19th International Fryderyk Chopin Piano Competition in Warsaw (Poland)

Statistical analysis of jury scores

author: mka

November 8, 2025

Abstract

This report presents a multi-angle analysis of jury scores[1] across all rounds of the 2025 Chopin Competition: distributions and use of the 1–25 scale; judge tendencies and consistency; correlations and clustering; leave-one-judge-out sensitivity; normalization method comparisons; PCA (contestants and judges); dispersion and outliers; rank agreement (Kendall's τ vs Pearson's r); and stability tests (jury-composition bootstrap and ± 1-point perturbations). The goals are to describe the structure of the scores, identify potentially sensitive areas, and verify how robust the outcomes are to reasonable technical and probabilistic changes.

General findings

Within the limits of the data and methods used, the picture is consistent:

- **Judging was fair.** No patterns suggest intentional action to benefit or harm particular contestants; score divergences are well explained by natural differences in calibration and taste, not coalitions or systemic bias.
- **The winners are clear.** Top placements remain stable under alternative normalizations, jury bootstrap, and small ± 1 perturbations. Distributions are narrow and the probabilities of holding position are high.
- **The mid-table is interchangeable.** Positions about 3–9 are more sensitive: small differences (e.g., ≤ 1 pt) between neighbors can swap order under random jury composition or slight noise in individual scores.
- **Normalization does not meaningfully change the final results.** Per-judge, per-stage normalization aligns scales across judges and does not affect the top hierarchy; some effects are visible mainly in the middle of the field, where tighter margins can lead to minor reshuffles. One of the laureates would not have reached the final.
- **Large spreads reflect taste, not strategy.** Contestants with the most polarized scores simply elicited strong artistic reactions; their outcomes are more sensitive to how points are aggregated (normalization, jury composition, or small perturbations).
- **No judge alliances.** Pairwise correlations are moderate; clustering does not reveal durable blocs.

1 Diversity/extremes and rank agreement are computed on raw per-judge stage scores (no corrections). Robustness of final outcomes is examined with two independent statistical simulations: (a) bootstrap resampling (with replacement) of judges for each stage, and (b) a sensitivity/noise test by adding random perturbations in the range $[-1, 1]$ to each judge's score. In each simulation, the hypothetical final outcome is recomputed strictly following the competition's official pipeline.

1 Score scale

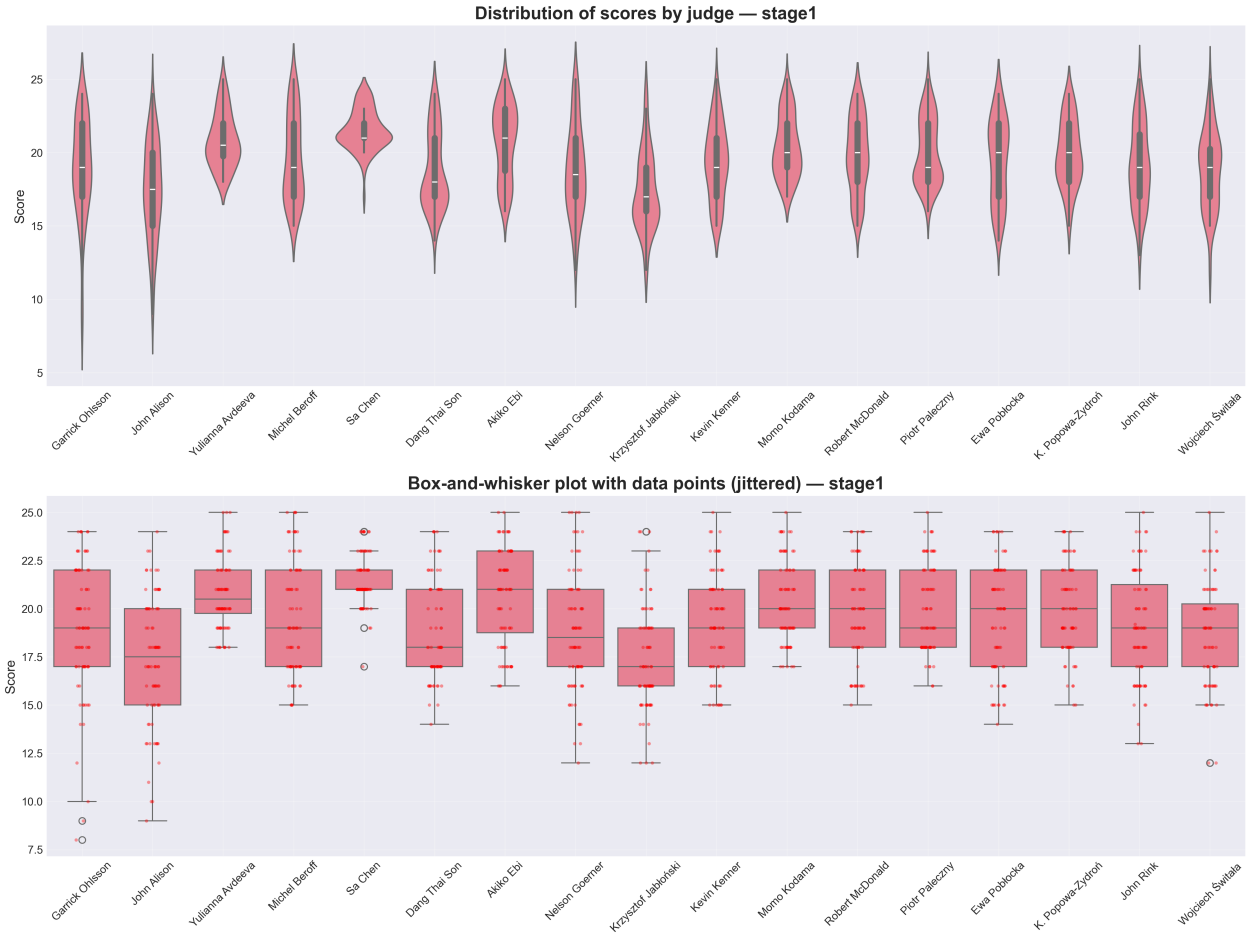


Figure 1: Distributions of judges' scores — Stage I.

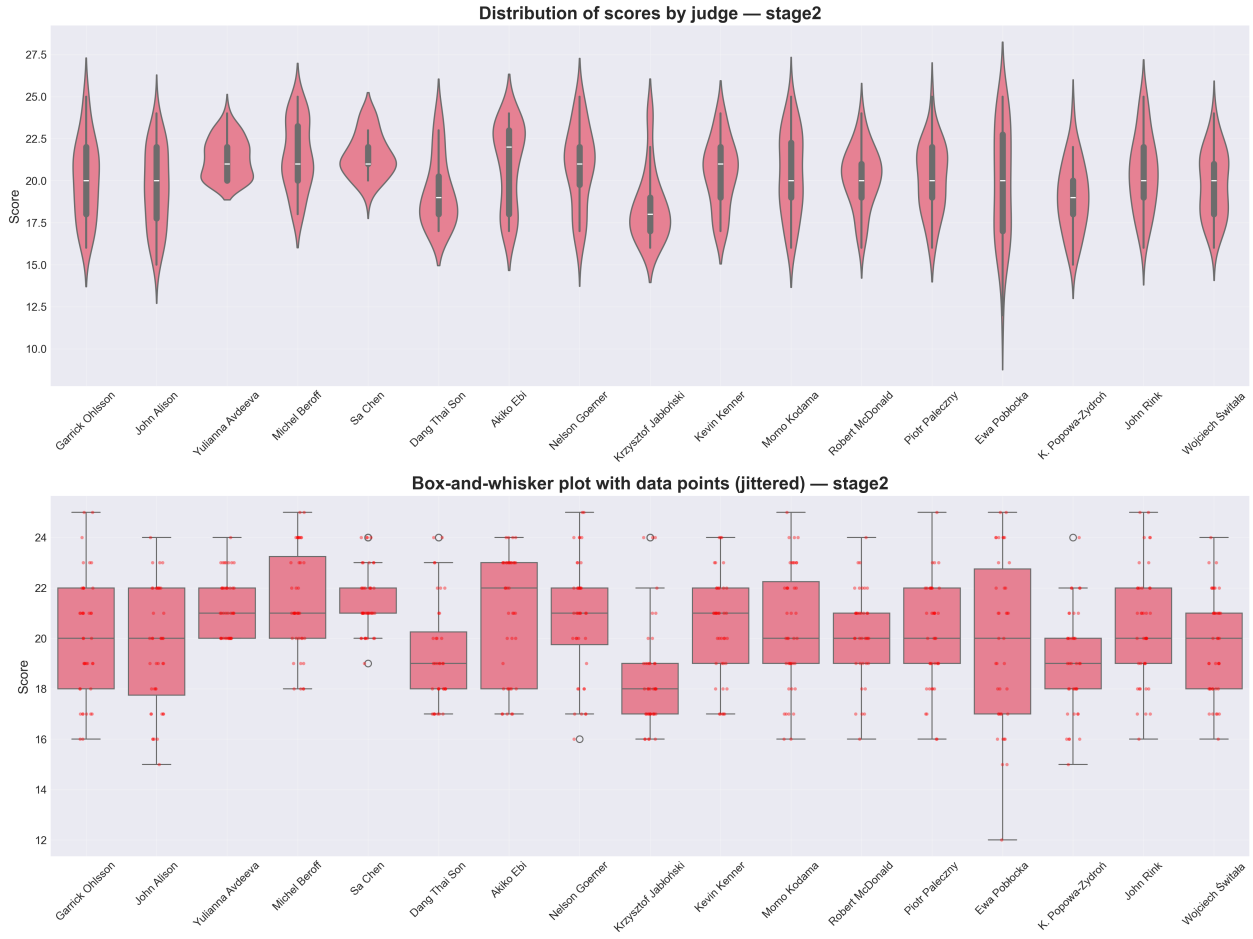


Figure 2: Distributions of judges' scores — Stage II.

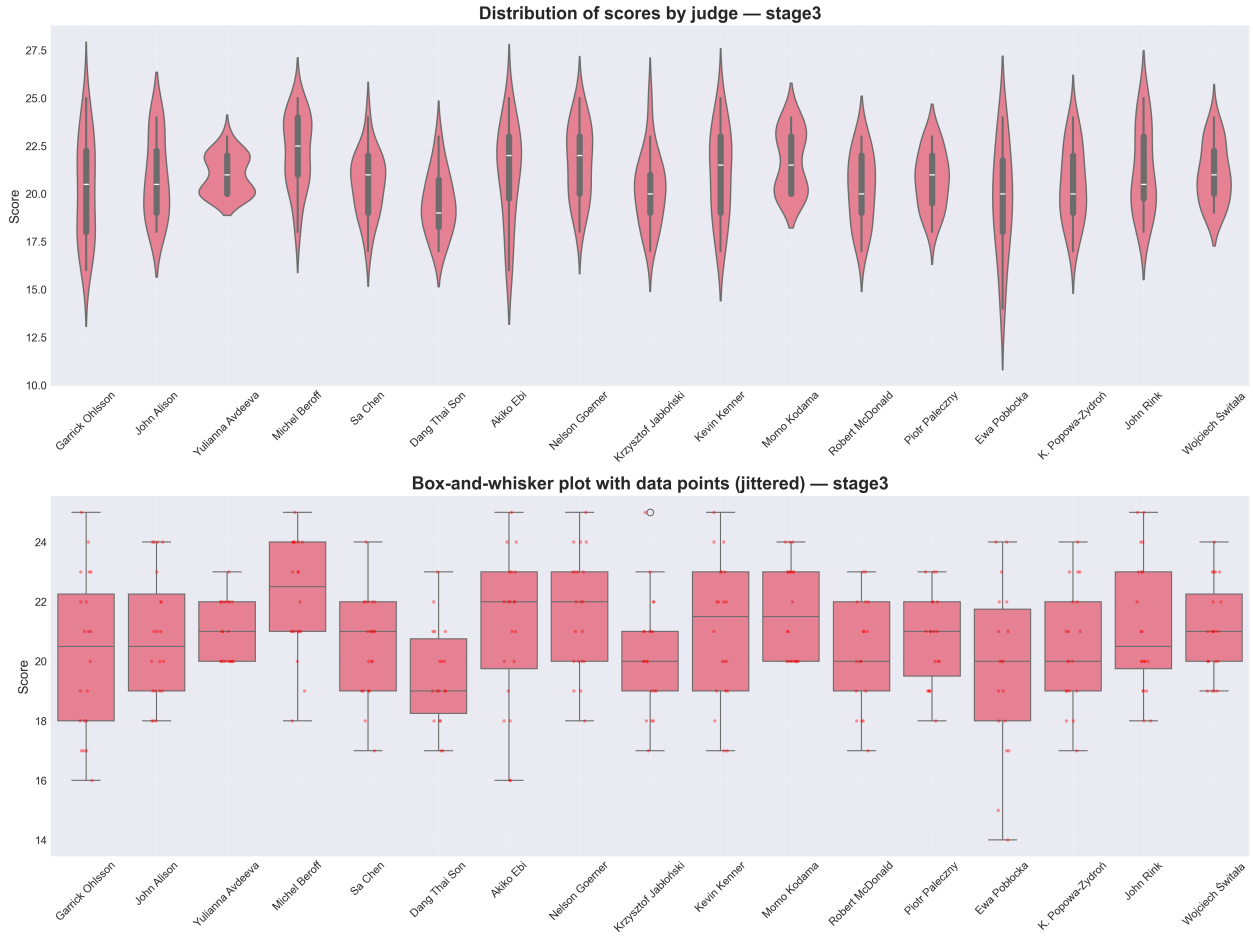


Figure 3: Distributions of judges' scores — Stage III.

