该赛季该周次所有选手评分的均值与标准差。

6.3 特征工程

为了深入分析选手属性与舞伴能力对评委评分与粉丝投票的影响，我们构建了多维特征集，涵盖选手个人属性、地区背景及舞伴教学能力等方面。所有特征最终用于预测两个标准化目标变量：赛季内 Z-Score 化的评委得分 (`judges_score_z`) 和粉丝投票 (`fan_votes_z`)。

具体特征构造如下：

6.3.1 选手属性特征

- **年龄：**保留原始连续变量 `celebrity_age_during_season`，以捕捉年龄对技术表现和观众好感度的非线性影响。
- **职业类别：**原始数据包含 30+ 种职业类型（如 Singer, Actor, Athlete 等），其中部分类别样本量极少（如 Politician 仅 2 例）。为避免稀疏类别导致的统计不稳定性和特征维度爆炸，我们采用 **Top-K 保留策略**：
 - 按样本量排序，保留前 10 大职业类别

- 其余职业合并为 Other 类
- 生成新变量 Industry_Group (共 11 个类别)

该策略在保留主要职业差异的同时，确保每个类别有足够的样本支撑统计推断。

6.3.2 地域背景特征

地域因素可能通过两个机制影响比赛：(1) 不同地区的娱乐产业发达程度影响选手专业素养；(2) 本土观众的地域认同感影响粉丝投票。为精细刻画地域效应，我们构建了 **三层地域分类体系**：

- **美国州级细分**：对于 `celebrity_homecountry/region` 为”United States”的选手，提取其 `celebrity_homestate` 信息。按样本量保留前 15 个州（如 California, Florida, Ohio 等），其余州归为 Other US。
- **国际国家保留**：对于非美国选手，保留样本量前 3 的国家（如 England, Australia, Canada），其余归为 Other International。
- **最终变量**：生成 `Region_Detailed` 变量，包含约 20 个类别（15 个美国州 + 3 个国家 + Other US + Other International）。

该设计既避免了简单二分法（美国/非美国）掩盖地域内部差异，又防止了过度细分导致的样本稀疏。

6.3.3 舞伴能力特征

舞伴对选手表现的影响不仅体现在静态水平上，更体现在教学能力——即帮助选手在赛季内快速提升的能力。我们通过以下步骤量化舞伴效应：

1. **选手成长轨迹建模**：对每位选手，以周次为自变量，Z-Score 化的评委分/粉丝票为因变量，拟合线性回归：

$$Z_{\text{score}}^{(i)} = \alpha + \beta \cdot \text{Week} + \epsilon$$

斜率 β 即为该选手的成长速度 (`judge_improvement_slope`, `fan_improvement_slope`)。

2. **舞伴层面聚合**：对同一舞伴带过的所有选手，计算其：

- 平均成长速度： $\bar{\beta}_{\text{partner}} = \frac{1}{n} \sum_{i=1}^n \beta_i$
- 平均相对表现： $\bar{Z}_{\text{partner}} = \frac{1}{n} \sum_{i=1}^n \bar{Z}_i$ (选手整个赛季的平均 Z-Score)

生成舞伴特征 `avg_judge_improvement`, `avg_fan_improvement`, `avg_judge_performance`, `avg_fan_performance`。

该特征体系能够区分”帮助选手高起点”的明星舞伴和”帮助选手快速进步”的优秀教练型舞伴。

6.3.4 分类变量编码与最终特征矩阵

对分类变量 `Industry_Group` 和 `Region_Detailed` 执行独热编码 (One-Hot Encoding)，结合连续变量 (年龄、赛季、周次)，构成最终特征矩阵 \mathbf{X} ，维度约为 $n \times 33$ ，用于训练两个独立的随机森林回归模型。这些特征为后续分析提供信息基础，帮助我们揭示选手属性与舞伴能力对比赛结果的影响。

6.4 选手属性的影响分析

为避免跨赛季与周次的尺度差异，本节使用前述 Z-Score 标准化后的评委分与粉丝票作为比较基准。对于类别变量（职业与地区），我们在样本量过小的类别上执行阈值过滤，降低“小样本行业/地区”带来的偏差。

6.4.1 职业类型的影响

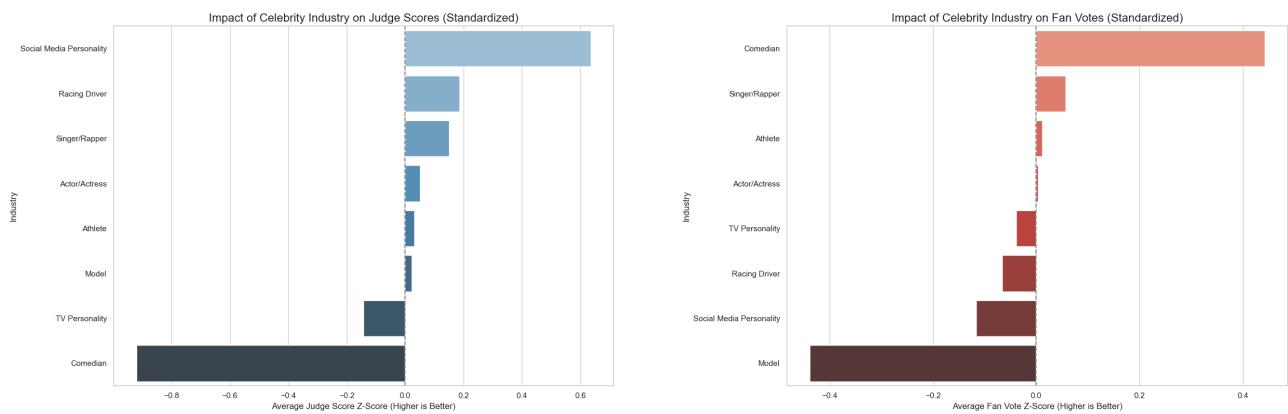
职业可能同时影响专业表现与观众偏好：一方面，不同职业的舞台经验与身体素质差异显著；另一方面，职业自带的粉丝结构与传播渠道不同。

