A Modular Framework for Bayesian Player and Team Rating in Professional *Valorant*

SABR Rating Project

Technical report, 2026-04

April 2026

Abstract

We describe SABR, a four-layer modular framework for rating players and teams in professional *Valorant*. Team ratings come from a per-match Glicko-1 system with a region-strength affine shrinkage; these ratings serve as the opponent-strength control for a per-stat hierarchical Bayesian regression (SIDO v3) fit over five per-round statistics. Fixed-anchor principal components of the player-level residual matrix yield two orthogonal rankings — a PC1 quality axis and a PC2 utility axis — each displayed on a common 0–100 scale. A per-map residual layer combines the same posterior components with an empirical-Bayes shrinkage step to produce per-match ratings with calibrated posterior intervals. Each layer communicates with the next through a narrow interface so that upgrades (Glicko-1 to Glicko-2, additional stats, joint multivariate likelihoods, series-level aggregation) can proceed without touching downstream code. The paper documents the shipped methodology, the key calibration decisions, and the validation bounds, including a variance decomposition showing that per-map utility signal is effectively unmeasurable with the current stat basis and must therefore be shipped as a season-level product only. All numbers are reported against a 16-window rolling panel covering 2024-11 through 2026-04.

1 Introduction

Professional *Valorant* has no standard, opponent-adjusted player rating system that separates individual contribution from team context while also supporting defensible per-map attribution. Box-score metrics (HLTV Rating 2.0 (HLTV.org, 2017)) are not opponent-adjusted; team-level Elo/Glicko systems say nothing about players; hierarchical Bayesian player models in sibling esports (Kinnell, 2024) give per-player coefficients but not calibrated per-match scores.

This paper documents **SABR**, a framework that yields three shipped products on a common 0–100 display scale:

1. a per-player *quality* ranking (PC1) and a parallel *utility* ranking (PC2);
2. per-match player ratings with Bayesian posterior intervals; and
3. per-match team ratings, both result-based and performance-based.

The central design claim is **modularity**: the framework is a four-layer stack in which each layer consumes the previous layer's output through a narrow interface and produces an interface for the next. Layers can be swapped in place — Glicko-1 can become Glicko-2, the stat basis can grow, the plug-in PCA composition can become a joint multivariate model — without downstream rewrites.

Section 2 makes the architecture explicit; Sections 3–7 work through the layers; Section 8 reports the validation and limitations; Section 9 sketches upgrade paths.

Scope. We do not treat match-outcome prediction as a headline: a parallel product (composite-v4) performs team-level win prediction and is reported elsewhere. The player and team ratings described here are descriptive — they say where a player sits on the skill distribution at a given horizon, and how their contribution on a given map departs from that baseline.

2 Architecture overview

Figure 1 shows the four layers and the interfaces between them. Each layer is a self-contained statistical model whose output schema is stable: downstream code never reaches back through a layer.

1. **Team-strength layer.** Per-match region-adjusted Glicko-1. Consumes a match stream; emits a pre-match rating snapshot per (team, match) pair.
2. **Player-stat layer.** Independent hierarchical Bayesian regressions over five per-round statistics (ADR, KAST, KPR, FKPR, APR), using the team ratings from layer 1 as opponent-strength and own-strength controls. Emits per-player per-stat posterior residuals $b_s^{\text{player}}(p, w)$ on rolling 3-month windows.
3. **Ranking layer.** Fixed-anchor PCA on the stacked residual matrix. Emits a PC1 quality score and an orthogonal PC2 utility score per player per window, rescaled to 0–100.
4. **Per-match residual layer.** Per-map residuals in natural stat space, composed through the fixed PC loadings, subjected to empirical-Bayes shrinkage, and rescaled to 0–100.

The interface between layers 1 and 2 is the column `opponent_rating_adj` in the per-map features file; between layers 2 and 3, the per-(player, stat, window) posterior means written to `sido_v3_multistat_bplayer.csv`; between layers 3 and 4, the fixed loadings file `pca_anchor.json`. Swapping any one layer requires only that the new implementation produces the same output schema.

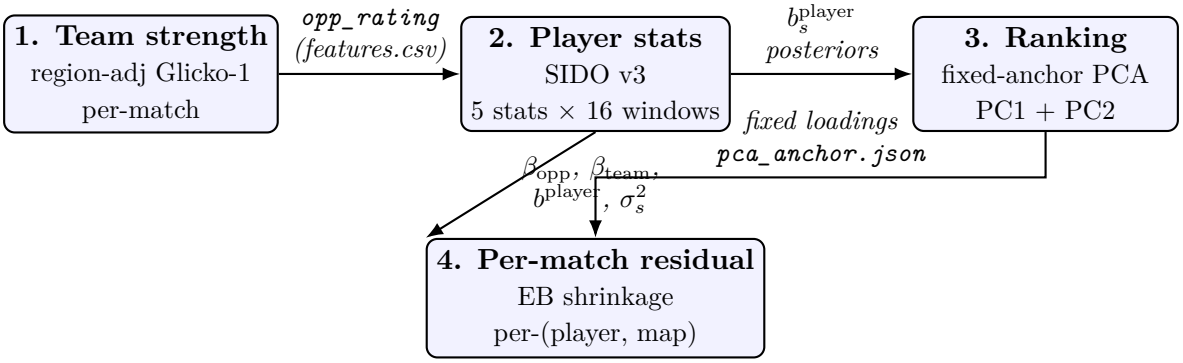

Figure 1: Four-layer architecture. Arrows label the narrow interfaces between layers; downstream code does not reach past its immediate upstream.

Notation. Throughout, p indexes a player, m a (team, map) row, w a rolling 3-month window, $s \in \{\text{ADR}, \text{KAST}, \text{KPR}, \text{FKPR}, \text{APR}\}$ a stat, and r_m the rounds played on map m.

3 Team ratings: region-adjusted Glicko-1

3.1 Per-match cadence and snapshot semantics

The team layer runs Glicko-1 (Glickman, 1995) with a rating period of one match. Each match updates both participating teams immediately; ratings for downstream consumers are always the *pre-match* snapshot. This avoids the common leak where a team's post-match rating (which already incorporates the match outcome) is visible to models predicting that match, and makes the team-rating column on the per-map feature table usable as an opponent-strength covariate without further adjustment.

We retain Glicko-1 rather than Glicko-2 (Glickman, 2012) because the volatility parameter σ is poorly identified at team-match volumes typical of *Valorant* pro play (teams have $\mathcal{O}(50\text{–}200)$ matches per season). Roster changes, the main source of abrupt team strength shifts, are handled as an observable event: the rating deviation RD is inflated when the trailing roster-turnover threshold is exceeded, which plays the role σ would in a Glicko-2 fit.