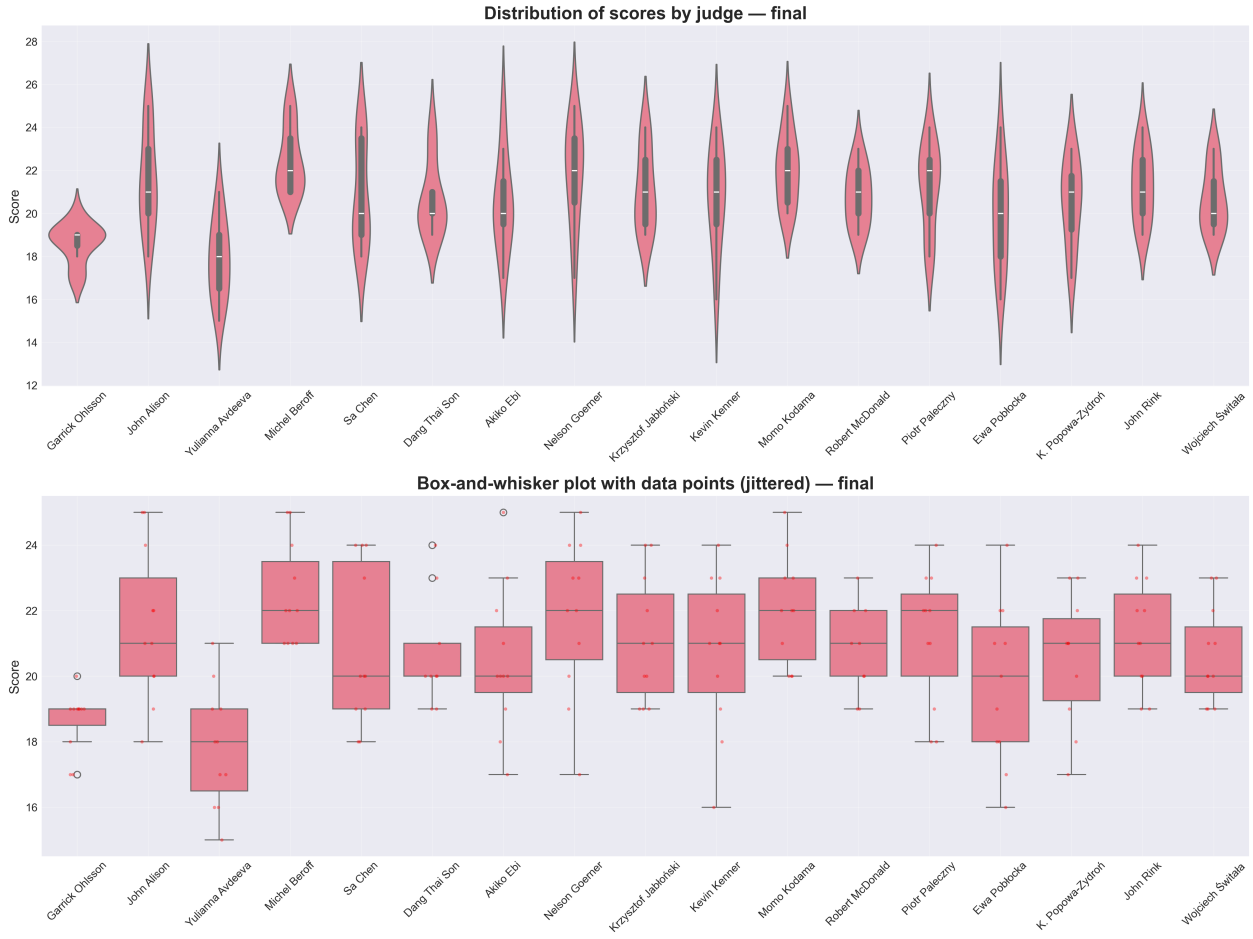


Figure 4: Distributions of judges' scores — Final.

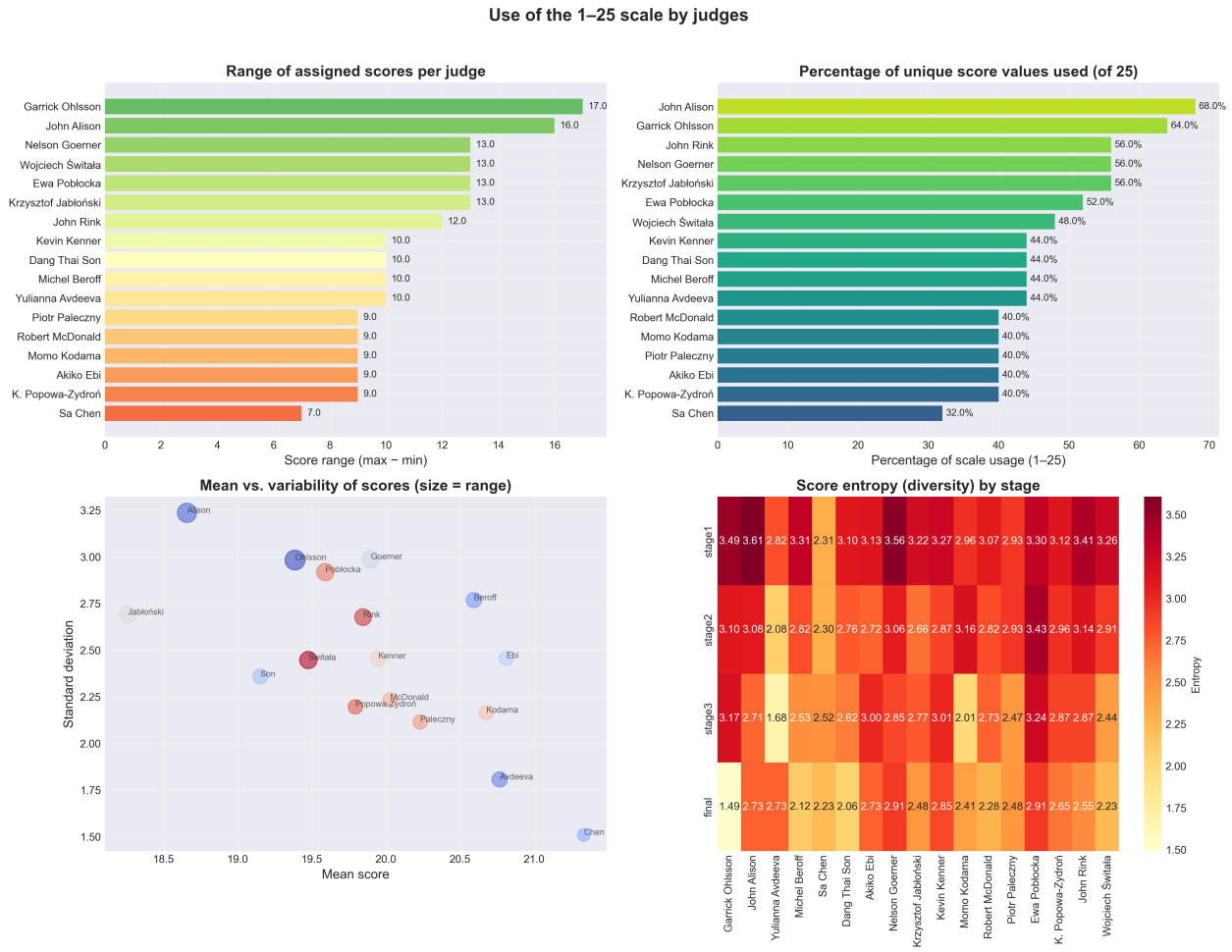


Figure 5: Use of the scale: (left) score range; (right) % of unique values; (bottom-left) mean vs SD (size = range); (bottom-right) score entropy by stage.

Differences in the width and location of distributions indicate distinct "listening philosophies." Some judges used a wide spread from the outset (e.g., Stage I), suggesting boldness in rewarding standouts and penalizing weaknesses; others kept a narrower band, preferring finer differentiation among comparable interpretations.

High score entropy[2] in a stage can reflect broader ability variance among contestants or repertoire that invites interpretive latitude.

2 Shannon entropy of the actual values a judge used in a given stage. Intuitively: how diversely the judge used the available 1–25 values. Higher entropy = more spread and balanced use; lower = more concentrated in a narrow band.

2 Judging tendencies

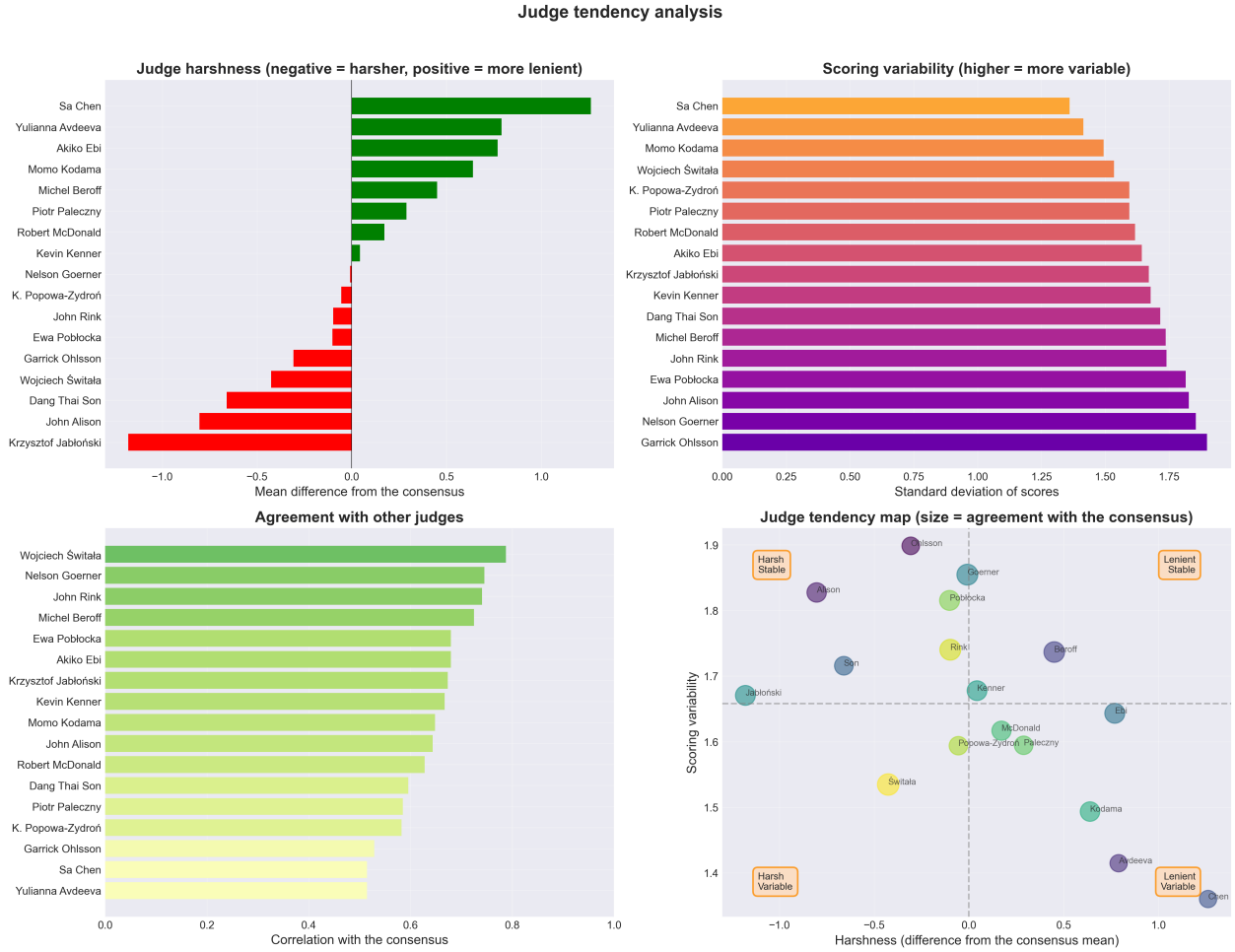


Figure 6: Tendencies: harshness/leniency (difference from consensus), variability, agreement with others (correlation with consensus), tendency map (size = agreement).