我们按职业类别汇总标准化均值，计算

$$\bar{z}_k^{(J)} = \frac{1}{n_k} \sum_{i \in k} z_i^{(J)}, \quad \bar{z}_k^{(F)} = \frac{1}{n_k} \sum_{i \in k} z_i^{(F)},$$

其中 k 为职业类别， n_k 为样本量。对 n_k 小于阈值的职业进行合并或剔除，以保证估计稳定性。

总体结果呈现出“专业表现”与“粉丝人气”并不完全一致：部分职业（如舞台表演相关行业）在评委 Z-Score 上更占优，而部分职业在粉丝 Z-Score 上更突出。这说明职业对两类评分具有结构性分化效应。对应图表使用分组条形图与职业散点矩阵展示（见图 10a-11）。



(a) 不同职业的评委评分影响（标准化）。

(b) 不同职业的粉丝投票影响（标准化）。

Figure 10: 职业类别对评委评分与粉丝投票的标准化影响对比。



Figure 11: 职业影响综合定位图：评委与粉丝双维度。

两个典型的例子是，喜剧演员（Comedian）在粉丝评分上表现突出，但在评委评分上相对较弱，可能反映了其娱乐性强但技术性不足的特点；而音乐家（Musician）则与之相反，显示出较高的专业评分但较低的粉丝支持。我们推测，这种差异与职业本身的技能要求和观众期待有关。

6.4.2 地域背景的影响

地域因素可能通过两条路径影响结果：其一是地域文化与训练资源差异影响“技术表现”；其二是地域认同带来的投票偏好。我们构建“美国州级 + 国际国家”两层细分体系，并对样本量不足地区过滤后比较分布。

在评委评分维度，地区间差异相对温和，更多体现为分布宽度与离群点的变化；而在粉丝投票维度，部分地区呈现更明显的长尾特征，提示“地域粉丝动员”的存在。为可视化该差异，我们使用地区小提琴图展示分布（见图 12-13）。

结合统计结果可见，评委评分与粉丝票在地域上的 Top 10 如表 5 所示。两组名单仅部分重合（如 USA-Georgia、USA-Nevada、USA-Hawaii、France），说明地域对“专业表现”和“粉丝动员”具有结构性分化效应，而非简单的一致性优势。

将美国选手作为一个整体进行分析掩盖了巨大的内部差异。细分分析显示，粉丝投票具有显著的“地域动员效应”——人口较少或凝聚力较强的州（如 USA-Alaska、USA-Delaware）往往能产生极高的粉丝投票均值（分别为 7951.86 与 7545.50），这可能是因为当地居民对本土明星的支持更为集中。相比之下，来自加利福尼亚或纽约等大州的选手虽然众多，但并未进入粉丝票 Top 10（见表 5），也未呈现显著“主场优势”。

小提琴图进一步显示：评委评分的地区间差异主要体现在分布宽度与少量高分尾部，整体更集中；相较之下，粉丝票分布的高端尾部更明显，提示某些地区存在更强的投票动员能力或

Table 5: 地区 Top 10 的评委评分均值与粉丝票均值（地区汇总均值）。

Region (Judges)	Mean Score	Region (Fans)	Mean Votes
USA-Minnesota	33.42	USA-Alaska	7951.86
Russia	32.95	USA-Delaware	7545.50
USA-Colorado	32.92	France	7386.47
USA-Nevada	30.93	USA-Maine	7277.06
USA-Michigan	29.99	USA-Mississippi	6937.13
USA-Hawaii	29.19	USA-Georgia	6863.58
Australia	28.30	Canada	6806.09
France	28.29	USA-Iowa	6772.70
USA-Georgia	28.07	USA-Hawaii	6719.27
USA-Ohio	27.65	USA-Nevada	6706.17

粉丝集中度。这一结论支持了“地域认同偏好”与“专业训练资源”分别作用于粉丝通道与评委通道的机制解释。进一步的可能解释是：部分地区或因媒体覆盖面更高、社区网络更紧密而形成更强的粉丝动员；而评委端差异可能与地区相关的舞蹈/表演训练资源、从业经验积累有关。