3.2 Region-strength affine shrinkage

Pro-scene regions play in mostly-closed circuits; raw Glicko ratings drift apart across regions in proportion to local pool strength. We shrink each team's raw rating toward its regional mean:

$$r_t^{\text{adj}} \;=\; \mu_{R(t)} \;+\; \beta\,\big(r_t^{\text{raw}} - \mu_{R(t)}\big), \qquad \beta = 0.72, \tag{1}$$

where $R(t)$ is team t's region and $\mu_{R(t)}$ its mean rating at the current snapshot. The coefficient β is fit to maximise cross-region match-win accuracy on a held-out split. Moving from a monthly rating period (**61.82%** match-win accuracy, cross-region test split) to the per-match cadence described above (**64.33%**, same split) captures the majority of the remaining predictive lift, and is the configuration downstream layers consume. Figure 2 shows both levers visually: panel (a) is the β-grid on the training set (Brier minimises at $\beta = 0.72$; accuracy is nearly flat in β, consistent with Brier being the more sensitive calibration signal), and panel (b) compares test-set Brier across monthly vs. per-match cadence at the raw ($\beta = 1$) and tuned ($\beta = 0.72$) settings. Both the region-shrinkage and the cadence upgrade reduce Brier independently; their combination takes test Brier from 0.2328 to 0.2202 (-5.4%).

3.3 Display transform

For publication we rescale the active pool (teams with ≥ 1 match in the trailing 90 days) to a 0–100 scale:

$$s_t \;=\; \text{clip}\left(50 + 15 \cdot \frac{r_t^{\text{adj}} - \bar{r}_{\leq 90}^{\text{adj}}}{\text{sd}(r_{\leq 90}^{\text{adj}})},\; 0,\; 100\right). \tag{2}$$

Figure 3 shows the raw, adjusted, and 0–100 distributions. The regional shrinkage tightens the global spread and centres active teams near the design mean of 50; the remaining right skew reflects a small elite subpopulation at the tournament front-runners.

4 Player stat-level model: SIDO v3

4.1 Per-stat generative model

For each stat s and each window w we fit an independent Bayesian regression on the n (player, map) observations in that window. Denote by $y_{s,m}$ the observed value of stat s on observation

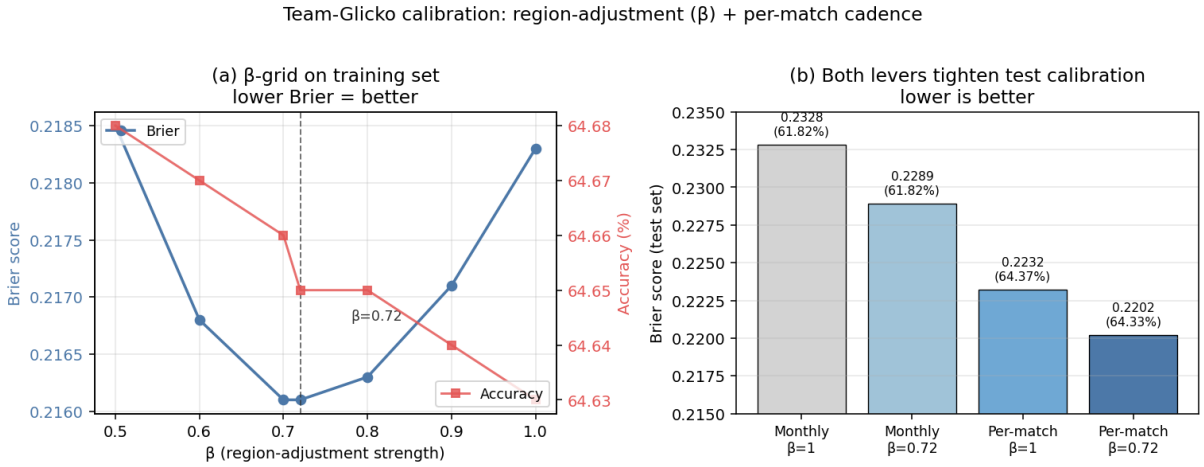

Figure 2: Team-Glicko calibration. (a) Training-set β-grid: Brier (left axis, blue) is minimised at $\beta = 0.72$; accuracy (right axis, red) is nearly constant in β. (b) Test-set Brier across the two configuration levers at the tuned $\beta = 0.72$: region-shrinkage and per-match cadence each lower Brier independently and compound to 0.2202 (per-match, tuned) from 0.2328 (monthly, $\beta = 1$); accuracy in parentheses on each bar.

m (map, player). SIDO v3 follows Kinnell (2024): a stat-specific likelihood with a shared linear predictor

$$\mu_{s,m} = \alpha_s + \beta_s^{\text{opp}} z_m^{\text{opp}} + \beta_s^{\text{team}} z_m^{\text{team}} + b^{\text{player}}_{s,\,p(m)} + b^{\text{agent}}_{s,\,a(m)} + b^{\text{map}}_{s,\,g(m)}, \tag{3}$$

where $z_m^{\text{opp}} = (r^{\text{adj}}_{\text{opp},m} - 1500)/400$ and z_m^{team} is the analogous own-team scaling; $p(m)$, $a(m)$, and $g(m)$ are the player, agent, and map identifiers for observation m. The hierarchical random effects b use a non-centred parameterisation with HalfNormal(0.3) priors on their scales.

Likelihoods:

- **ADR**: log-Normal on $\log y$.
- **KAST**: Beta on $y/100$ (Beta mean reparameterisation with $\mu = \sigma(\cdot)$, $\kappa \sim$ HalfNormal).
- **KPR, FKPR, APR**: log-Normal on $\log(y + 1/r_m)$ (Laplace smoothing of zero-counts at a rate of one event per map).

The Laplace offset keeps zero-count rows on the support of the log-Normal while contributing a negligible shift at typical rates. All fits use PyMC (Abril-Pla et al., 2023) with NUTS (4 chains $\times$ (1500 warmup + 1000 draws), `target_accept`=0.99). Across the 16 rolling 3-month windows $\times$ 5 stats = 80 fits, 80 converged clean on first attempt ($\hat{R} < 1.02$, $\text{ESS}_{\text{min}} > 290$).

4.2 Opponent-adjusted player coefficients

The posterior mean of $b^{\text{player}}_{s,p}$ is SIDO's central output: an opponent- and own-team-adjusted residual stat-skill coefficient for player p in window w. Because z^{opp} and z^{team} are fit jointly, the residual is a *double-sided* opponent adjustment: it controls for both the strength of the team being faced and the strength of the teammates providing the match context. Downstream per-match ratings inherit this opponent-adjustment without needing their own regression.