The tendency map summarizes calibration and consistency: to the right are judges who, on average, score above the consensus (more lenient), and to the left those below (harsher). Vertical position encodes variability; point size encodes agreement with others. The map describes calibration/consistency, not aesthetic preferences per se.

3 Correlations and clustering of judges

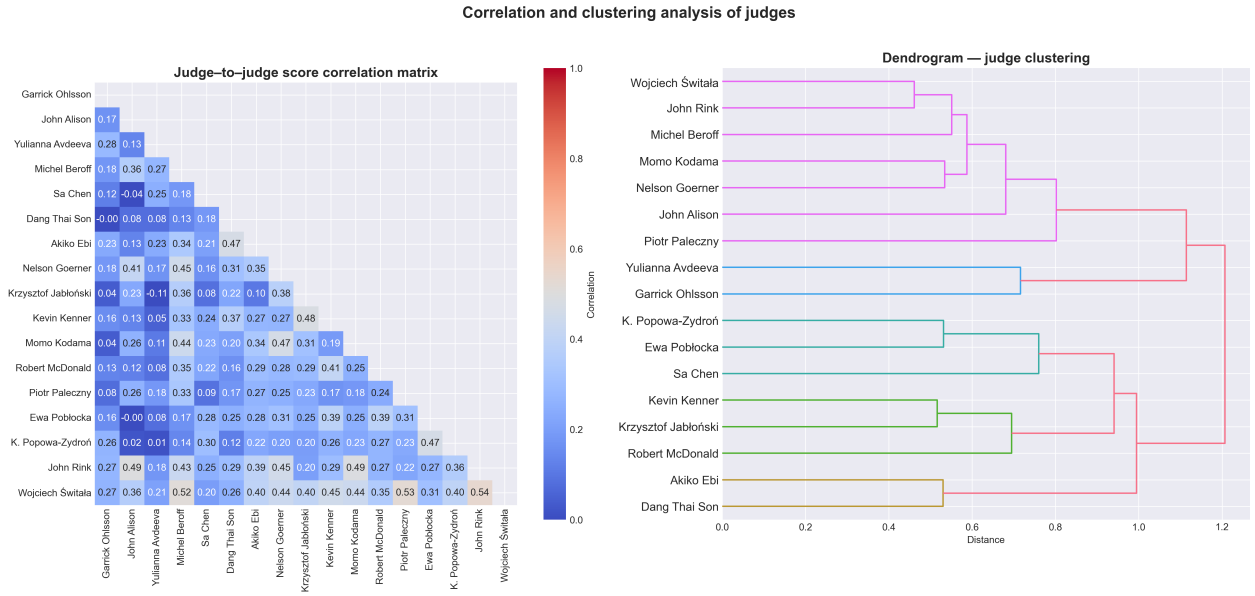


Figure 7: Score-correlation matrix and judge-clustering dendrogram. *No clear alliances.*

The correlation matrix and dendrogram do not confirm stable "aesthetic blocs"—rather ad-hoc proximities. This matters: divergences in scores are largely *substantive*, reflecting genuine interpretive preferences rather than factional discipline.

4 Impact of removing a judge (leave-one-out)

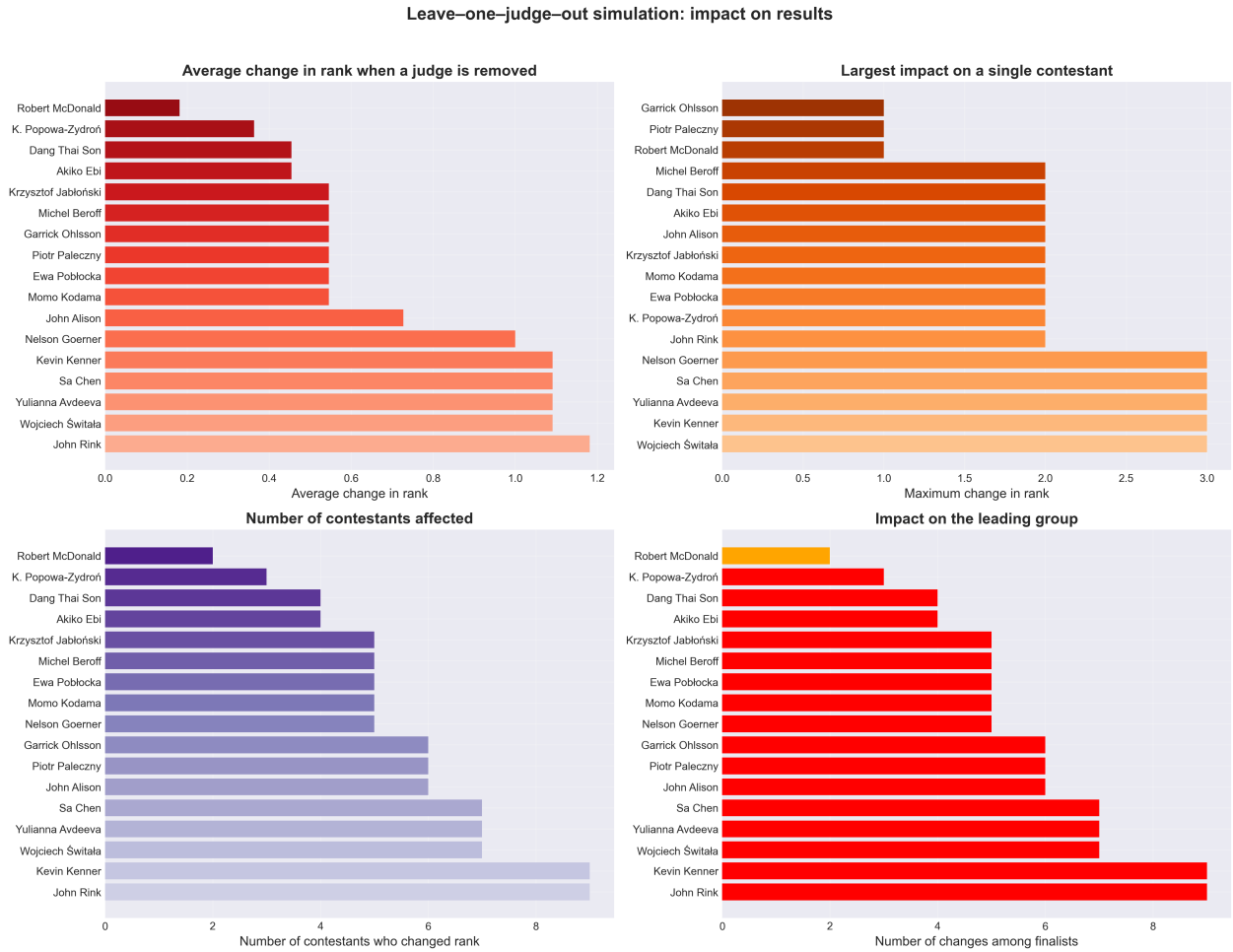


Figure 8: Sensitivity of results to removing a judge: average impact, maximum impact on an individual, number affected, impact on the leading group.

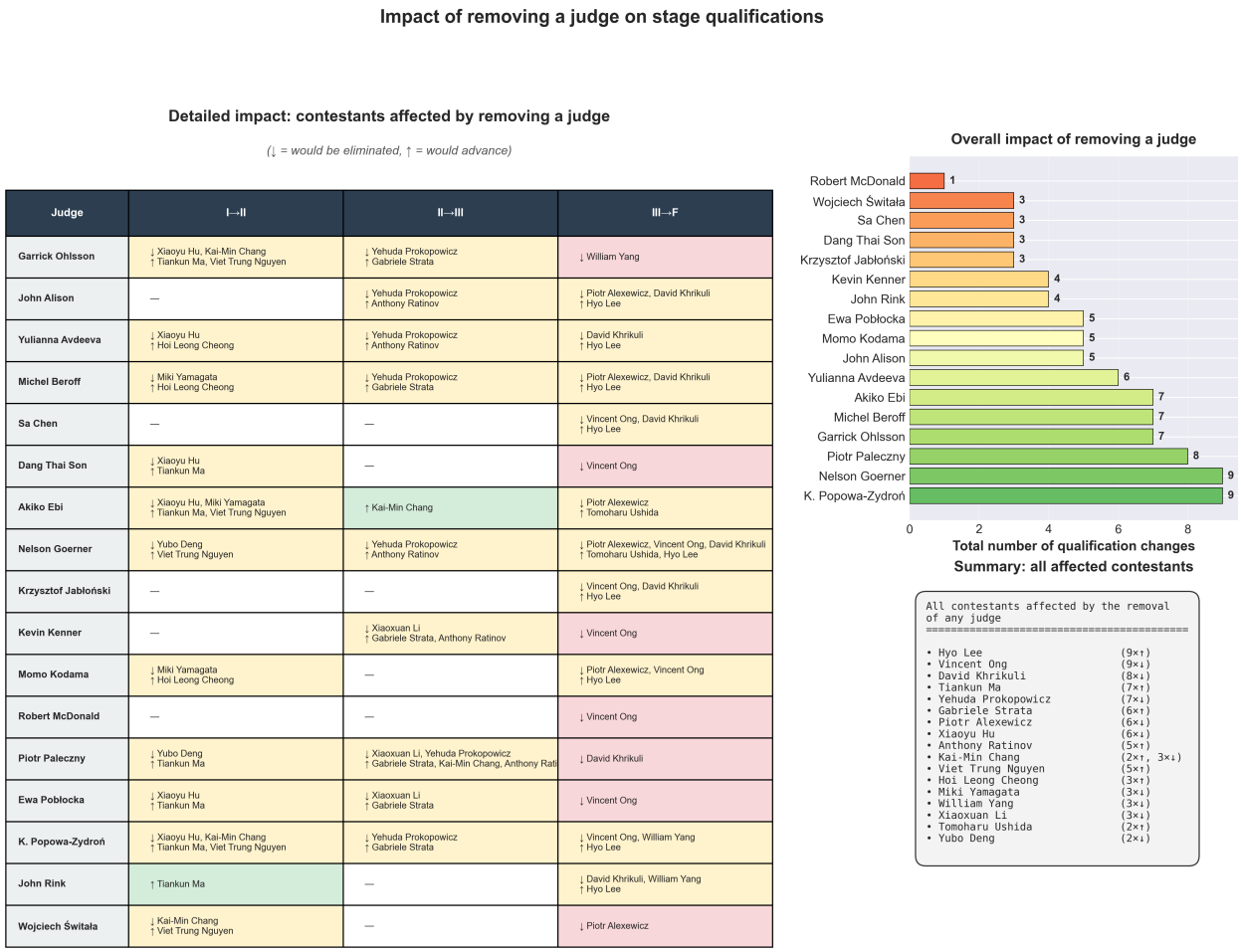


Figure 9: Impact of removing a judge on qualifications (I→II, II→III, III→F) and the total number of changes.