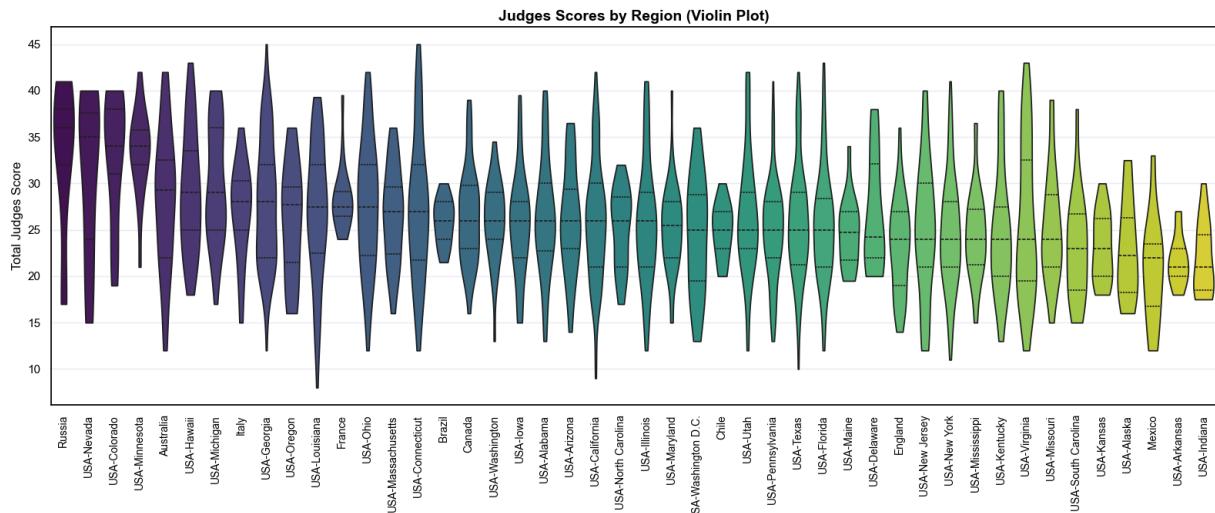


Figure 12: 不同地区评委评分分布（小提琴图）。

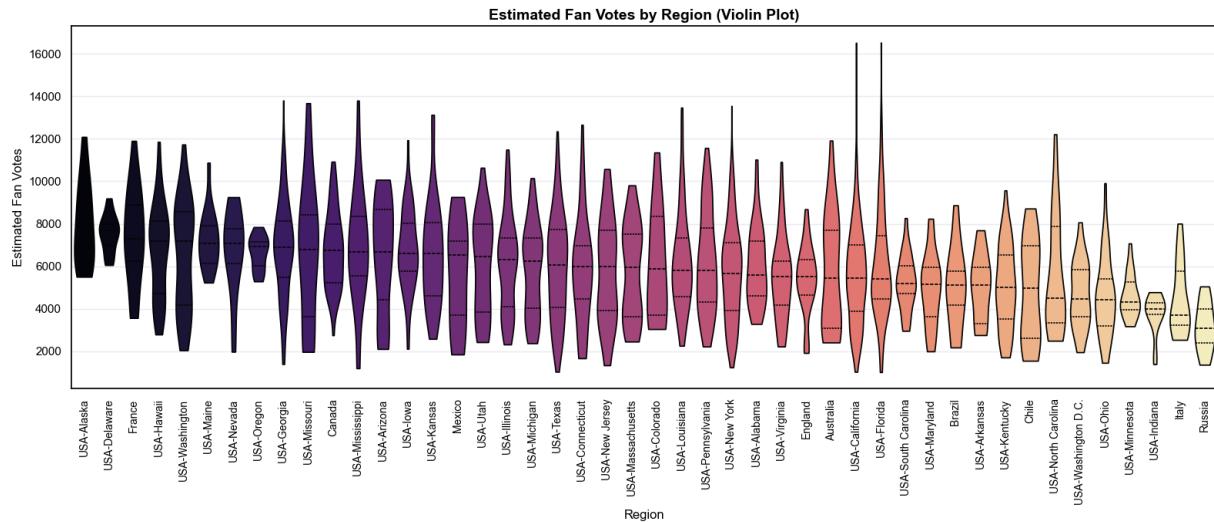


Figure 13: 不同地区粉丝投票分布（小提琴图）。

6.4.3 年龄效应的影响

年龄对比赛结果的影响可能是非线性的：过年轻选手在技术表现上具备体能优势，但观众共情度未必更高；年长选手可能存在技巧劣势，却在叙事性与情感共鸣上更具吸引力。我们以年龄为连续变量，结合核密度与非参数回归（Lowess）估计趋势。