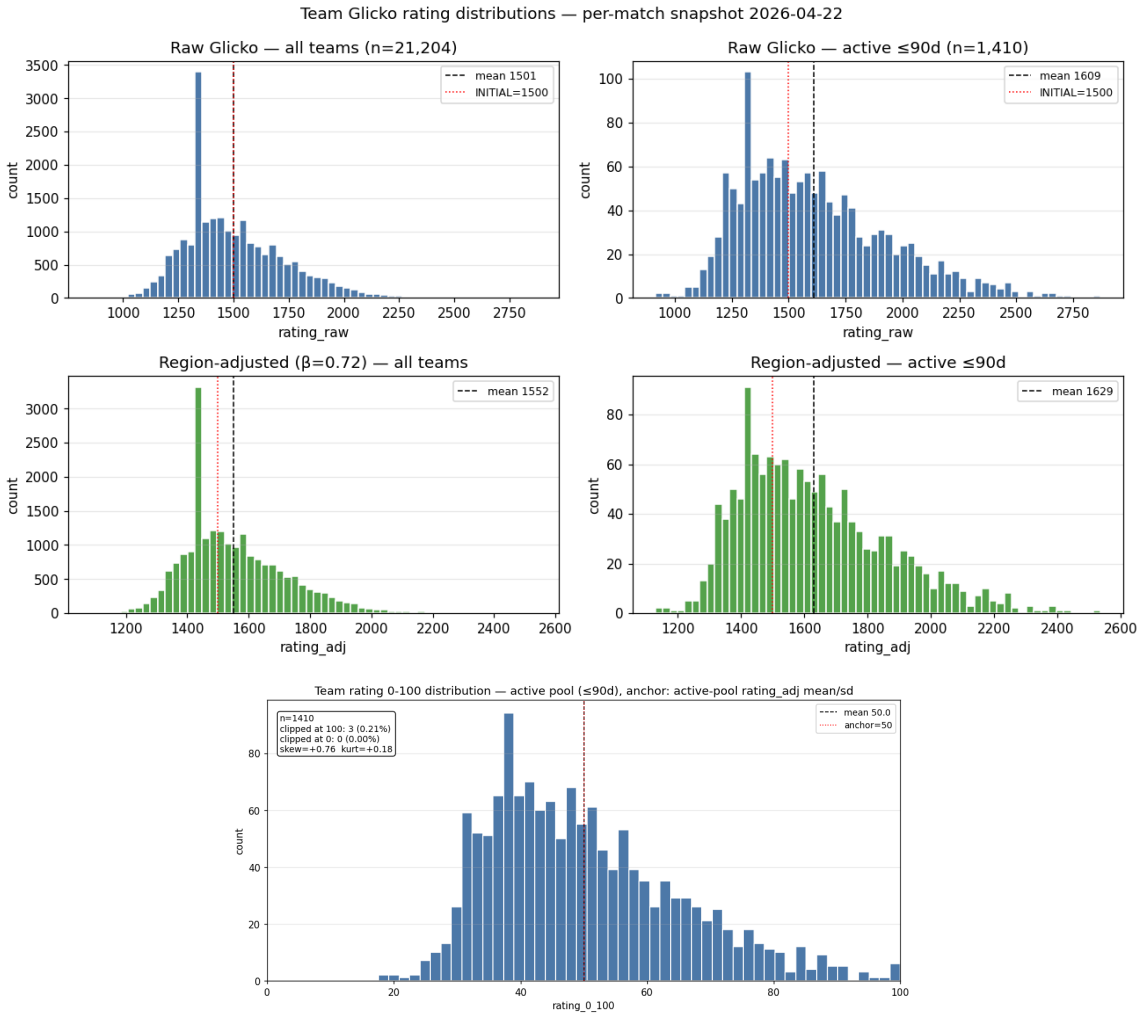

Figure 3: Team rating distributions. Top: raw Glicko and region-adjusted Glicko for all teams (left panels, $n = 21{,}204$) and the active pool within 90 days (right panels, $n = 1{,}410$). Bottom: the 0–100 display transform centred on the active pool, mild right skew (skew = +0.76, kurt = +0.18).

4.3 Validation: forecast stability and external cross-check

We evaluate each stat's ratings with a 1-month-lag forecast: for each adjacent window pair (w_t, w_{t+1}) we Spearman-correlate $b_s^{\text{player}}(p, w_t)$ with the window-w_{t+1} per-player residual. Mean over 15 pairs: $\rho_{\text{ADR}} = +0.810$, $\rho_{\text{KAST}} = +0.770$. As an external check at the anchor window, we compare the PC1 composite (Section 5) against VLR's R2.0 box-score rating on the 86-player pool intersection: Spearman $\rho = +0.74$. This is high without being mechanical: PC1 uses opponent-adjusted residuals from five stats, R2.0 uses a public fragging-weighted formula, and the two nonetheless agree on player ordering at the top end.

5 Player rankings: fixed-anchor PCA

5.1 Anchor-window eigenstructure

Each window yields a $(n_w \times 5)$ matrix of per-player posterior residuals $b_{s,p}^{\text{player}}(w)$. We z-score each column against the anchor-window pool $(\bar{b}_s, \text{sd}_s)$ and run a standard PCA (Jolliffe, 2002) *at the anchor window only*; the resulting loadings are then applied unchanged to every other window. The anchor is the most recent complete 3-month window (2026-01-31 through 2026-04-30), $n = 121$ players with ≥ 10 maps.

PC1 explains 49.95% of residual variance; PC2 adds 24.24% for a cumulative 74.2%. Figure 4 shows the loadings. PC1's loadings are all positive with ADR and KPR carrying the heaviest weight; reading it as *overall quality* is unambiguous. PC2 loads positively on APR and KAST and negatively on FKPR and KPR with ADR near zero — a clean *utility-vs-fragging* axis: players whose contribution is assist- and survival-heavy score high, players whose contribution is first-kill-heavy score low.

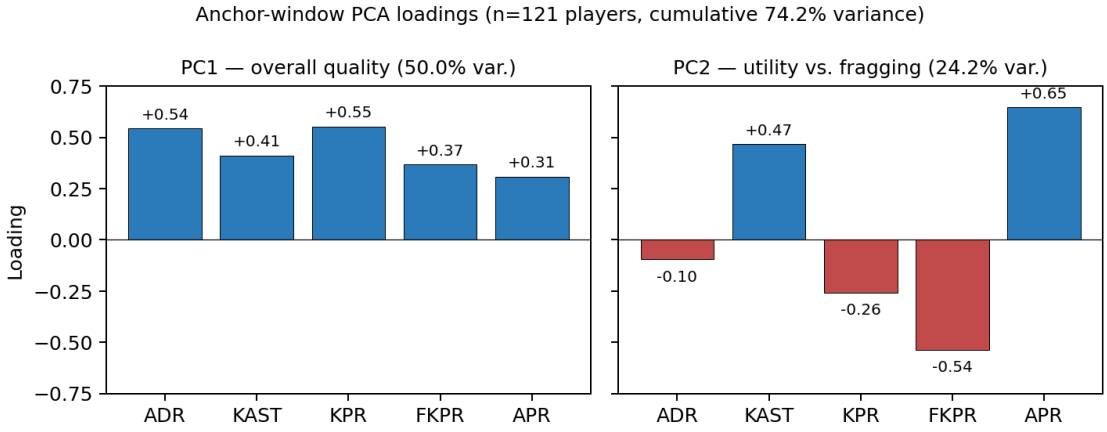

Figure 4: Fixed anchor-window PCA loadings. PC1 (left) loads positively on all five stats and admits an "overall quality" reading. PC2 (right) loads positively on APR and KAST, negatively on FKPR and KPR, with ADR ≈ 0 — a utility-vs-fragging axis orthogonal to PC1 by construction.