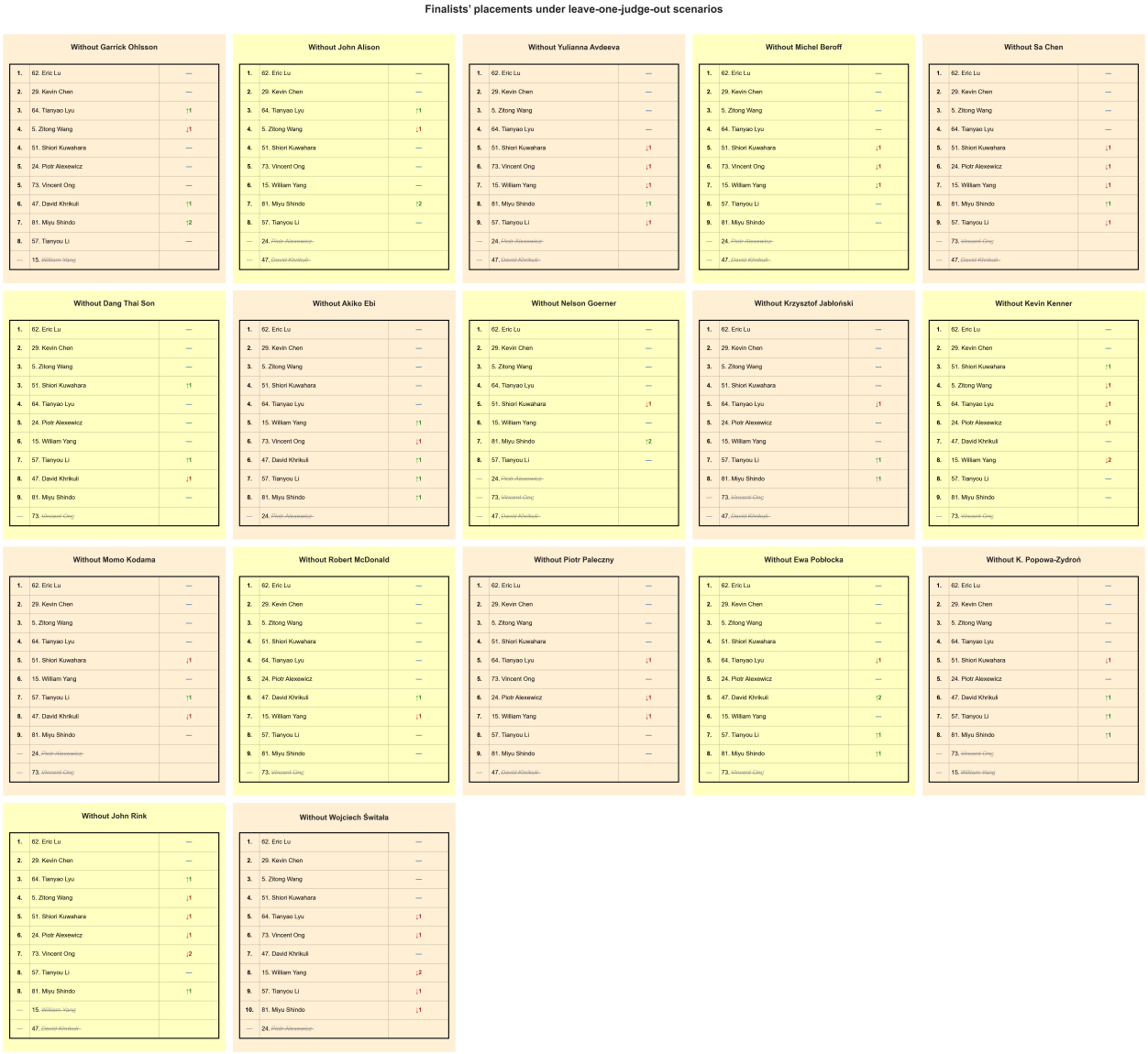
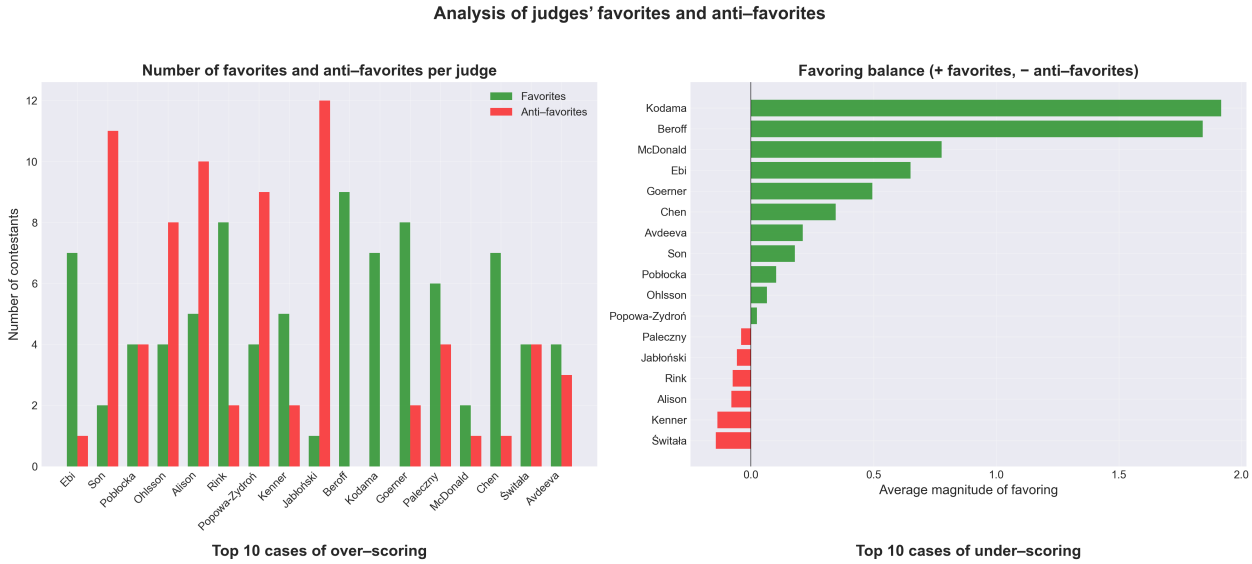


Figure 10: Finalists' placements under leave-one-judge-out scenarios ("without judge X").

Simulations show that a single judge can noticeably shift contestants with distinctive profiles (e.g., highly expressive tempo shaping, or conversely very restrained playing). Impact on the leaders is greatest for judges whose calibration and "taste" deviate most from the median. At the same time, the average effect on the overall outcome remains moderate—extremes tend to *offset* in the collective decision.

5 Favorites and anti-favorites



Top 10 cases of over–scoring

Judge	Contestant	Mean difference	Stages
Garrick Ohlsson	Yang (Jack) Gao	+2.33	3
Garrick Ohlsson	Piotr Pawlak	+2.33	3
Garrick Ohlsson	Yehuda Prokopowicz	+2.33	3
John Alison	Piotr Pawlak	+2.33	3
Michel Beroff	Hyuk Lee	+2.33	3
Ewa Pobłocka	Xiaoxuan Li	+2.33	3
Akiko Ebi	Hyo Lee	+2.31	3
Sa Chen	Tianyou Li	+2.25	4
Akiko Ebi	Piotr Alexewicz	+2.25	4
Michel Beroff	Yifan Wu	+2.18	3

Top 10 cases of under–scoring

Judge	Contestant	Mean difference	Stages
Garrick Ohlsson	Eric Guo	-2.33	3
Krzysztof Jabłoński	Yang (Jack) Gao	-2.33	3
John Alison	Vincent Ong	-2.25	4
Ewa Pobłocka	David Khrikuli	-2.25	4
Garrick Ohlsson	Piotr Alexewicz	-2.22	4
Dang Thai Son	Yehuda Prokopowicz	-2.22	3
Garrick Ohlsson	Kevin Chen	-2.21	4
Krzysztof Jabłoński	Zitong Wang	-2.19	4
Krzysztof Jabłoński	Piotr Alexewicz	-2.16	4
Krzysztof Jabłoński	Piotr Pawlak	-2.16	3

Figure 11: Favoring balance (plus/minus), number of favorites/anti-favorites per judge, and top cases (tables).

Every judge evaluated at least a few contestants differently from the rest. The tables highlight the most pronounced judge–contestant pairs recurring across ≥ 3 stages—either a shared interpretive wavelength or the opposite.

6 Normalization comparison

Figure 12: Comparison of final rankings across three normalization methods (Z-score, Min–Max, Rank).

Judges used the available scale differently: some employed broader ranges, others narrower; some were sparing with high scores, others awarded them more readily. Normalization mitigates these differences. The tables show how each method would affect the competition results.

7 PCA of contestants

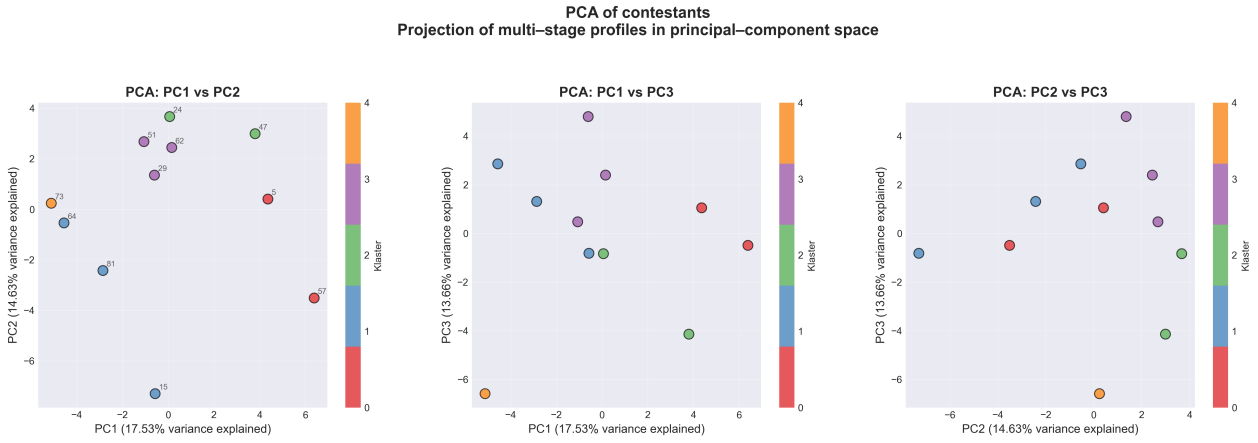


Figure 13: PCA of finalists (profile projection; clusters), with a cluster-membership table.

PCA groups finalists by similarity of multi-stage score profiles. PC1 primarily separates by *overall scoring level* (largest variance), while PC2/PC3 capture second-order differences (e.g., uneven vs

stable progression across stages). Clusters indicate finalists with similar round-to-round trajectories; small clusters/outliers mark distinctive profiles.