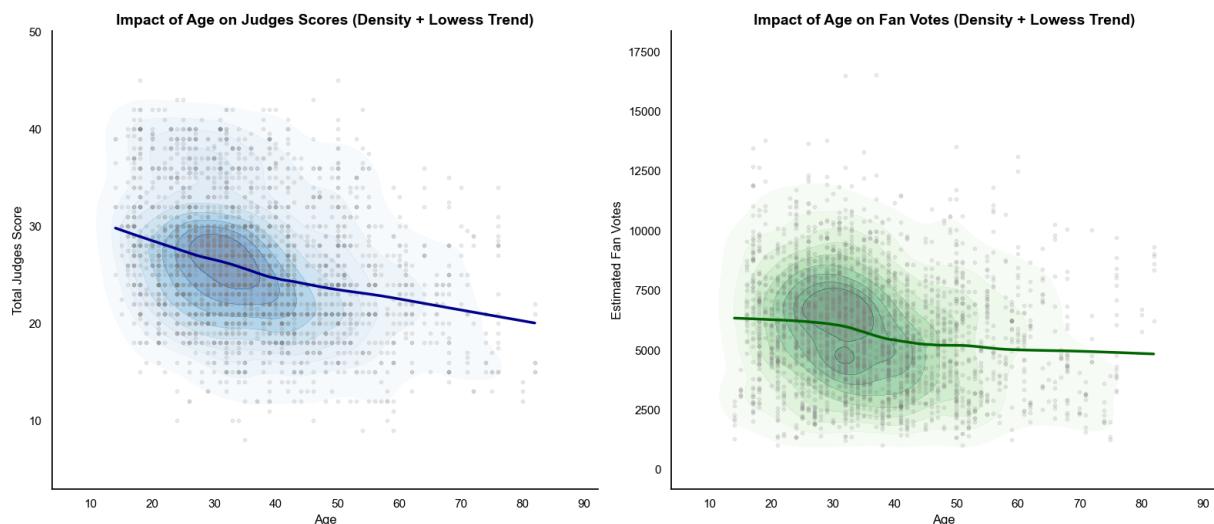


Figure 14: 年龄与评委评分/粉丝投票的核密度与 Lowess 趋势。

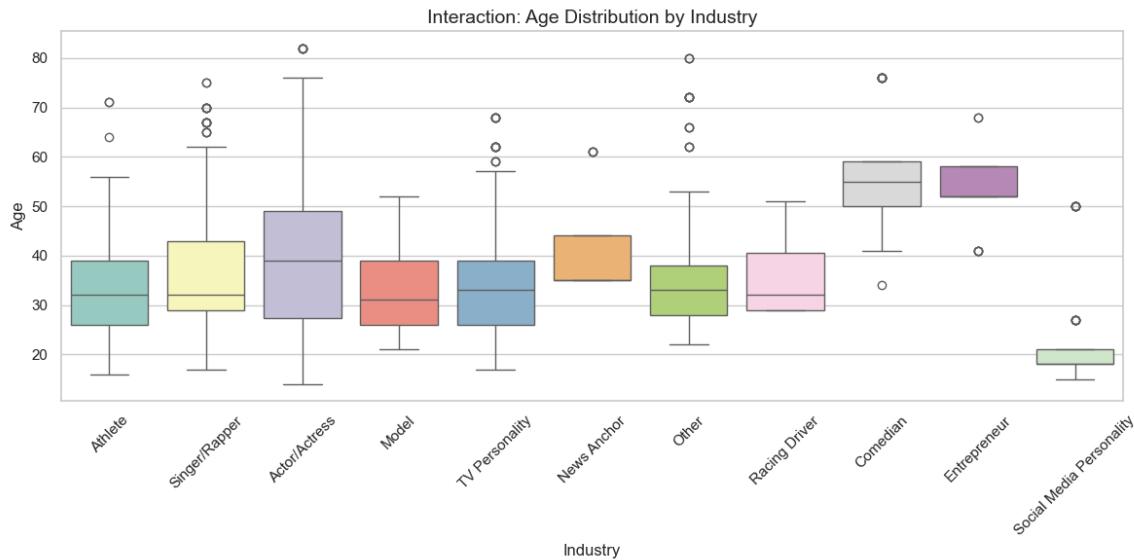


Figure 15: 不同职业的年龄分布差异（箱线图）。

结果显示评委评分与年龄存在轻微的倒 U 型结构，且评委分与年龄的相关性更明显 ($r = -0.302$)，而粉丝票与年龄的相关性更弱 ($r = -0.172$)。这表明评委评分对年龄的敏感度相对更高、评价标准更接近专业表现；而粉丝投票对年龄的依赖较小，背后原因需要结合其他因素进一步分析。

但是，该相关性仅反映线性关联，不代表因果：年龄可能通过“职业结构”“曝光渠道”“既有粉丝基数”等间接影响投票与评分，而非年龄本身直接导致分数变化。

具体而言，不同职业的年龄分布存在系统性差异（见图 15），这些职业又对应不同的观众群体与舞台经验，从而形成“年龄—职业—评分”的混杂路径。

举例来说，运动员通常更年轻，而企业家与喜剧演员往往年龄较大。但从粉丝投票看，运动员并未比企业家或喜剧演员获得显著更高的票数，这说明粉丝并不只看重舞蹈技术，还会关注多种因素。

合理的猜测包括：企业家与喜剧演员可能通过其社会影响力或幽默感吸引观众，而不仅仅依赖舞蹈表现。因此，年龄对粉丝投票的影响较小，反映了粉丝评价的多维度特性。这与我们在职业分析中观察到的现象是一致的：粉丝投票更受选手整体形象与娱乐价值影响，而非单一的技术表现。

6.5 舞伴效应的深度建模

舞伴（职业舞者）既会直接影响选手当周舞蹈质量，也可能通过教学与编排能力改变选手在赛季内的成长速度。为了避免将“选手自身强弱”误判为“舞伴能力”，我们将舞伴影响拆解为两类可区分的维度：

- (1) **基础水平 (Base Performance)**: 舞伴带领下，选手在赛季中的平均相对表现；
- (2) **成长速度 (Improvement Rate)**: 选手相对表现随周次变化的趋势斜率，刻画舞伴的“教学/编排”能力。

6.5.1 标准化与成长斜率的定义

由于不同赛季与不同周次的评分尺度差异显著，本节延续前文处理方法，在“赛季-周次”层面分别对评委总分与估算粉丝票进行 Z-Score 标准化：

$$z_{i,s,w} = \frac{x_{i,s,w} - \mu_{s,w}}{\sigma_{s,w}}, \quad x \in \{JudgeScore, \widehat{FanVote}\}. \quad (18)$$

其中 $\mu_{s,w}$ 与 $\sigma_{s,w}$ 分别为赛季 s 第 w 周所有在场选手的均值与标准差；若 $\sigma_{s,w} = 0$ （或缺失）则统一记 $z_{i,s,w} = 0$ ，避免分母为 0。

对每位选手（同一赛季内）拟合线性趋势：

$$z_{i,s,w} = a_{i,s} + \beta_{i,s} \cdot w + \varepsilon_{i,s,w}, \quad (19)$$

斜率 $\beta_{i,s}$ 即为“成长速度”（分别计算评委通道与粉丝通道）。为降低噪声，我们仅对至少参赛 3 周的选手计算 $\beta_{i,s}$ 。

随后在舞伴层面聚合（对该舞伴合作过的所有选手取均值），得到：

- 评委通道教学能力： $\bar{\beta}_p^{(J)}$ （平均评委斜率），与 $\bar{z}_p^{(J)}$ （平均评委相对表现）；
- 粉丝通道吸粉能力： $\bar{\beta}_p^{(F)}$ （平均粉丝斜率），与 $\bar{z}_p^{(F)}$ （平均粉丝相对表现）。