5.2 Two products, not one composite

Because PC1 and PC2 are orthogonal by construction (empirical Pearson $r = -2.75 \times 10^{-12}$ on the anchor pool), a weighted composite of the two would erase a genuine dimension of player value. We ship them as *separate* rankings. The four quadrants this produces are populated in practice:

derke (PC1 = 90.5, PC2 = 98.0) is the complete package; *zmjjkk* (PC1 = 79.7, PC2 = 5.7) is a pure entry fragger; *munchkin* (PC1 = 58.6, PC2 = 100.0) is a utility specialist with moderate fragging. A win-prediction-weighted composite, which we also computed as a baseline, collapses onto a 59%-ADR weight and loses the utility axis entirely, illustrating why variance-based (PC) weights are the right primitive here.

5.3 Anchor normalisation: drift and scale

A naive per-window z-score (v3) makes cross-window ranks comparable but not cross-window scores. SIDO v3.2 is a post-processing pass that upgrades the ranking from rank-space to absolute-scale trajectories. It decomposes into two effects, both taken *from the anchor window* rather than the current window:

Drift. For each pair of adjacent windows $(t, t+1)$, compute $\delta_s(t{\to}t{+}1) = \frac{1}{|A|}\sum_{p\in A}\big(b_s(p,t{+}1) - b_s(p,t)\big)$ in natural space (log-rate for rate stats, logit for KAST) over an anchor set A of 12 stable players (3 per role). Accumulate $\Delta_s(t) = \sum_{t'\le t}\delta_s(t')$ outward from the anchor and subtract $\Delta_s(t)$ from every player's b_s at window t. This is a *pool-wide additive shift* that does not change relative ranks within a window.

Scale. After drift adjustment, z-score against the *anchor-window* pool $(\bar{b}_s, \mathrm{sd}_s)$, not the current window's. Anchor sd_s differs from typical per-window sd (anchor/mean-pool ratios: ADR 0.98, KAST 1.07, KPR 0.84, FKPR 0.80, APR 0.94) — windows with wider per-pool variance get less compressed. This *does* change relative ranks within non-anchor windows, intentionally.

Rank agreement at the anchor window itself is $\rho = 1.0000$ by construction (v3 and v3.2 share the same PCA loadings and the same reference statistics at that window). Off-anchor, drift and scale both contribute; the largest single-stat correction in the current panel is a $+0.55 \to +0.01$ z-shift on FKPR for *aspas* in the 2025-Q2 window where FKPR pool drift peaked, consistent with the "small-pool inflation" artefact that motivated the upgrade.

5.4 Career trajectories

Figure 5 plots PC1 for six notable players over the 16 anchor-normalised windows. Because v3.2 places all windows on a common reference scale, vertical shifts between points are absolute-skill statements rather than rank-space artefacts. A few arcs illustrate the utility of the product: *marteen*'s late-2025 surge from a near-anchor-median player to the top of the anchor-window pool; *derke*'s sustained top-tier stability across nearly all windows; *aspas*'s mid-2025 peak followed by a correction that the v3.2 drift adjustment partially attributes to pool-wide inflation rather than individual decline; and *kicks*'s multi-window trough consistent with a rebuild period. These plots match known public narratives about each player's recent competitive history, providing a qualitative face-validity check on the anchor-normalisation output.

6 Per-match player ratings (PMR)

Given the layer-2 posteriors, producing a per-map rating is mechanically straightforward — but the calibration of the *display scale* is not, and it is where the methodological weight of the product sits. We document three successive versions (v1, v1.1, v1.2), because each represents a distinct class of calibration decision.

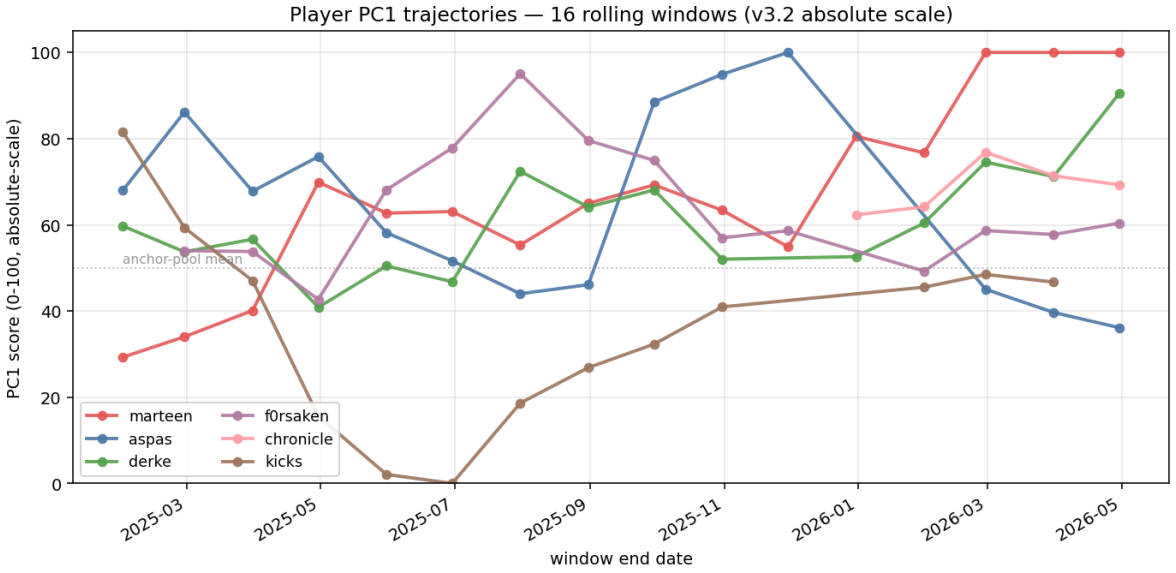

Figure 5: PC1 trajectories over 16 rolling 3-month windows for six players with substantial match history. Horizontal dotted line marks the anchor-pool mean of 50. Cross-window comparisons are on a common absolute scale after the v3.2 drift and scale corrections.