8 PCA of judges

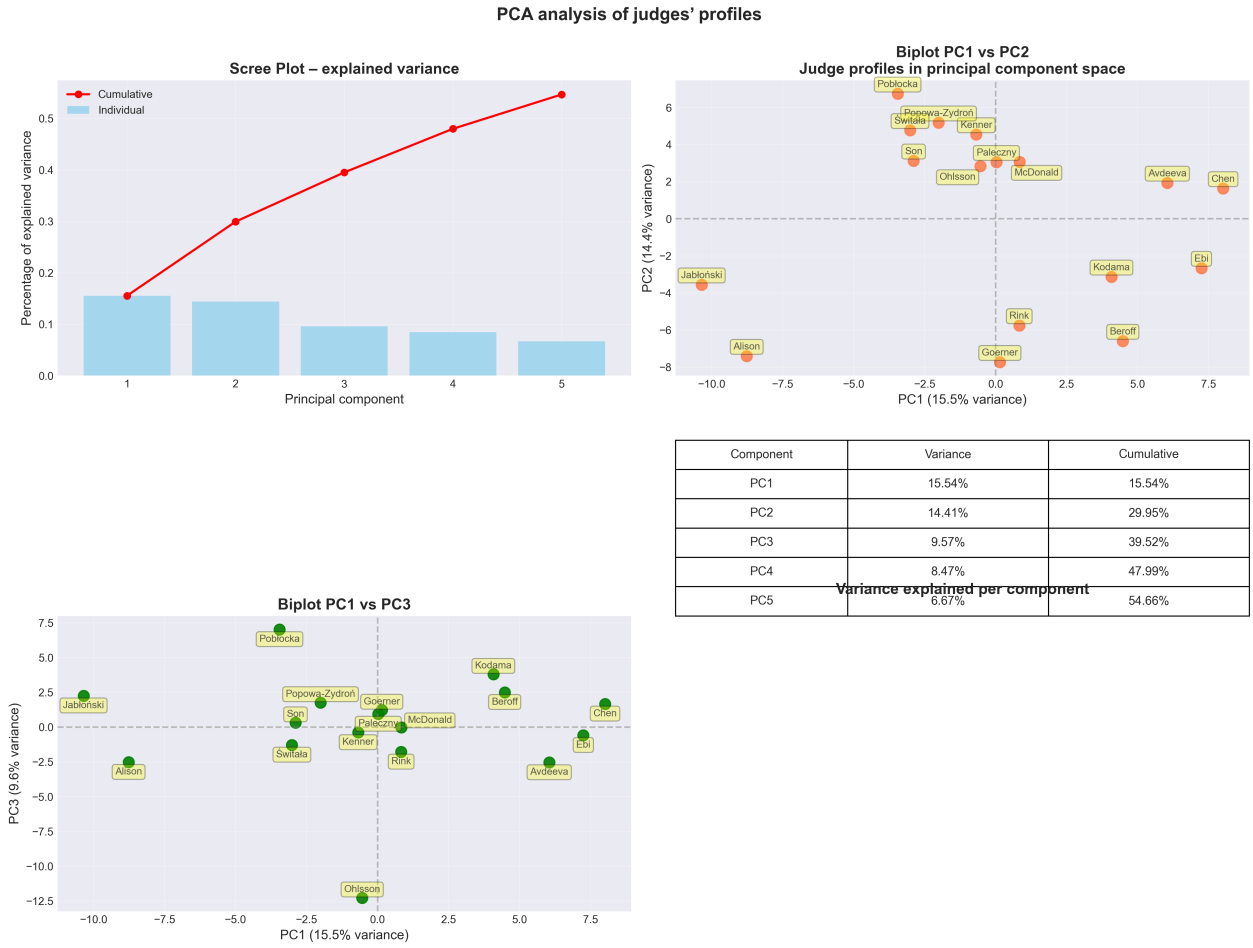


Figure 14: **PCA of judges' profiles (extended)**: scree plot, PC1–PC2 and PC1–PC3 biplots, table of explained variance.

Judges' PCA reveals diversity of *scoring styles*. PC1 behaves like a *calibration* axis (lenient vs harsh overall), while PC2 reflects *spread/consistency* of scale usage (more vs less variable scores). The scree plot supports using the first 2–3 components; biplots reveal groupings of similarly scoring judges.

9 Score heatmap

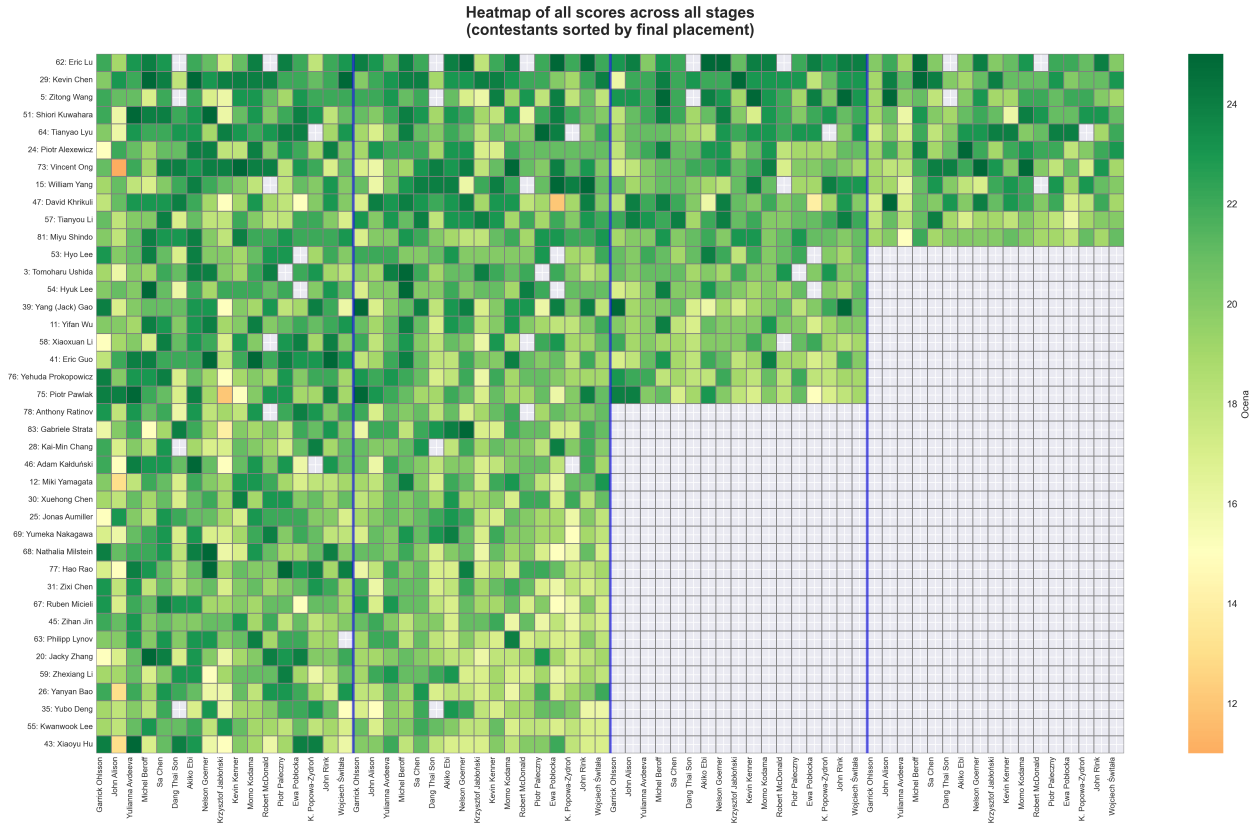


Figure 15: Heatmap of all scores (contestants sorted by final rank).

The full heatmap (columns = judges, rows = contestants sorted by final outcome) shows where the jury was more aligned (uniform colour bands) and where opinions diverged (mixed colours). Upper rows (higher-ranked finalists) typically show darker fields; lower rows, lighter. Column differences show individual judge calibration; striking orange cells mark extremely low scores from single judges.

10 Consistency of judging

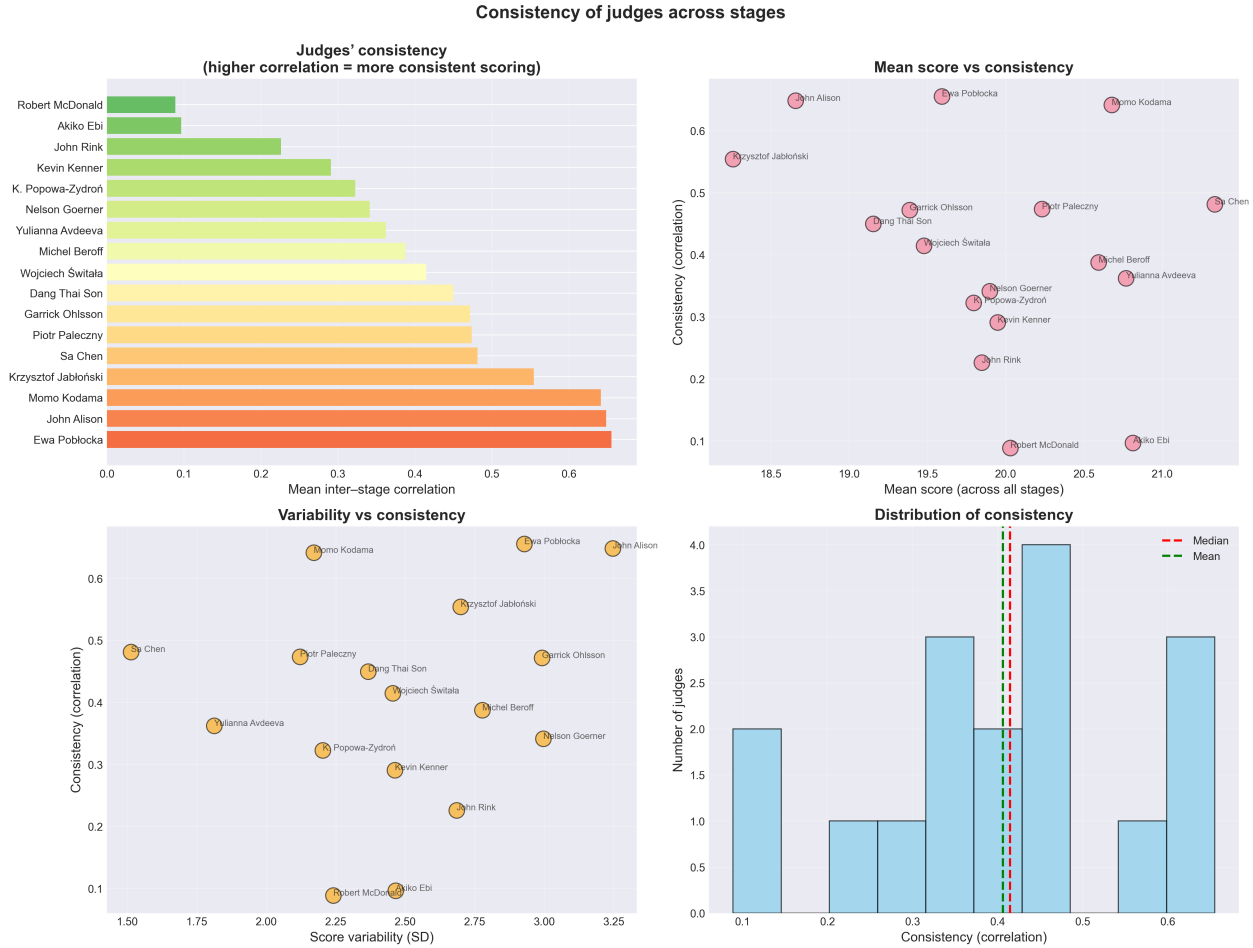


Figure 16: Consistency across stages: mean correlation per judge, mean vs consistency, distribution of consistency.

Inter-stage consistency (mean correlations) reveals two stances: (1) *more consistent* judges—higher agreement across stages and lower dispersion; (2) *more reactive* judges—lower agreement, higher sensitivity to repertoire/context. The "mean score vs consistency" relation shows that leniency/harshness alone does not guarantee consistency.

11 Finalists' trajectories across stages

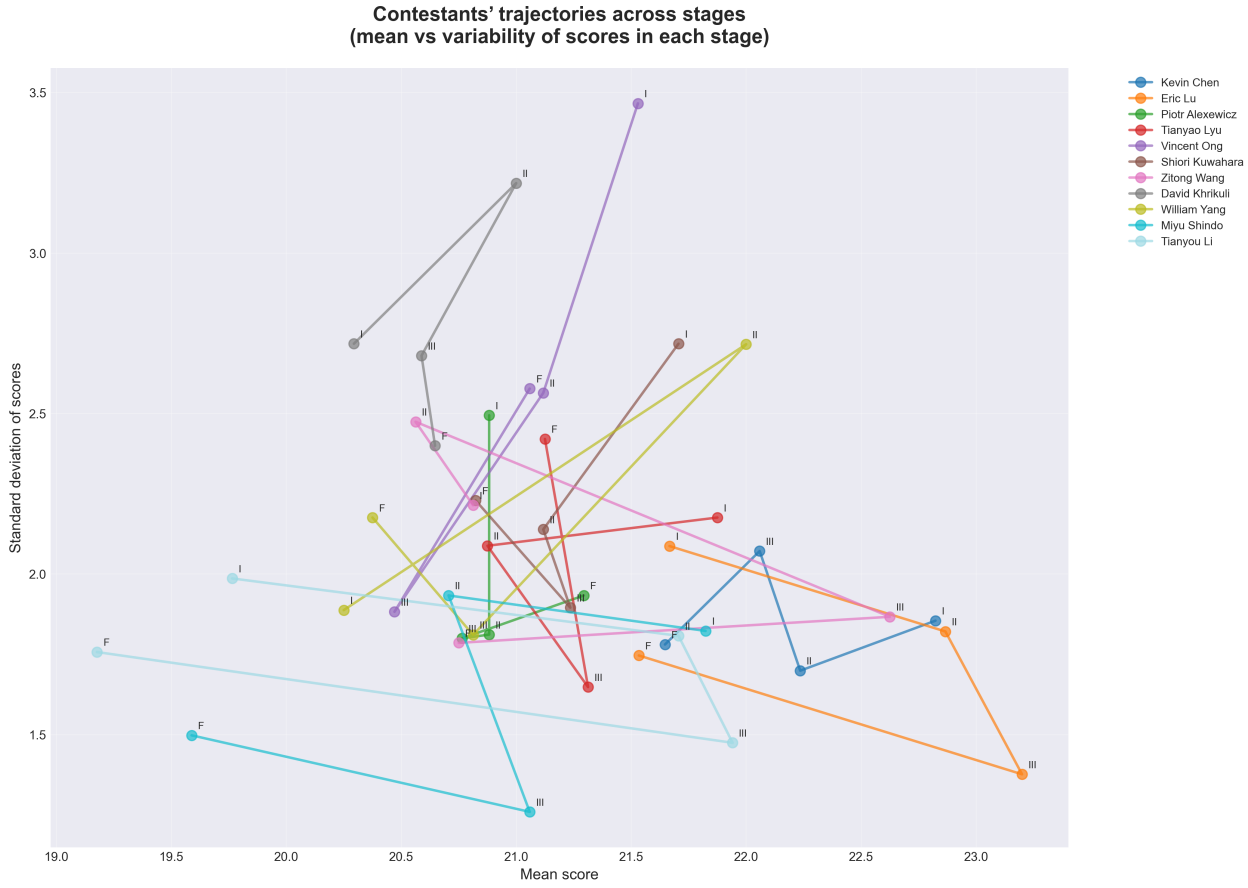


Figure 17: Trajectories (mean vs variability of scores) across stages for finalists.