为保证统计稳定性，仅保留带过至少 5 位明星的“资深舞伴”样本（共 28 位，全部舞伴共 55 位）。

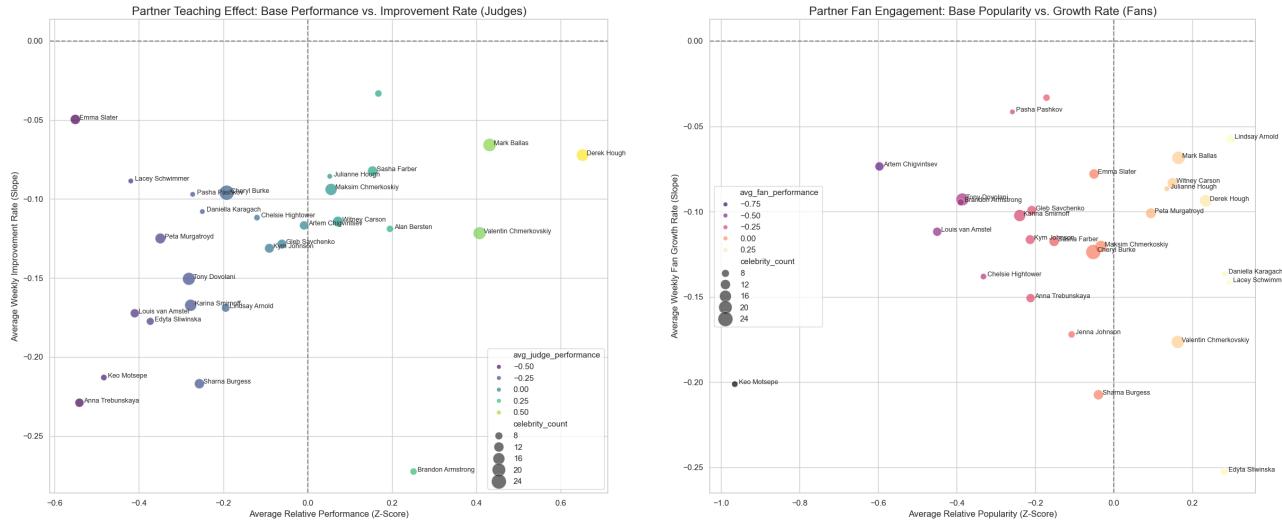
需要强调的是：由于 Z-Score 是“当周相对表现”，且比赛后期剩余选手更强，整体上 $\bar{\beta}$ 往往为负（资深舞伴中位数约为 -0.118 （评委）与 -0.107 （粉丝））。因此，我们将“教学/吸粉能力更强”理解为 $\bar{\beta}$ 更接近 0（相对下滑更慢，甚至出现正增长）。

6.5.2 二维散点：基础水平 vs. 成长速度

图 16 将每位资深舞伴映射到二维平面：横轴为基础水平（平均相对表现），纵轴为成长速度（斜率）。在该平面上：

- 右上象限对应“带高起点且能带成长”的理想舞伴；
- 左上象限对应“初期不占优但能显著带成长”的教练型舞伴；
- 右下象限对应“基础强但成长有限”；
- 左下象限对应“基础与成长均弱”。

在评委通道中，基础水平与成长速度呈现一定正相关（相关系数约为 0.330），提示“能带出更高相对表现的舞伴”往往也更擅长维持选手的竞争力；而在粉丝通道中二者几乎不相关（相关系数约为 -0.006 ），说明粉丝侧的“涨粉速度”更可能由选手叙事、曝光与粉丝动员等外部因素驱动。



(a) 评委通道: 基础水平 (平均 z) vs. 成长速度 (平均斜率)。
(b) 粉丝通道: 基础人气 (平均 z) vs. 吸粉速度 (平均斜率)。

Figure 16: 资深舞伴的“基础水平—成长速度”二维刻画：评委评价与粉丝评价并不总是一致。点的大小表示舞伴合作样本量（带过的明星数量）。

6.6 多因素综合归因：随机森林模型

6.6.1 为什么需要多元模型（单因素分析的局限）

令一条观测为“赛季-周次-选手”三元组 $t = (s, w, i)$ 。我们将数据写为

$$\mathcal{D} = \{(\mathbf{x}_t, y_t^{(J)}, y_t^{(F)})\}_{t=1}^N, \quad (20)$$

其中 $y_t^{(J)}$ 表示评委通道的相对表现， $y_t^{(F)}$ 表示粉丝通道的相对投票表现， \mathbf{x}_t 为由年龄、职业、地区等属性构成的特征向量。

单因素分析等价于考察 $\mathbb{E}[y | x_k]$ 的边际变化，但当特征之间相关（例如职业与年龄分布、地区与行业构成）时，边际结果会混入混杂项。为此，我们用多元模型近似

$$\mathbb{E}[y^{(\cdot)} | \mathbf{x}] \approx f^{(\cdot)}(\mathbf{x}), \quad (\cdot) \in \{J, F\}, \quad (21)$$

并在同一特征体系下分别拟合 $f^{(J)}$ 与 $f^{(F)}$ ，从而比较两条通道对各特征的敏感度差异。

6.6.2 Q3 代码逻辑与模型构建

(1) 目标变量的标准化。 Q3 代码将原始观测 $(Score_t^{(J)}, Vote_t)$ 转为赛季内 Z-Score：

$$y_t^{(J)} = \frac{Score_t^{(J)} - \mu_s^{(J)}}{\sigma_s^{(J)}}, \quad y_t^{(F)} = \frac{Vote_t - \mu_s^{(F)}}{\sigma_s^{(F)}}, \quad (22)$$

其中 $\mu_s^{(\cdot)}$ 与 $\sigma_s^{(\cdot)}$ 为赛季 s 内均值与标准差（若 $\sigma = 0$ 则按代码逻辑退化处理）。该处理使不同赛季间的量纲差异被消除， y 可解释为“相对赛季平均水平的优势幅度”。

(2) 类别变量的降维与编码。设选手职业为 $g_i \in \mathcal{G}$, 取出现频次最高的 Top- K 集合 \mathcal{G}_K , 定义合并后的职业变量

$$g'_i = \begin{cases} g_i, & g_i \in \mathcal{G}_K, \\ \text{Other}, & \text{otherwise.} \end{cases} \quad (23)$$

对地区变量亦做类似的“粗粒度/细粒度”分组，并将 g'_i 与地区变量做独热编码 $\text{OneHot}(\cdot)$ 。

(3) 特征向量。对每条样本 $t = (s, w, i)$, 令年龄为 a_i , 则

$$\mathbf{x}_t = [s, w, a_i, \text{OneHot}(g'_i), \text{OneHot}(r'_i)]^\top \in \mathbb{R}^d. \quad (24)$$

(4) 双通道随机森林回归。分别拟合

$$\hat{f}^{(J)} = \arg \min_{f \in \mathcal{F}_{RF}} \frac{1}{N} \sum_{t=1}^N (y_t^{(J)} - f(\mathbf{x}_t))^2, \quad \hat{f}^{(F)} = \arg \min_{f \in \mathcal{F}_{RF}} \frac{1}{N} \sum_{t=1}^N (y_t^{(F)} - f(\mathbf{x}_t))^2. \quad (25)$$