6.1 Residuals and composition

For each (player, map) row we compute the observed-minus-expected residual in each stat's natural space:

$$\varepsilon_{\mathrm{ADR},m} = \log y_m - \mu_{\mathrm{ADR},m}, \tag{4}$$

$$\varepsilon_{\mathrm{KAST},m} = (y_m - \mu_{\mathrm{KAST},m}) / \big(\mu_{\mathrm{KAST},m}(1-\mu_{\mathrm{KAST},m})\big), \tag{5}$$

$$\varepsilon_{s,m} = \log(y_m + 1/r_m) - \mu_{s,m} \qquad (s \in \{\mathrm{KPR}, \mathrm{FKPR}, \mathrm{APR}\}). \tag{6}$$

The KAST form is a delta-method linearisation of the logit residual that removes a +0.087 Jensen bias at the pool's typical (μ, κ). Residuals are z-normed against the anchor-pool cross-player sd_s of b_s^{player} and composed via the fixed PC loadings of Section 5:

$$\mathrm{PC1}_m^{\mathrm{res}} = \sum_s L_s^{\mathrm{PC1}} \cdot \frac{\varepsilon_{s,m}}{\mathrm{sd}_s}, \qquad \mathrm{PC2}_m^{\mathrm{res}} \text{ analogous.} \tag{7}$$

6.2 v1: unit-consistent-but-wrong scale

Version 1 chose the 0–100 display slope to preserve additivity with the window ranking: a 15-point match rating above overall rating corresponds to one anchor-pool sd of PC1 ($\mathrm{sd}_{\mathrm{PC1}} = 1.58$). The consequence was catastrophic: per-map residual SD is roughly 5.44× larger than cross-player SD, so routine noisy maps registered as ±3 cross-player sigma events. **55.2%** of rows clipped at 0 or 100 (Figure 6, left).

6.3 v1.1: empirical σ_{match} calibration

Version 1.1 swapped the denominator for the empirical per-match residual standard deviation, $\sigma_{\mathrm{match,PC1}} = \mathrm{SD}(\mathrm{PC1}^{\mathrm{res}}) = 8.598$ on the 18,013-row panel. By construction this gives SD(match −

overall) = 15 and an approximately $\mathcal{N}(\text{overall}, 15)$ per-map distribution. Clipping dropped to **4.6%**. The unit-consistency with the ranking's "15 points = 1 anchor-pool sd" rule was lost; the scale interpretation became "15 points = 1 per-match sd" instead, which is the right primitive for a per-match product.

6.4 v1.2: empirical-Bayes shrinkage

v1.1 still treated every residual with the same step size, independent of the round-count evidence behind it. Short maps carry more sampling noise per observation than long maps, so their residuals should be pulled harder toward the prior. v1.2 formalises this as an empirical-Bayes posterior mean (Efron and Morris, 1975; Gelman et al., 2014).

Decompose the observed residual variance as

$$\sigma^2_{\text{match}} \;=\; \sigma^2_{\text{true}} \;+\; \mathbb{E}\big[\sigma^2_{\text{samp}}(r)\big], \tag{8}$$

where $\sigma^2_{\text{samp}}(r)$ is the sampling variance of PC^{res}_m propagated through the fixed loadings from stat-level sampling variances (NegBin-delta for rate stats, Beta-Binomial for KAST, log-Normal residual for ADR). The signal variance σ^2_{true} is a global constant (or per-role scalar for the role variant) fit once. The posterior mean of true per-map performance is then

$$\tilde{y}_m \;=\; \lambda(r_m)\,\text{PC}^{\text{res}}_m, \qquad \lambda(r_m) \;=\; \frac{\sigma^2_{\text{true}}}{\sigma^2_{\text{true}} + \sigma^2_{\text{samp}}(r_m)}. \tag{9}$$

The display rating keeps v1.1's denominator so that "15 points $\approx$ 1 per-match sd" reads as before, but with a narrower effective scale:

$$\text{match\_pc1\_100}_m \;=\; \text{overall\_pc1\_100}_{p(m)} \;+\; 15 \cdot \frac{\tilde{y}_m}{\sigma_{\text{match,PC1}}}, \tag{10}$$

with posterior SD $15\sqrt{\lambda(r_m)\sigma^2_{\text{samp}}(r_m)}/\sigma_{\text{match,PC1}}$, tighter than v1.1's by factor $\sqrt{\lambda}$. Over the global panel, SD(match − overall) drops from 15.00 (v1.1) to **8.84**; clipping drops to **2.1%**; mean CI width contracts by 22.8%.

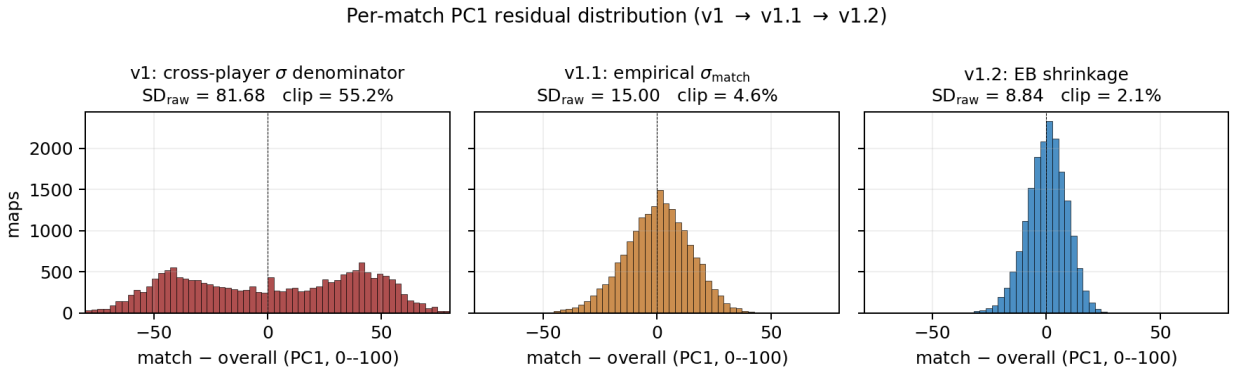

Figure 6: Per-match PC1 residual distribution (match − overall) through the calibration arc. The v1 panel is dominated by the 0/100 clip pileup; v1.1 recovers an approximately normal shape; v1.2 tightens it further via empirical-Bayes shrinkage. Raw SD (pre-clip) drops 81.7 → 15.00 → 8.84; clipping drops 55.2% → 4.6% → 2.1%.

6.5 Variance decomposition: why PC2 cannot be per-map

The same decomposition surfaces a structural finding about the PC2 axis. On PC1, $\sigma^2_{\text{match}} = 73.9$, mean $\sigma^2_{\text{samp}} = 29.9$, so $\sigma^2_{\text{true}} = 44.0$ — **60% signal, 40% sampling noise**, mean $\lambda_{\text{PC1}} \approx 0.60$. On PC2, $\sigma^2_{\text{match}} = 38.03$, mean $\sigma^2_{\text{samp}} = 37.59$, so $\sigma^2_{\text{true}} = 0.44$ — **99% sampling noise**, mean $\lambda_{\text{PC2}} \approx 0.01$. Per-map PC2 collapses to per-player overall PC2 under shrinkage (Figure 7, right panel). The physics is that PC2's heaviest loadings are on APR (Poisson on a 13-30 trial count with rate ~ 0.3) and on KAST (Beta-Binomial on the same trial count); their per-observation variance is too large for per-map PC2 composition to leave signal after the weighted sum. Utility is a *style* over many maps, not a per-map *behaviour*. This is the key justification for shipping PC2 only as a season-length ranking and refusing a per-match PC2 scoreboard.