Trajectories *(mean vs variability)* across rounds reveal who climbed *steadily* (small shifts, low dispersion) and where "quality jumps" occurred (material shifts in mean or changes in SD).

12 Score variation / “controversiality”

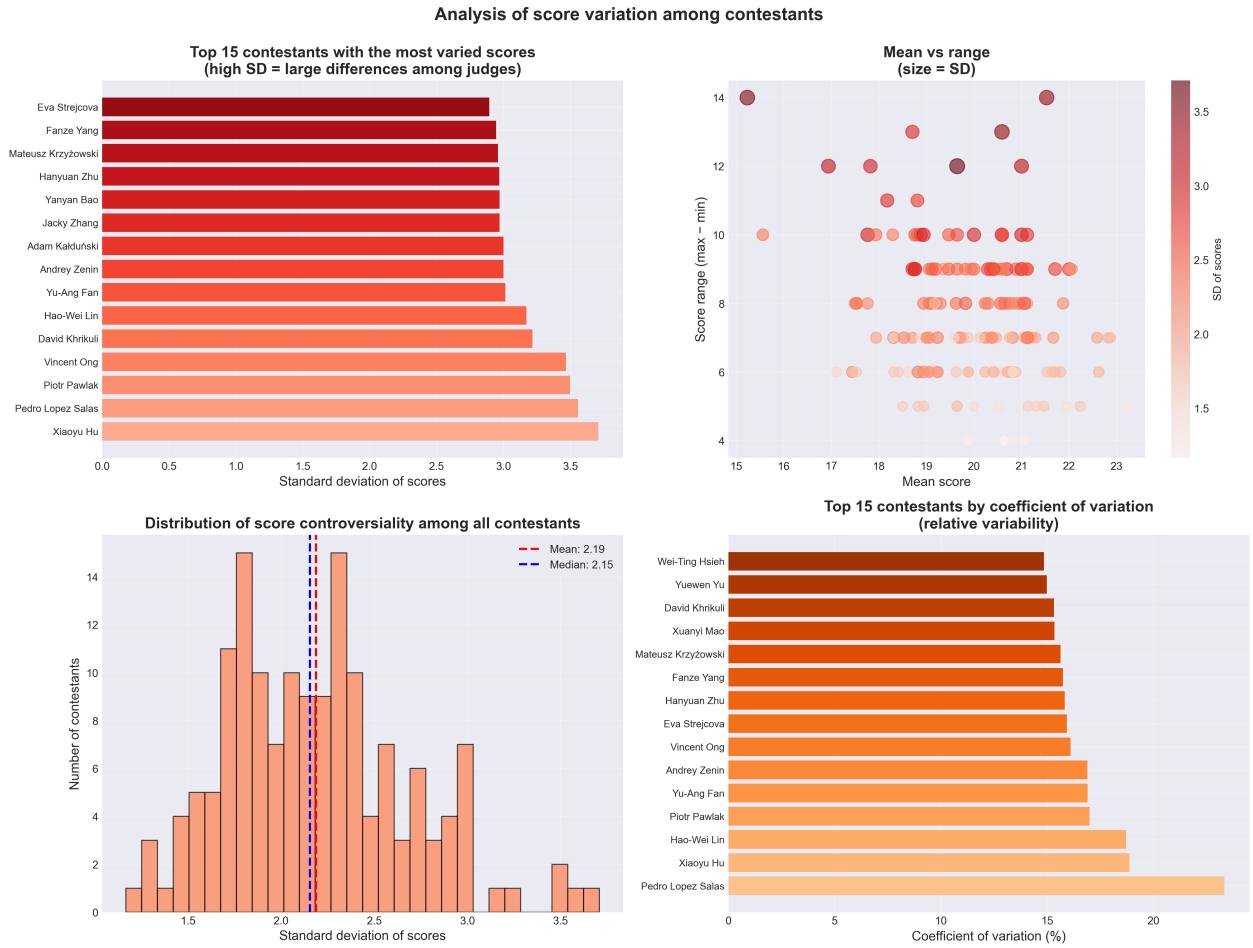


Figure 18: Highest SD, mean vs range (size = SD), SD histogram, top by coefficient of variation (CV).

High SD (top-left) indicates large spread among judges for a performance; wide *range* (max–min, top-right) flags extreme differences even if most scores cluster near the mean. The SD histogram (bottom-left) shows a narrow typical band with a few standouts. CV (bottom-right) highlights where spread was large *relative to the mean level*.

13 Outliers

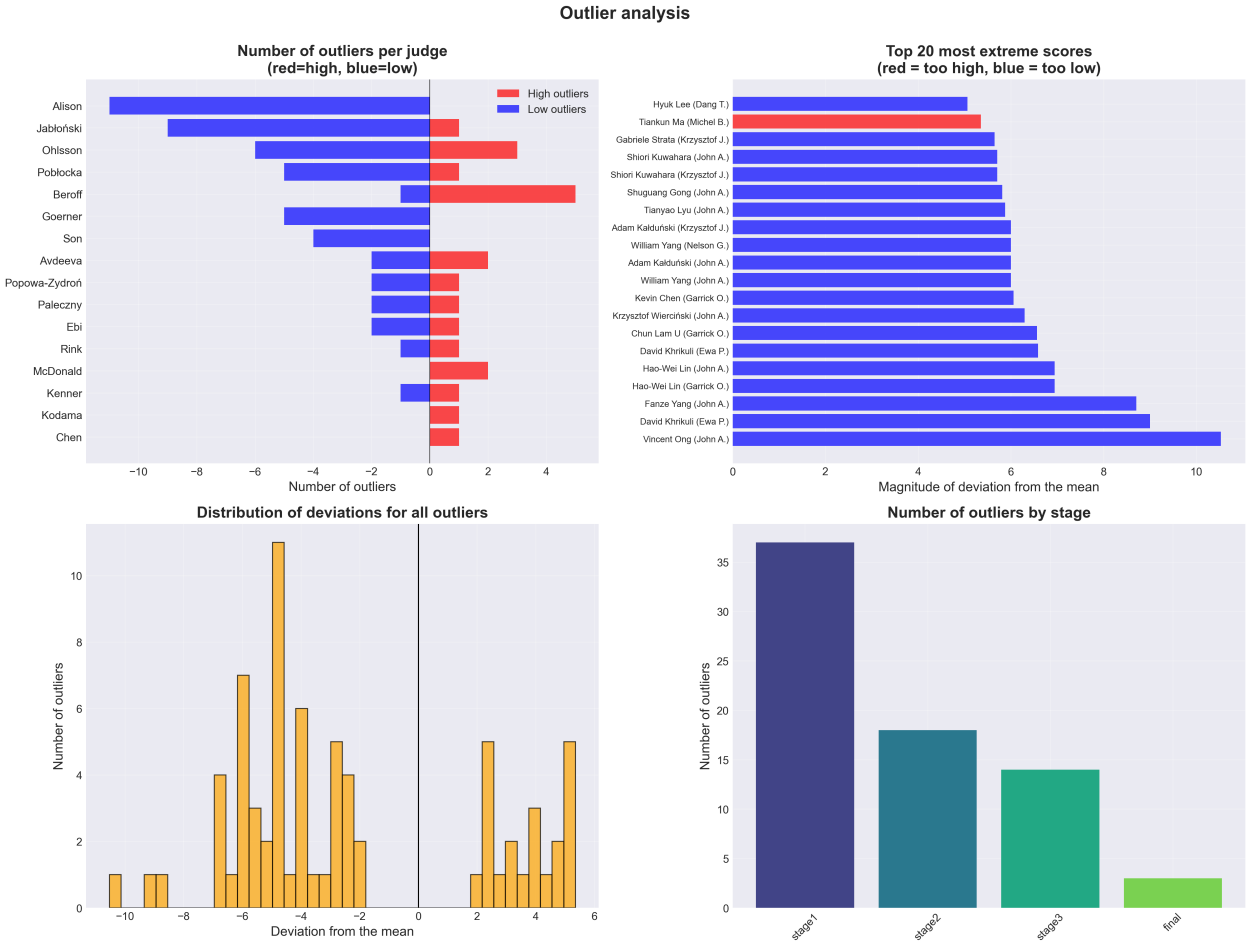


Figure 19: Outliers: counts per judge (high/low), top-20 extremes, distribution of deviations, number of outliers by stage.

Outliers are scores that markedly deviate from the consensus within a stage. The "high/low" splits and top-N lists show where a judge's calibration differed most from the rest. Outliers concentrate where natural spread is largest (often earlier rounds or polarizing performances).

14 Rank agreement

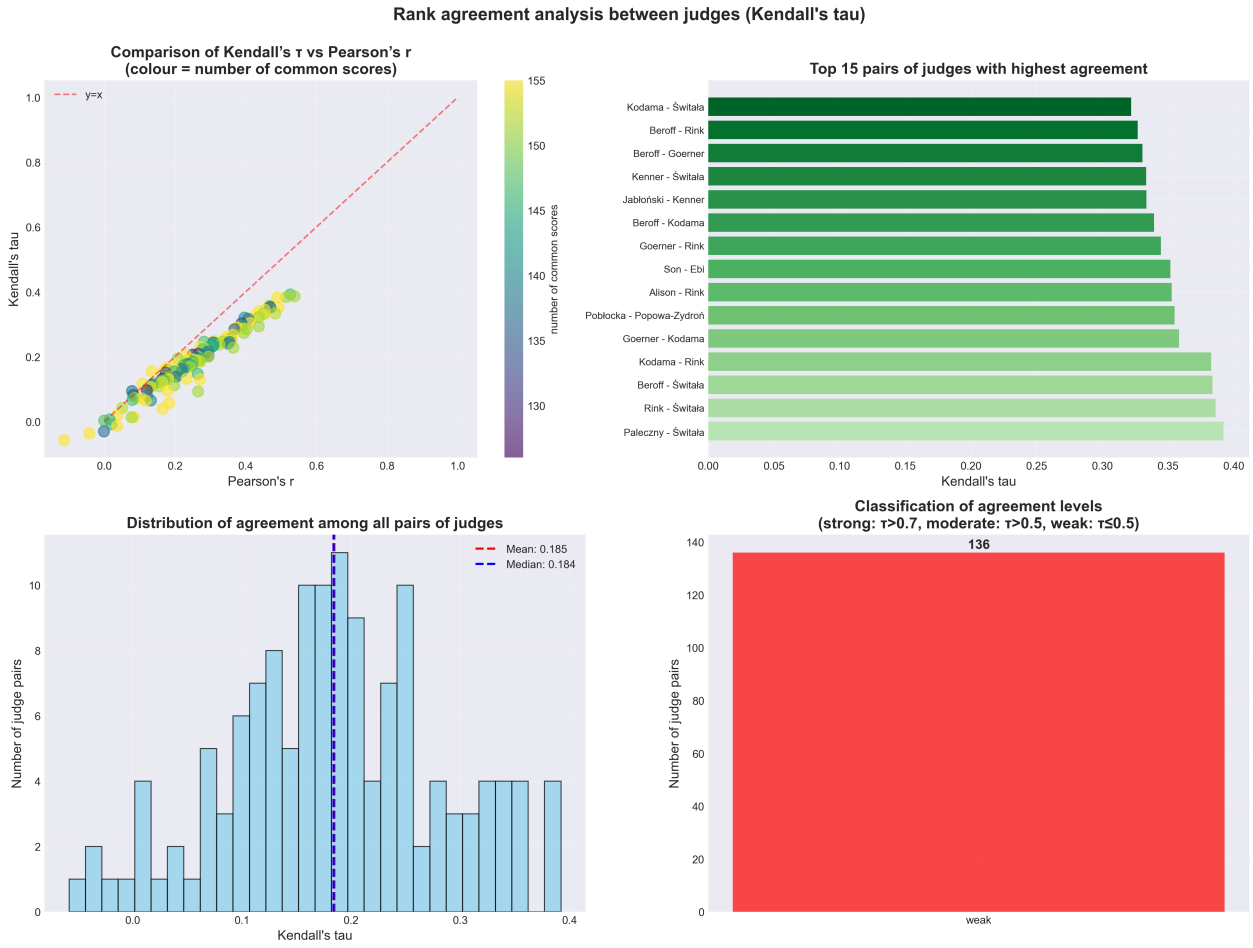


Figure 20: Rank agreement: Kendall's τ vs Pearson's r (comparisons and distributions).