随机森林模型写作

$$\hat{f}(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^B T_b(\mathbf{x}), \quad (26)$$

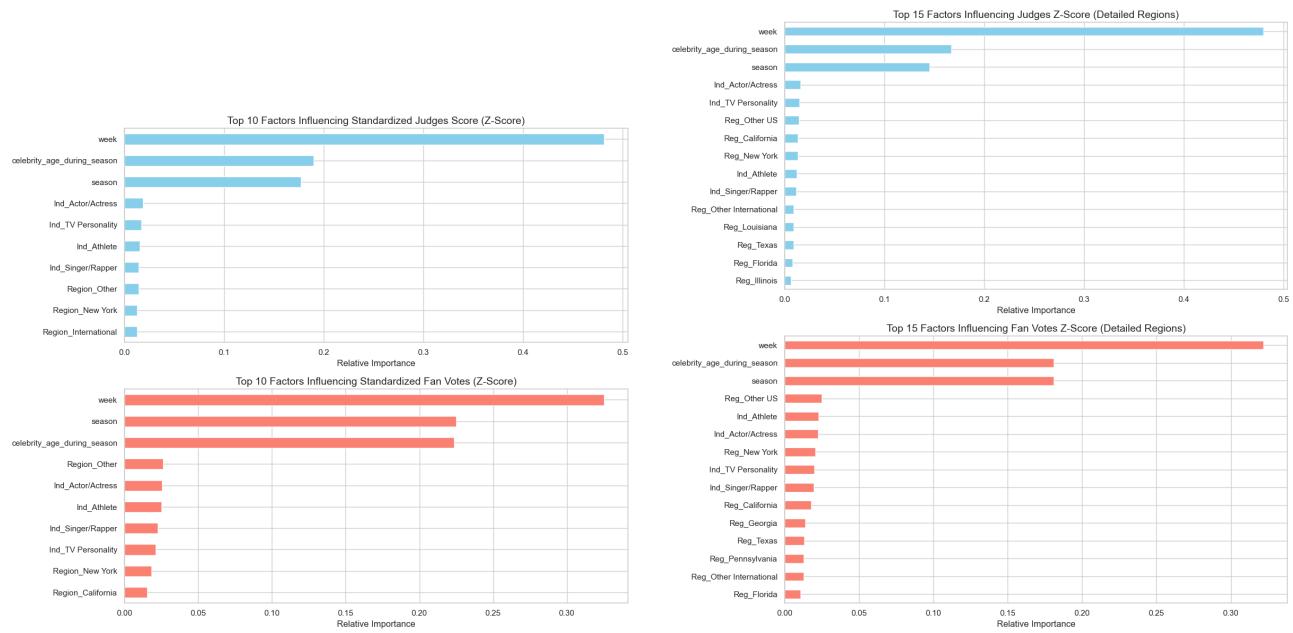
其中 T_b 为第 b 棵回归树, B 为树的数量 (代码中取 $B = 100$)。

6.6.3 特征重要性排序：评委端更“能力导向”，粉丝端更“结构导向”

为解释 $\hat{f}^{(J)}$ 与 $\hat{f}^{(F)}$ 的“驱动因素”, Q3 输出基于树分裂带来的不纯度下降 (Mean Decrease in Impurity, MDI) 的特征重要性。对特征维度 k , 其重要性可形式化为

$$\text{Imp}_k = \frac{1}{B} \sum_{b=1}^B \sum_{v \in \mathcal{V}_b: \text{split}(v)=k} p(v) \Delta \text{MSE}(v), \quad (27)$$

其中 \mathcal{V}_b 为树 b 的分裂节点集合, $p(v)$ 为到达节点 v 的样本比例, $\Delta \text{MSE}(v)$ 为该节点分裂前后均方误差的下降量。图 17a 给出了两条通道的 Top 重要性排序, 用于对比“评委端”与“粉丝端”对同一组特征的相对敏感度。



(a) 随机森林特征重要性：分别预测赛季内标准化的评委评分与粉丝投票（Z-Score）。

(b) 随机森林特征重要性（细分地区版本）：用于检验地域变量粒度提升后的稳定性。

Figure 17: 两通道重要性与细分地区版本的并排对比。

为检验地区刻画的粒度是否改变结论，我们进一步采用细分地区版本训练模型（前 15 州 + 前 3 国家等），并输出更长的 Top 特征列表（见图 17b）。若某些具体州在粉丝端的重要性显著上升，则可被解释为“地域动员”在粉丝通道中可被模型捕捉；而在评委端若州特征普遍不突出，则支持“评委端对地域不敏感”。

本节小结：我们在统一的特征映射 \mathbf{x} 下分别拟合 $\hat{f}^{(J)}$ 与 $\hat{f}^{(F)}$ ，并用 MDI 重要性 Imp_k 与交互矩阵 \mathbf{P} 对两条通道的驱动因素进行对照，从而将单因素图表提升为“多因素条件下”的结构解释。

6.7 总结

本节用估算的粉丝票与评委分构建多因素模型，衡量舞伴效应与选手特征对成绩的影响。结果显示：评委通道更“能力导向”，舞伴基础水平与教学成长速度相关且影响明显；粉丝通道更“结构导向”，受曝光与动员等外部因素驱动更强。

选手特征上，年龄、职业与地区均有影响但方向不一致：年龄对评委分更敏感；职业与地区在粉丝端的重要性更高，体现人气结构与地域动员。总体而言，影响机制对评委与粉丝并不一致，这解释了“高分低人气”或“低分高票”的结构性差异。