The λ_{PC1} range 0.56 (at $r = 13$) to 0.63 (at $r = 28$) is narrower than the pure-Poisson intuition would suggest. The reason is that ADR's log-Normal likelihood has an r-independent residual sigma, and ADR carries PC1 loading 0.544 — it clamps the r-variation of the total sampling variance. A joint multivariate v4 with per-round-derived ADR would lift this ceiling.

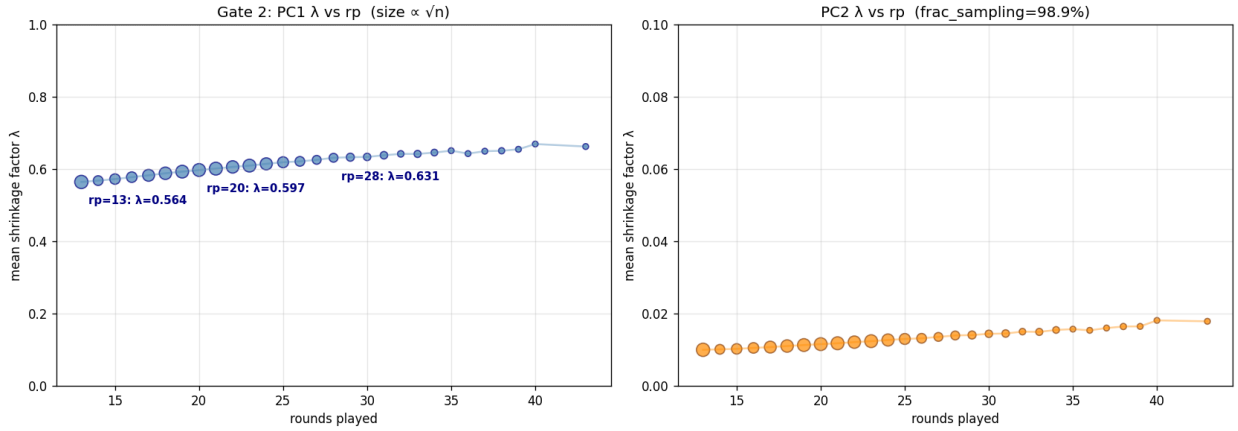

Figure 7: Shrinkage factor λ by rounds played r, PC1 (left) and PC2 (right). λ_{PC1} is monotone and ranges 0.56–0.63, bounded from above by the rp-independent ADR residual component. $\lambda_{\text{PC2}} \approx 0.01$ across all r (right panel, note expanded vertical range): the PC2 signal is essentially unrecoverable per map.

7 Per-match team ratings

A per-map team rating combines two views of team performance, each constructed directly from earlier layers.

Result-based. Fit the one-feature linear model

$$\text{margin}_m \;=\; \alpha \;+\; \beta \cdot \left(r^{\text{adj}}_{t,m} - r^{\text{adj}}_{t',m}\right) \;+\; \epsilon_m, \tag{11}$$

on 244,550 (team, map) rows (antisymmetric long-format; α is identically zero by construction, $\hat{\beta} = 0.01156$: 100 rating points $\approx$ 1.16 rounds). The z-anchored residual $(\text{margin} - \hat{\text{margin}})/\sigma_{\text{res}}$ with $\sigma_{\text{res}} = 6.07$ is rescaled by $50 + 15z$ and clipped. $R^2 = 0.118$: at BO1 map-margin level a single rating-gap feature does about what one can expect. The result-based rating is zero-sum per match

by construction; a "strategic win" is any match where the winning team scored above 50 on this rating by more than the prediction already implied.

Performance-based. Average the five starting players' shrunk PC1 PMR scores (Section 6) for that (team, map). This rating is on the same 0–100 absolute scale as the individual PMRs and is *not* zero-sum: both teams can over- or under-perform their baselines together. Coverage is SIDO-pool-limited: 3,660 rows, versus 244,550 for the result-based rating.

Divergence is informative. On the 3,660 overlap rows, Pearson $r = +0.44$. Median $|\text{result} - \text{perf}| = 9.8$; p95 $= 28.2$. Four interpretable quadrants (Figure 8): both high (players strong + team won big, 33%), both low (33%), result>50 with perf<50 (*strategic or clutch wins* where the team won bigger than the individual stats predicted, 18%), result<50 with perf>50 (*heroic losses*, 16%). The within-team player residual correlation is $\approx +0.67$ (inferred from the observed SD of the team-mean PMR against the naïve $1/\sqrt{5}$ prediction), consistent with teammate residuals covarying through shared match state — a team-performance signal has about one-third the independent-observation efficiency a naïve average would assume.

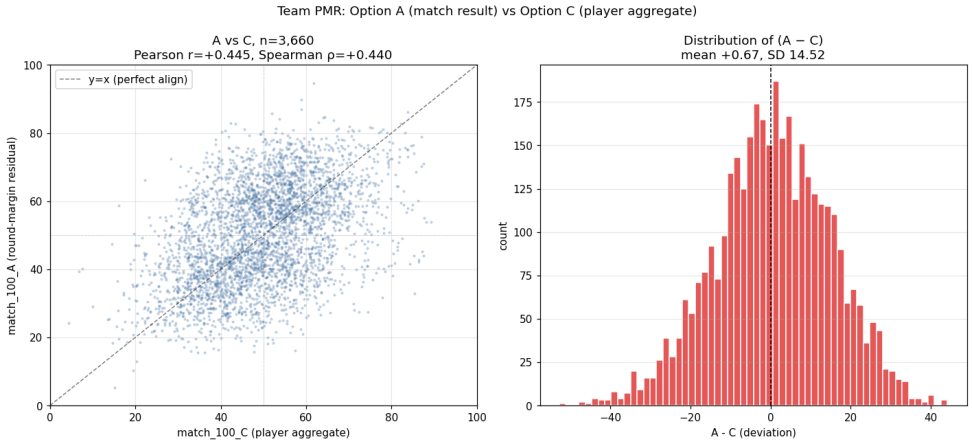

Figure 8: Team per-match ratings: result-based (horizontal) vs. performance-based (vertical) on the 3,660 overlap rows. Pearson $r = +0.44$; ∼33% of maps show substantive divergence. Upper-left quadrant is "strategic wins" (team margin beat individual stats); lower-right is "heroic losses" (individuals strong, team still lost or won small).

8 Validation and limitations

8.1 Stickiness bound on PMR

A natural question for any per-match player rating is whether the residual on map t predicts the residual on map $t+1$ within a (player, window) group — i.e., is there form/streak/fatigue signal? We fit the AR(1) regression $\varepsilon^{\text{PMR}}_{t+1} = \alpha + \rho\,\varepsilon^{\text{PMR}}_{t} + e$ pooled across all (player, window) groups with ≥ 5 maps (1,652 groups, 14,138 consecutive pairs), with cluster-robust SE on (player, window).