Comparing Pearson's r and Kendall's τ separates two notions of agreement: r reflects *scaling* (linear agreement in values), whereas τ reflects *ranking* (order of contestants). Many pairs with $r > \tau$ imply judges can score "along a similar line" without agreeing on places. "Shared taste" should be inferred cautiously from τ, not just r.

15 Stability of finalists' results (bootstrap)

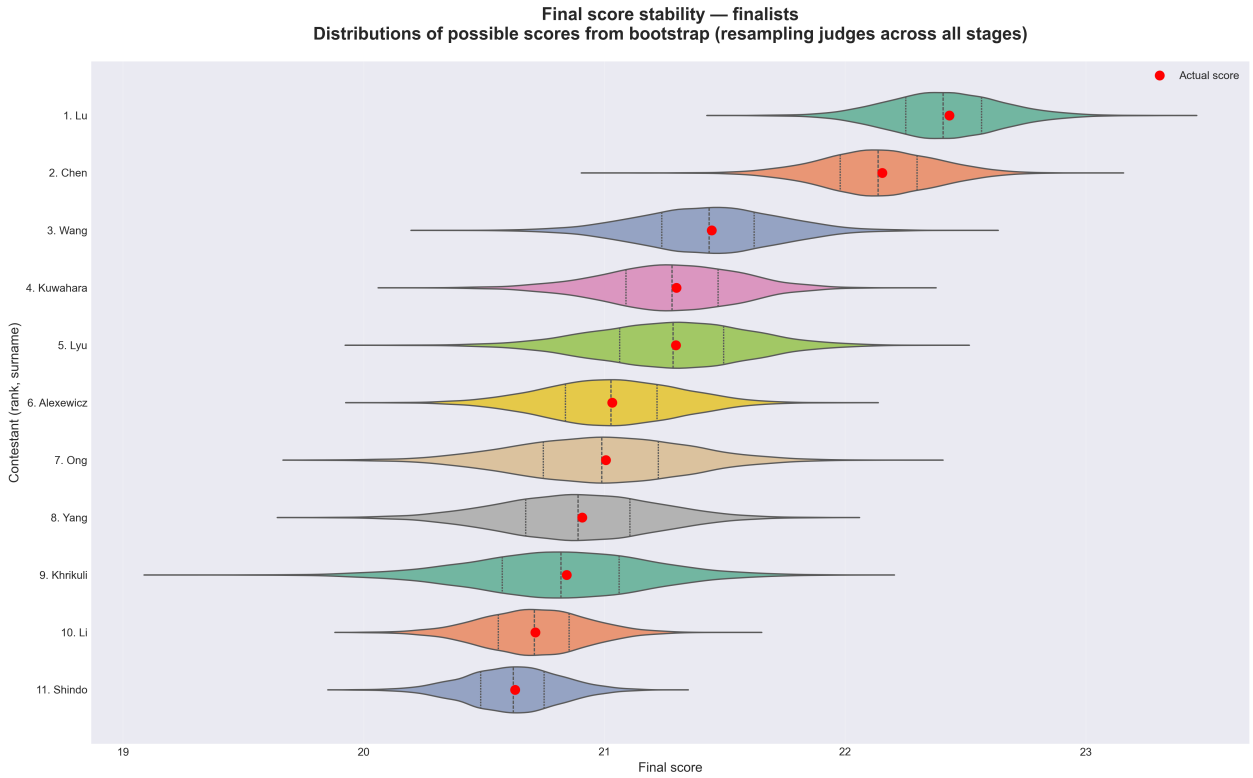


Figure 21: Distributions of possible scores from the bootstrap; red dots = actual results.

0.75

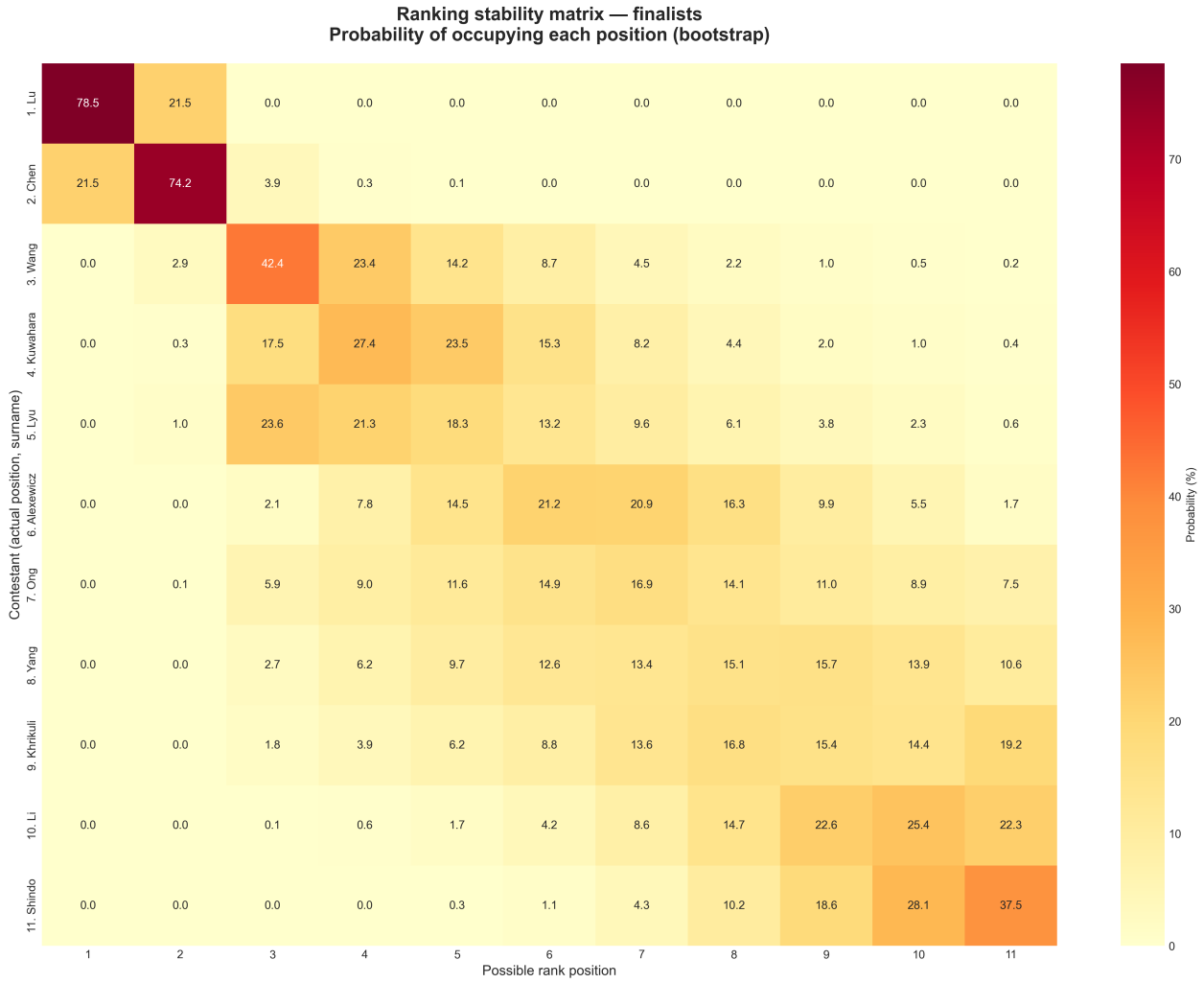


Figure 22: Ranking-stability matrix: probabilities of occupying each position (finalists).

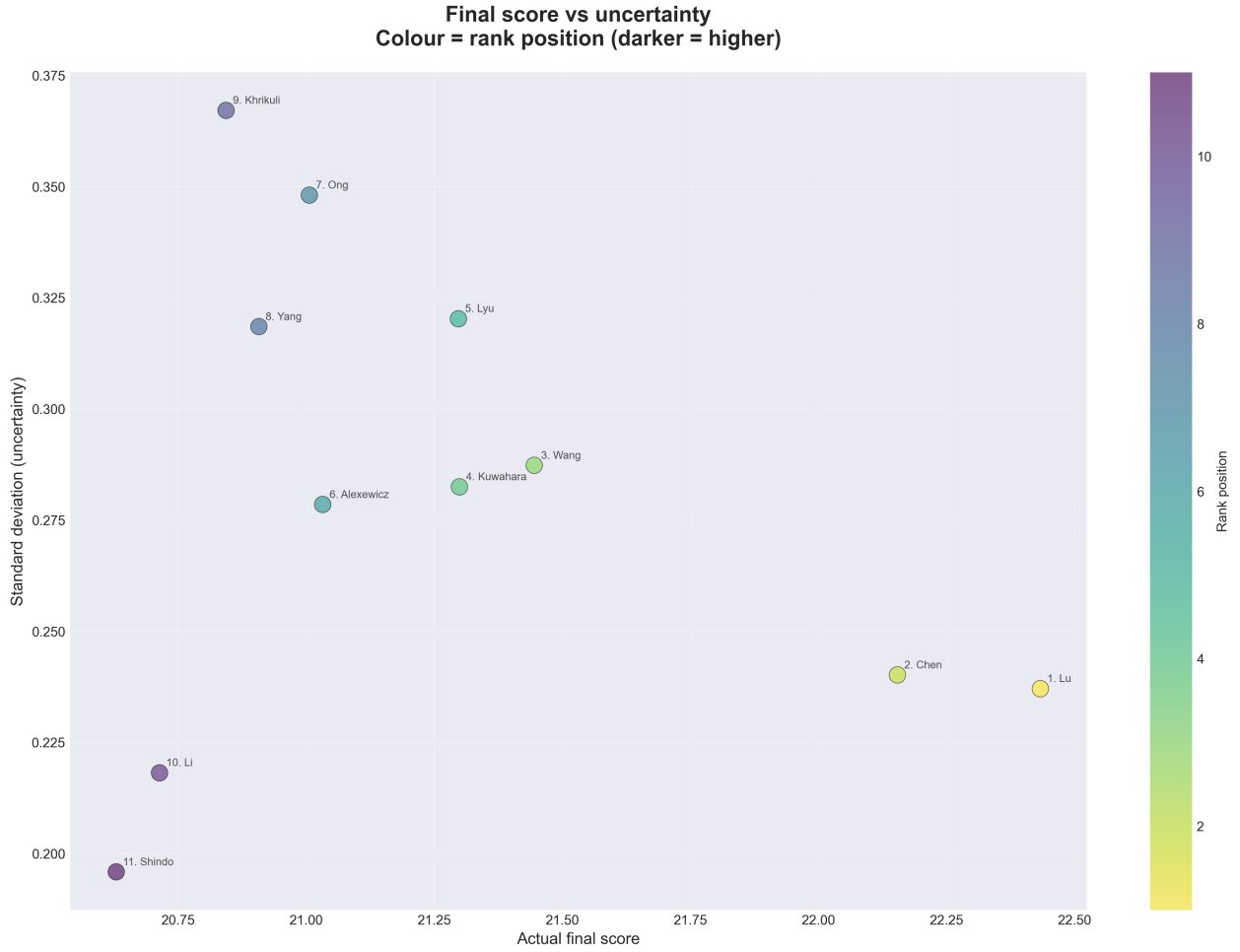


Figure 23: Final score vs uncertainty (bootstrap SD); colour = rank position.