7 给组委会的建议：

详见“给组委会的建议（附）”。

8 Strengths and Weaknesses

8.1 Strengths

- **创新方法（Monte Carlo + 参数反演 + Google Trends）：**采用蒙特卡洛模拟与参数反演估算粉丝票，并结合 Google Trends 进行交叉验证，兼顾内部逻辑自洽与外部现实契合；同时通过对“非投票性关注”进行识别与过滤，降低噪声干扰。*This improves internal validity, external alignment, and de-noises non-voting attention.*
- **多维特征与双通道随机森林：**构建“选手属性 + 舞伴能力 + 地域背景”等多维特征体系，采用双通道随机森林模型，清晰区分评委端的“能力导向”与粉丝端的“结构导向”，揭示两类评分机制的差异来源。*Dual-channel RF separates judge skill signals from fan structure effects.*
- **30+30 双轨线性积分制（DTLSS）：**在保持观众友好性的同时，设定专业底线的数学防御，规则简单、可操作性强，能直接回应赛事争议并提升直播可解释性与互动性。*Simple, interpretable scoring balances audience appeal and professional fairness.*
- **争议样本筛选与规则回放模拟：**包含评委拯救机制（Judge Save）的对比试验，量化比较不同规则的偏袒性与稳定性，结论扎实可靠。*Rule replay quantifies bias and stability, supporting robust conclusions.*

8.2 Weaknesses

- **粉丝基数静态假设：**假设选手“基础粉丝量全程不变”，未充分考虑赛季中的突发舆情与话题事件对粉丝基数的动态影响，可能与真实场景存在偏差。*Static fan-base assumption may miss time-varying shocks.*
- **参数区间依赖试运行：**参数范围采用“先宽后窄”的收敛策略虽能降低主观偏差，但收缩区间仍依赖小规模试运行结果，存在一定统计偶然性。*Narrowing ranges based on small pilots can introduce chance effects.*
- **特征维度有待扩展：**模型暂未纳入社交媒体互动质量（如评论情感倾向）、舞种类型适配度等变量，未来可进一步丰富特征以提升解释力与预测性能。*Add sentiment/engagement and style-fit variables to improve modeling.*

9 Conclusion

本文研究 DWTS 评分机制中的结构性矛盾：如何在专业评委的技术评价与观众的人气投票之间取得稳健平衡。在缺乏直接投票数据的现实约束下，我们从制度与机制角度分析规则运行与偏差，并提出兼顾公平性、可解释性与参与度的改进方案。核心发现表明，争议源于制度的放大效应，需从规则层面加以修正。

总体结论有三点：其一，评委分与粉丝票反映不同维度的价值，问题在于对粉丝影响缺乏合理约束，导致技术评价被稀释；其二，百分比制与排名制均存在内生偏向，在“人气—技术”差异显著时易放大极端结果；其三，在关键节点引入有限评委干预或设置粉丝分上限，能在不削弱参与的前提下提升公平性与稳定性。总体结论如下：

- 评委分与粉丝票反映不同维度的价值；关键在于对粉丝影响施加合理约束，避免技术评价被稀释。
- 百分比制与排名制均存在内生偏向；在人气与技术差异显著时，易放大极端结果。
- 在关键节点引入有限评委干预，或设置粉丝分上限，可在不削弱参与的前提下提升公平性与稳定性。

为此，本文提出“双轨线性积分系统”，以对称、线性的“30+30”框架并行整合专业与大众，减少隐含放大并提升可解释性与接受度。该思路可推广至其他“专业评价+大众投票”的赛事；未来可在保持核心结构的前提下，随参与方式与数据来源的丰富持续优化，实现公正性、娱乐性与黏性的共赢。

References

- [1] S. D' Angelo, T. B. Murphy & M. Alfò. "Latent Space Modeling of Multidimensional Networks with Application to the Exchange of Votes in Eurovision Song Contest." *arXiv preprint arXiv:1807.06517*, 2018.
- [2] M. Blangiardo & G. Baio. "Evidence of bias in the Eurovision song contest: modelling the votes using Bayesian hierarchical models." *arXiv preprint arXiv:1310.3501*, 2013.
- [3] R. Fairstein, A. Lauz, K. Gal & R. Meir. "Modeling People's Voting Behavior with Poll Information." *arXiv preprint arXiv:1902.04118*, 2019.
- [4] L. Chen, P. Xu & D. Liu. "Experts versus the Crowd: A Comparison of Selection Mechanisms in Crowdsourcing Contests." *SSRN Electronic Journal*, 2015. DOI: 10.2139/ssrn.2631317.
- [5] D. Zhang. "Methods and Rules of Voting and Decision: A Literature Review." *Open Journal of Social Sciences*, vol. 8, no. 9, pp. 310–326, 2020.
- [6] L. S. Shapley & M. Shubik. "A Method for Evaluating the Distribution of Power in a Committee System." *American Political Science Review*, vol. 48, no. 3, pp. 787–792, 1954.
- [7] "Voting Matters." McDougall Trust, <https://www.mcdougall.org.uk/>.
- [8] "Probabilistic voting model." In *Voting Theory*, https://en.wikipedia.org/wiki/Probabilistic_voting_model.
- [9] "Google Trends." Google LLC, <https://trends.google.com/>.
- [10] General Mills. "pytrends: Unofficial API for Google Trends." GitHub repository, <https://github.com/GeneralMills/pytrends>.

给组委会的建议（附）

基于本研究对 DWTS 历史计分规则（排名法与百分比法）的量化评估，我们发现复杂的权重计算往往是导致观众困惑与结果争议的根源。为了在提升节目娱乐性的同时捍卫舞蹈竞技的专业底线，我们建议组委会采用一种全新的“双轨线性积分系统”（**Dual-Track Linear Scoring System, DTLSS**）。

该机制的设计哲学在于“解耦与制衡”：将专业评分与大众投票解耦为两条平行的评价赛道，并通过线性映射实现两者权重的精确制衡。

新积分方式：“30+30” 计分模式

我们建议摒弃当前的“票数百分比”或的“综合排名”计算，转而采用我们提出的积分叠加制。该模型设定两条赛道拥有完全一致的权重上限（30 分）：

- **评委赛道 (Judges' Track)**: 维持现状。三位评委给出的分数直接累加，满分 30 分。这代表了选手的技术硬上限。
- **粉丝赛道 (Fans' Track)**: 将粉丝投票的排名直接按序转化为分数，满分同为 30 分。这代表了选手的人气硬上限。