The aggregate estimate is $\hat{\rho}_{\text{PC1}} = +0.047$ (95% CI $[+0.031, +0.062]$) — statistically nonzero but tiny ($\rho^2 \approx 0.22\%$ of next-map variance). Decomposing by the calendar gap between consecutive maps reveals the structure (Figure 9): same-day pairs (73% of the sample, and in practice maps within the same BO3/BO5 series) have $\hat{\rho} = +0.075$; the 1–7-day and $>$ 7-day buckets are

both statistically zero or slightly negative. **PMR is descriptive per-map, not predictive cross-match.** A form-adjusted rating layer weighting recent map residuals would not propagate measurable signal to the next match. The natural product opportunity unlocked is instead a *series-level PMR aggregate* that rolls up the 2–5 maps of a BO3/BO5 into a single match-level number with the within-series correlation factored in rather than double-counted.

As a specification check, $\hat{\rho}_{\mathrm{PC2}} = -0.006$ (null), a second confirmation of the 99% sampling-noise finding via an independent angle. v1.1 (unshrunk) and v1.2 (shrunk) AR(1) coefficients differ by < 0.003 — AR(1) is scale-invariant under a near-constant multiplier.

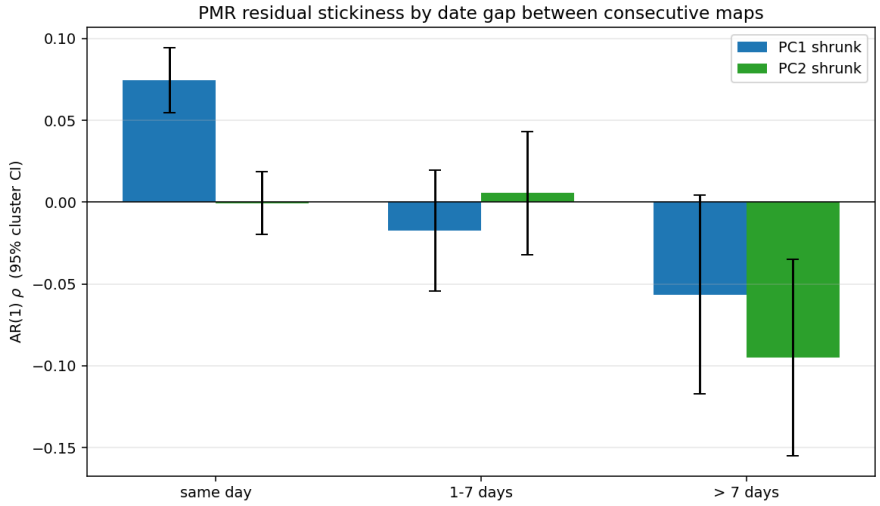

Figure 9: AR(1) serial correlation of PMR PC1 residuals within (player, window), split by calendar gap between consecutive maps. Same-day pairs — almost always within the same BO3/BO5 series — drive the aggregate. Cross-day buckets are statistically null. Error bars are 95% cluster-robust CIs.

8.2 Data coverage

The player pool is bounded by the VLR ingest scope and by the availability of per-map ADR in the feature table; the current panel averages ∼200–250 active players per window, ∼120 of whom clear the $\geq$ 10-map headline threshold. A complementary Grid round-level data source is available on ∼80 matches from 2026-02 onward. Grid ADR disagrees with VLR ADR in absolute scale (Grid/VLR = 0.79±0.07) due to different measurement semantics (Grid reports effective damage clipped to remaining HP; VLR reports raw weapon damage), but rank agreement is $r = +0.95$. The two sources are not unioned in a single fit to avoid injecting the scale offset as spurious player-skill variance; bridging is future work.

Team-PMR history is bounded by the Glicko rerun cadence (currently ending 2026-04-06 in the published CSV). Transfer-triggered RD inflation is dormant in the current snapshot due to under-coverage of the `serving_export.transfers` table; the correction would affect recently-transferred teams only and is unlikely to move the aggregate distribution.

8.3 Structural limitations

- **ADR sampling variance is r-independent.** Because ADR carries PC1 loading 0.544, this clamps the rp-range of $\sigma^2_{\text{samp,PC1}}$ and narrows λ_{PC1} to $[0.56, 0.63]$. A joint multivariate v4 that casts ADR as a per-round count-derived stat would lift this ceiling materially.
- **Per-map PC2 is unmeasurable** with the current stat basis (Section 6); utility is therefore shipped only at window-scale.
- **rounds_played is estimated, not measured.** The per-map round count is reconstructed from total map kills via an empirical calibration (median 6.844 kills/round, 95% within ± 2 rounds of ground truth on a 198-map sample). All ten players on a map share the same estimate, so the error does not contaminate cross-player comparisons on the same map.
- **log-Normal on count stats** is an approximation; a Poisson/NegBin likelihood with `rounds_played` exposure is more principled but was not required for a stable PC1.
- **Plug-in rather than joint composition.** The five stats are fit independently and then combined. Per-map residuals are positively correlated across stats, so the propagated posterior intervals slightly overstate uncertainty.
- **Clipping asymmetry.** Players whose overall rating is near the 0 or 100 boundary have asymmetric clipping in PMR by construction; use the `_raw` (unclipped) column for edge-case reads.

9 Discussion and future work

Four upgrade paths are queued, each confined to a single layer. **(i) Joint multivariate SIDO v4.** A MVN-plus-Beta likelihood over the five stats would tighten PC1 posterior intervals, propagate correctly correlated uncertainty, and (critically) allow casting ADR as a per-round-derived stat — lifting the λ_{PC1} ceiling. **(ii) Role-normalised PC2.** PC2 currently favours role archetypes whose stat profile aligns with the utility axis; a within-role normalisation would answer who is a utility *outlier* within their role. **(iii) Series-level PMR aggregate.** The within-series $\rho \approx +0.075$ finding motivates rolling up per-map residuals into a series-level score that counts the correlation rather than double-counting it. **(iv) Richer team-PMR features.** Side bias, map specialization, and home/away variables could push the result-based R^2 from 0.12 toward 0.25–0.30 at BO1 level.

Each of these lives within a single layer of the architecture of Section 2 and requires no changes to the others.

Acknowledgements

We thank the data-engineering team for the per-match feature snapshot pipeline, and the Valorant analyst community for conversations that sharpened the framing of the utility-vs-fragging axis.

References

O. Abril-Pla, V. Andreani, C. Carroll, L. Dong, C. J. Fonnesbeck, M. Kochurov, R. Kumar, J. Lao, C. C. Luhmann, O. A. Martin, M. Osthege, R. Vieira, T. Wiecki, and R. Zinkov. PyMC: A

modern and comprehensive probabilistic programming framework in Python. *PeerJ Computer Science*, 9:e1516, 2023.

B. Efron and C. Morris. Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70(350):311–319, 1975.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis*. CRC Press, 3rd edition, 2014.