Jury-composition bootstrap[3] shows how (un)stable finalists' placements are under alternative random juries. Violin plots indicate the top places are most stable, while the mid-table (≈3–9) shows wider, overlapping distributions—sub-1-point differences between neighbors are not decisive.

The position-stability matrix quantifies probabilities (diagonal = "holding" a place), and the score–uncertainty relation confirms that extremes (top/bottom) are more robust.

3 **Bootstrap of jury composition** — Sensitivity to the *choice of judges*. For each stage s with original judge set J_s (size $|J_s|$), B=10,000 replications are generated as follows:

1. For each stage s, sample *with replacement* a new judge set $J_s^{(b)}$ of size $|J_s|$ from J_s.
2. From the $J_s^{(b)}$ scores, final results are computed to obtain the final score $Y_i^{(b)}$ and rank $R_i^{(b)}$ for each contestant i.

Medians and 95% CIs (2.5–97.5%) are derived from $\{Y_i^{(b)}\}_{b=1}^{B}$; the probability matrix of ranks is built from $\{R_i^{(b)}\}$. This answers: *if the jury were another sample from the same population of judges, how much would results change?*

16 Stability of finalists' results (perturbation)

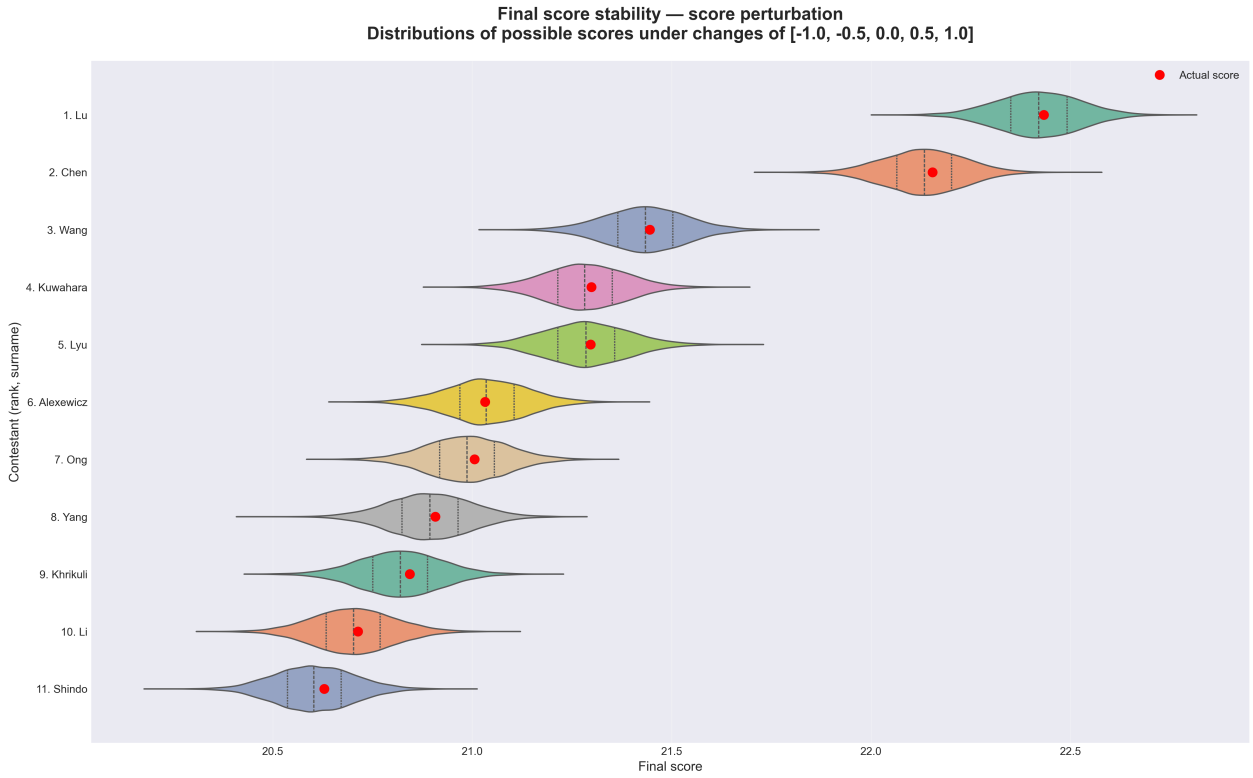


Figure 24: Distributions of possible results from score perturbations; red dots = actual results.

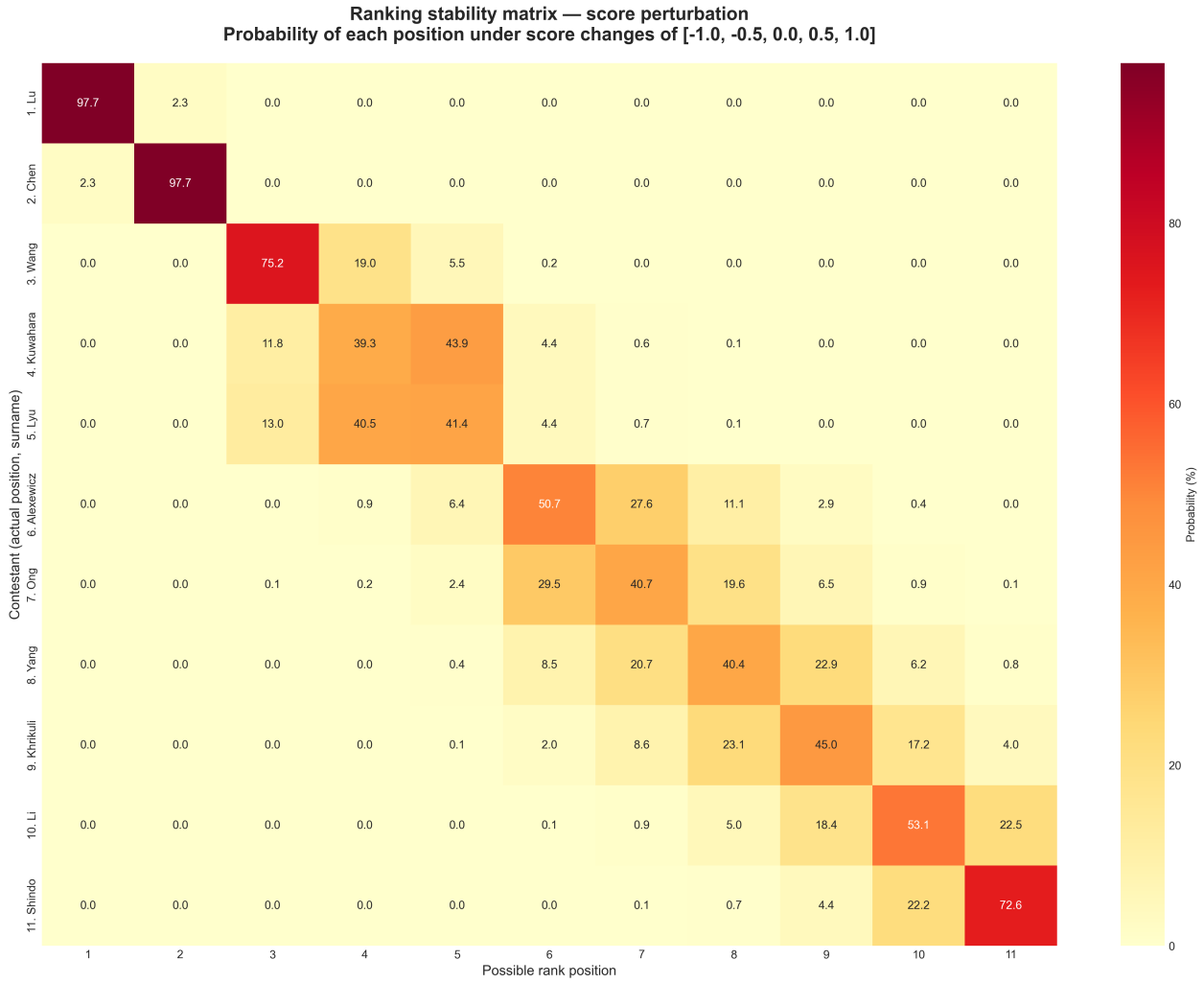


Figure 25: Probability matrix of final positions under per-score perturbations in $[-1, 1]$.

The **score-perturbation** test[4] shows how outcomes would change if judges had awarded slightly different points. Red dots (actual results) typically align with medians, indicating the observation was not a "lucky draw." Top places (1–2) have narrow, right-shifted distributions—*robust* even with ± 0.5–1.0 changes. The middle ($\approx$4–9) is less stable: small shifts can move contestants by 2–3 ranks; the very bottom is also relatively stable due to separation.

4 **Score perturbation** — Local uncertainty with fixed jury. For each raw score x_{ijs} (contestant i, judge j, stage s) an independent $\varepsilon \in \{-1.0, -0.5, 0, +0.5, +1.0\}$ (equal probabilities, $\mathbb{E}[\varepsilon] = 0$) is added, then the score is *clipped* to the 1–25 scale:

$$x'_{ijs} = \mathrm{clip}\big(x_{ijs} + \varepsilon,\ 1,\ 25\big).$$

All results are recalculated exactly as in the official pipeline, and the procedure is repeated B=10,000 times to obtain $\{Y_i^{(b)}\}$ and $\{R_i^{(b)}\}$.

Summary

Overall picture

The comparisons of score distributions, correlations between jurors, PCA of profiles, and stability tests lead to a consistent conclusion: **no patterns of conscious action "for" or "against" specific pianists are visible**. Extreme scores are well explained by the *individual calibration* and *diversity of preferences* among the jurors. These are differences in taste, not coalitions or systemic biases.

The top tier

- **Eric Lu** won **indisputably**. In all normalization variants, in the bootstrap, and with score perturbations, his result and position remain exceptionally stable; the distributions are narrow and shifted relative to the rest.
- **Kevin Chen maintains second place** with a large margin of safety. Stability tests do not undermine this position.

Who sparked the greatest discrepancies?

- Among the finalists, **Vincent Ong** and **David Khirkuli** provoked the **most polarized reactions** among the judges (high standard deviations, wide score ranges, presence of outliers on both sides).

Juror style

- Analysis of tendencies shows real differences in *calibration* (more lenient vs. more strict) and *scatter* (frequency of using wide scale fluctuations).
- **Krzysztof Jabłoński** was the **strictest** (on average lower relative to the consensus), while **Sa Chen** was the **most lenient** (on average higher).
- The most individual judges (scores inconsistent with the consensus) were **Yulianna Avdeeva**, **Sa Chen**, and **Garrick Ohlsson**, despite using extremely different score ranges.
- **No jury alliances were detected**: correlations are moderate, and the dendrogram does not indicate stable, repeatable blocks.

Normalization

- The application of one of the three normalization methods studied would have resulted in the laureate **Vincent Ong** being excluded from the final. The remaining results would not have changed significantly.

Excluding a single juror

- **Every juror had a large impact on the competition's fate**: the absence of any single judge would have affected qualification for the next round for at least 1 pianist, and a maximum of 9.

- Among the finalists, **Vincent Ong** would have been missing if one of as many as nine judges had been absent, and **David Khirkuli** if one of eight judges had been absent. The absence of one of six judges would have prevented **Piotr Alexewicz** from reaching the final. If one of three judges had been missing from the jury, **William Yang** would have finished the competition in the 3rd round.
- **Eric Lu** and **Kevin Chen** would have remained in first and second place, respectively, if any single judge had been removed from the jury.

The scoring system

In the author's opinion, the **system is objective and effective**: top results are **clear and robust**, while mid-table positions show **limited sensitivity** to minor fluctuations.