计算逻辑与规则

粉丝赛道的得分 S_{Fan} 仅取决于选手在当周观众投票中的相对排名 $Rank_{Fan}$ 。其计算公式如下：

$$S_{Fan} = S_{max} - \delta \times (Rank_{Fan} - 1) \quad (28)$$

其中， $S_{max} = 30$ 为粉丝赛道满分， $\delta = 2$ 为排名步长。即：观众投票第 1 名直接获得 30 分，第 2 名获得 28 分，以此类推。

最终裁定分数为两者之和，总分最低者淘汰：

$$TotalScore = Score_{Judge} + Score_{Fan} \quad (29)$$

优势一：极致的观众友好度 (Audience Accessibility)

与以往观众无法直观判断“几百万票能抵消多少评委分”的黑箱状态不同，DTLSS 机制具有极强的可解释性 (**Interpretability**) 和电视表现力：

- **直观的激励机制**：规则转化为简单的线性奖励——“在观众投票中每提升一名，总分增加 2 分”。这种清晰的反馈回路比复杂的百分比算法更能有效调动观众的投票热情。
- **悬念可视化的直播体验**：在直播环节，屏幕可分屏显示确定的“评委分（基数）”与动态滚动的“粉丝分（变数）”。观众可以清晰地感知竞争态势：“选手 A 目前评委分落后 2 分，只需在粉丝排名中超过选手 B，即可实现反超。”这种直观的加法运算将极大地增强节目的紧张感与互动性。

优势二：强制性的结构平衡 (Structural Balance)

我们在 Section 6.4 的分析中指出，Season 27 的争议源于粉丝票数在百分比法下的无限膨胀效应。新机制通过“量纲统一”彻底解决了这一隐患，构建了数学层面的防守机制。

1. 流量封顶机制 (Influence Capping)

无论某位明星的粉丝基数有多庞大（即使出现如 Bobby Bones 式的离群点），他在粉丝赛道的收益上限被严格锁定为 30 分。这从根本上限制了超额选票的边际效用，防止人气因素过度稀释专业评分的权重。

2. 专业底线的数学防御

我们可以通过以下情景回测来验证该机制的鲁棒性：

- **情景假设：**选手 X 舞技极差（评委分 15 分），但人气极高（粉丝分满分 30 分），其总分为 $15 + 30 = 45$ 分。
- **对比选手：**选手 Y 表现优异（评委分 27 分），但人气中等（粉丝排名第 3，获 26 分），其总分为 $27 + 26 = 53$ 分。
- **结果判定：** $53 > 45$ ，选手 Y 胜出。

上述推演证明，即使在人气完全碾压的情况下，过低的技术分依然会导致淘汰。DTLSS 机制在数学原理上杜绝了“严重偏科”选手夺冠的可能性，确保了比赛结果的公正性。

总结

“双轨线性积分系统”并非削弱粉丝的参与感，而是将抽象的支持转化为看得见、算得清的“积分奖励”。它用最简单的加法逻辑，在复杂的社会选择问题中实现了专业评判与大众喜爱的完美制衡。我们建议组委会在 Season 35 优先试行此方案，作为解决长期以来“评分争议”的破局之策。

AI Use Statement

1. Did you use AI tools in your work? Yes.

To enhance non-native language expressions in the course of academic paper writing, our research team has adopted AI technology. Before integrating any AI-assisted output into designated sections of the paper, we subject it to multiple layers of human review to ensure it meets academic standards. Moreover, to guarantee compliant use of the technology, we begin by establishing a comprehensive semantic framework and then guide AI-generated content through meticulously crafted prompts. Throughout this process, we strictly adhere to research ethics, ensuring there is no misuse of AI technology.

2. Which tools did you use and how?

- 1) **ChatGPT-5:** We used ChatGPT-5 to polish the language of the Introduction section.

Query:

Polish this paragraph in an empirical tone: Dancing with the Stars (DWTS) is a classic television program that blends professional dance competition with popular entertainment elements. In each season, celebrities from diverse fields partner with professional dancers to compete in an elimination-style contest through weekly dance performances. A contestant's ultimate fate is determined by two core dimensions: first, the professional scores awarded by judges based on dance technique, choreographic creativity, and other criteria; second, the audience votes that reflect public preferences.

Output:

Polished Version—Dancing with the Stars (DWTS) is a long-running television program that integrates professional dance competition with popular entertainment. In each season, celebrities from diverse professional backgrounds are paired with professional dancers and compete in an elimination-style format through weekly dance performances. A contestant's progression and eventual outcome in the competition are jointly determined by two core components: professional judges' scores, which evaluate dance technique, choreography, and artistic expression, and audience votes, which reflect public preferences.

- 2) **ChatGPT-5:** After completing Question 1, we used GPT to generate the code for consolidating all the data of Seasons 1 to 34 into a single CSV file.

Query:

I need to consolidate the estimated fan volume data for Seasons 1 to 34 into a single file, with the requirement to keep all original column names unchanged. Please provide the corresponding code.

Output (partial code):

```
if len(all_data) == 0:  
    raise ValueError("All files failed to read or no valid data exists, merge")  
# Merge all DataFrames, ignore original indices, and reset a new continuous index  
merged_df = pd.concat(all_data, ignore_index=True, sort=False)
```

```
# Step 4: Final data sorting and cleaning
# Sort by original Chinese columns: Season → Week → Contestant Name for su
merged_df = merged_df.sort_values(by=["Season", "Week", "Name"], ignore_in
# (Optional) Remove duplicate rows (if duplicate data exists for the same
merged_df = merged_df.drop_duplicates(subset=["Season", "Week", "Name"], k
# Step 5: Save the merged result
merged_df.to_csv(Config.OUTPUT_FILE, index=False, encoding="utf-8-sig")
print(f"\n\u25b6Merging\u25bccompleted!\u25bcThe\u25bcfull-season\u25bcdata\u25bchas\u25bcbeen\u25bcsaved\u25bcto:\u25bc{Con
```

- 3) **Grammarly:** We used Grammarly to correct minor grammatical errors in the text.