M. E. Glickman. The Glicko system. *Boston University*, 1995. http://www.glicko.net/glicko/glicko.pdf.

M. E. Glickman. Example of the Glicko-2 system. *Boston University*, 2012. http://www.glicko.net/glicko/glicko2.pdf.

HLTV.org. Introducing Rating 2.0. https://www.hltv.org/news/20695/introducing-rating-20, 2017.

I. T. Jolliffe. *Principal Component Analysis*. Springer, 2nd edition, 2002.

T. Kinnell. SIDO: Hierarchical Bayesian models for opponent-adjusted player performance in competitive CS:GO. Technical report, Unpublished technical report, 2024.

A SIDO v3 prior and sampler detail

Per-stat priors (all scales in the natural-response space of the stat's likelihood):

$$\begin{aligned}
\alpha_s &\sim \text{Normal}(0,\ 1),\\
\beta_s^{\text{opp}},\ \beta_s^{\text{team}} &\sim \text{Normal}(0,\ 0.5),\\
\tau_s^{\text{player}},\ \tau_s^{\text{agent}},\ \tau_s^{\text{map}} &\sim \text{HalfNormal}(0.3),\\
b_s^{\text{player}},\ b_s^{\text{agent}},\ b_s^{\text{map}} &\sim \text{Normal}(0,\ \tau),\\
\sigma_s^{\text{obs}} &\sim \text{HalfNormal}(1) \quad \text{(log-Normal stats)},\\
\kappa_{\text{KAST}} &\sim \text{HalfNormal}(20).
\end{aligned}$$

All hierarchical effects use a non-centred parameterisation ($b = \tau \cdot \tilde{b}$, $\tilde{b} \sim \text{Normal}(0, 1)$). Sampler: NUTS in PyMC 5 (Abril-Pla et al., 2023) with 4 chains × (1500 warmup + 1000 draws), `target_accept`=0.99. A fit is flagged for retry at $\hat{R} > 1.05$ or $\text{ESS}_{\text{min}} < 100$; 80/80 fits over the 16-window × 5-stat panel converged on first attempt. Worst observed $\hat{R}$ across the panel was 1.019; worst ESS_{min} was 296.

B v3 → v3.2 anchor-chain worked example

The v3→v3.2 change at each window decomposes into a drift component (uniform additive shift, preserves within-window ranks) and a scale component (compresses or amplifies extremes at windows where per-window pool variance differs from anchor variance). Working through a single player at a peak-drift window makes this concrete.

Take *aspas* at the 2025-01-31→2025-04-30 window, where cumulative FKPR drift peaks at $\Delta_{\text{FKPR}} = +0.070$ log-rate (roughly 7% inflated baseline relative to the anchor) and cumulative KPR drift peaks at $\Delta_{\text{KPR}} = +0.060$. The per-stat z shifts are:

stat	z_{v3}	$z_{\text{v3.2}}$	Δz	comment
ADR	+1.12	+0.95	−0.17	small; ADR drift near zero
KAST	+1.99	+2.28	+0.30	scale effect; anchor KAST sd wider
KPR	+2.45	+2.43	−0.02	drift and scale near-cancel
FKPR	+0.55	+0.01	−0.54	headline: FKPR was entirely pool drift
APR	+0.17	−0.24	−0.41	drift + scale; below anchor baseline

Composed through the PC1 loadings, $\text{pc1}_{\text{v3}} = +3.03$ drops to $\text{pc1}_{\text{v3.2}} = +2.72$ ($\Delta = -0.31$; on the 0–100 scale, 79.1 → 75.8). The rank at that window is unchanged (2 of 151). The reader takeaway is that the mid-2025 FKPR "spike" attributed to aspas under v3 was pool-wide meta inflation; his genuine improvements on KPR and KAST both survive v3.2.

C PMR v1 → v1.1 → v1.2 case comparison

Five representative (player, map) rows showing how the display number evolves across the three calibration versions. Per-stat residuals and rounds played r are identical across versions; only the rescaling differs.

direction	player / map	r	overall	v1	v1.1 (SE)	v1.2 (SE)	λ
short over	akeman / Split	13	30	100 (clip)	64 (10)	**49** (8)	0.56
long over	guang / Breeze	19	22	98	36 (10)	**30** (7)	0.59
long under	marteen / Corrode	18	100	18	84 (9)	**90** (7)	0.60
short under	benjyfishy / Corrode	13	60	0 (clip)	35 (10)	**46** (8)	0.56
on expect.	xeus / Pearl	24	46	46	46 (9)	**46** (7)	0.62

Directionality is preserved across all rows (over- remains over-, under- remains under-); clipped extremes (v1) move to interpretable near-edge values (v1.1) and then contract toward the prior under v1.2 in proportion to sampling noise. The on-expectation row barely changes: the multiplicative shrinkage of a near-zero residual is a near-zero effect.

D Appendix D: Top-20 anchor-window PC1 players

Table 1 lists the top 20 players by PC1 score in the anchor window (2026-01-31 to 2026-04-30) at the $n_{\text{maps}} \geq 10$ ship threshold. "PC2 rank" is the player's rank within the same pool on the orthogonal utility axis; the large PC1/PC2 rank divergences (e.g. *zmjjkk* PC1 rank 3, PC2 rank 121; *derke* PC1 rank 2, PC2 rank 2) illustrate the two-products argument from Section **??**.

#	Player	Team	Role	Maps	PC1	PC2	PC2 rank
1	marteen	Gentle Mates	duelist	22	100.0	63.9	22
2	derke	Team Vitality	duelist	11	90.5	98.0	2
3	zmjjkk	EDward Gaming	duelist	21	79.7	5.7	121
4	scales	TYLOO	controller	13	76.6	23.2	118
5	hyunmin	KIWOOM DRX	duelist	12	75.6	38.6	97
6	miniboo	Team Liquid	duelist	22	75.6	55.7	39
7	erv	TYLOO	controller	13	71.2	60.5	26
8	hiro	Natus Vincere	sentinel	13	69.8	42.6	85
9	chronicle	Team Vitality	sentinel	11	69.3	66.0	13
10	verno	MIBR	initiator	16	68.5	69.8	7
11	valyn	G2 Esports	controller	24	67.4	56.7	34
12	dos9	Karmine Corp	controller	13	66.6	51.0	59
13	kamo	Team Liquid	duelist	22	66.5	69.7	8
14	buzz	T1	duelist	18	66.2	52.8	51
15	nats	Team Liquid	controller	22	65.9	51.5	54
16	yetujey	FUT Esports	controller	15	65.5	40.0	95
17	zekken	MIBR	controller	16	65.3	46.7	71
18	jemkin	Rex Regum Qeon	duelist	17	64.2	44.3	82
19	mada	NRG	duelist	30	63.7	31.2	110
20	splash	TYLOO	duelist	13	63.2	54.5	43

Table 1: Top 20 anchor-window players by PC1 score. All scores on the 0–100 display transform centred on the anchor-pool mean of 50. PC2 rank is out of 121 anchor-pool players.