RESEARCH NOTE

Prospective Evaluation of a Results-Only Cross-League Football Rating Model

UEFA Champions League and Europa League league-phase forecasts, 2024/25-2025/26

Author: Sten Ruediger, Rank-O-Football

Report date: 15 July 2026

Study design: Complete recorded ledger cohort; retrospective scoring of realized 90-minute outcomes

Primary comparison: Ranko versus day-before BWIN probabilities in Champions League matches

Abstract

Objective. To evaluate Ranko in the inter-league domain for which it was designed. **Design.** We scored every 2024/25 or 2025/26 BetsAnalysis row labelled Champions League (CL) or Europa League (EL) with complete Ranko odds, BWIN odds and a known 90-minute result. Ranko and BWIN odds were normalized symmetrically. **Results.** The cohort contains 64 CL and 14 EL matches. In CL, Ranko selected 40/64 outcomes correctly (62.5%) and BWIN 36/64 (56.3%): a paired difference of +6.25 percentage points (95% paired bootstrap interval -7.81 to +20.31; exact McNemar p=0.503). Ranko exceeded BWIN by two correct picks in each season. CL Brier scores were nearly identical (0.5570 versus 0.5539), RPS was nearly identical (0.2121 versus 0.2124), and log loss favored BWIN (1.0433 versus 0.9432). In the 60-match CL closing-market sensitivity subset, Ranko and the closing average each selected 38 outcomes correctly. EL results were unfavorable to Ranko (5/14 versus 9/14). **Conclusion.** Ranko displayed competitive observed CL top-pick accuracy and comparable Brier/RPS performance in this recorded cohort. The data do not establish superiority or formal equivalence.

Plain-language result. In 64 Champions League matches, Ranko made four more correct top picks than BWIN, with the same +2 difference in each season. That repeat is encouraging, but the paired uncertainty interval is wide enough to include no advantage. The defensible statement is competitive observed CL performance, not 'Ranko beats bookmakers.'

Keywords: football forecasting; cross-league rating; proprietary model; Champions League; proper scoring rules; paired validation

1. Introduction

Most domestic football prediction tasks compare teams that meet repeatedly within the same league. Ranko was designed for a different problem: estimating the relative strength of clubs from England, Germany, Spain, Italy and France on a common scale so that teams from different domestic competitions can be compared. Its public input description is limited to completed domestic and UEFA match results; the forecast-generation implementation is proprietary.

The UEFA league-phase format is therefore a direct test environment. It generates many inter-league fixtures while preserving the practical difficulty that clubs do not share a complete domestic schedule. High-quality archived prediction datasets for this specific use case are scarce. Ranko's own ledger is valuable because it contains match-level forecast odds and a contemporaneous BWIN reference, but it is not a preregistered public archive and does not contain independently verifiable cell-level timestamps.

The primary question is whether Ranko's observed CL top-pick accuracy is competitive with a serious market reference. The secondary questions concern probability quality, season-to-season stability, a separately matched closing-market subset and exploratory EL performance. The analysis is deliberately paired: every method is scored on the same match.

2. Forecast model and disclosure boundary

2.1 Intended use and public input description

Ranko is a proprietary results-only cross-league rating and forecast system. Its declared inputs are finished matches from the Premier League, Bundesliga, Serie A, La Liga and Ligue 1, plus UEFA fixtures involving those clubs. It does not use bookmaker odds, lineups, injuries, transfers, suspensions, weather, expected goals or event-level match statistics.

Proprietary probability conversion. Relative Ranko values are converted to H/D/A probabilities by a confidential calibrated mapping. Its functional form, fitted parameters, weighting rules and source code are not disclosed.

2.2 Reproducibility boundary

This is a validation report, not a model-development disclosure. The published ledger supplies the fixed forecast probabilities, benchmark probabilities and realized outcomes required to reproduce every evaluation statistic. Readers can therefore audit the reported accuracy, proper scores, paired intervals and sensitivity analyses without receiving the proprietary forecast-generation algorithm. The recorded forecasts are treated as fixed black-box outputs throughout the study.

2.3 Why the evaluation is domain-specific

The distinctive claim is not that a results-only model should dominate a bookmaker on ordinary domestic fixtures. It is that a system designed for cross-league comparison can make useful forecasts when clubs from different domestic leagues meet. CL results are therefore the primary domain. EL is retained because it is also inter-league, but its smaller sample and potentially different rotation incentives make it exploratory.

3. Study design and reproducible analysis

3.1 Source ledger and cohort

The source ledger was the internal BetsAnalysis sheet. The public validation dataset is a row-level extract containing only the 78 included matches and the fields required to reproduce the primary Ranko-versus-BWIN evaluation. Inclusion required: (i) season 2024/25 or 2025/26; (ii) competition label CL or EL; (iii) three Ranko odds; (iv) three BWIN odds; and (v) a known 90-minute score. Rows labelled CO were excluded. Closing-market availability was not an inclusion criterion for the primary analysis.

Season	Competition	n	Coverage
2024/25	CL	30	Matchdays 4-8
2024/25	EL	6	Matchdays 4-8
2025/26	CL	34	Matchdays 2-7
2025/26	EL	8	Matchdays 2-7
Total	CL + EL	78	64 CL; 14 EL

The ledger export preserves forecast values but not cell-level edit timestamps. Its pre-kickoff status rests on the documented operational workflow used to prepare day-before selections. This is weaker evidence than an immutable timestamped archive and is treated as a limitation.

3.2 Probability normalization and outcome

For each decimal-odds triplet, the implied home/draw/away probabilities were normalized by dividing each inverse odd by the sum of all three inverse odds:

$$p_j = \frac{o_j^{-1}}{\sum_{k \in \{H,D,A\}} o_k^{-1}}, j \in \{H, D, A\}.$$

The same transformation was applied to Ranko and BWIN. Only the 90-minute H/D/A result was scored; extra time and penalties were irrelevant.

3.3 Scoring rules

Let y_j be the one-hot indicator for the realized outcome and let p_y be the probability assigned to that outcome. Top-pick accuracy is one when the outcome with the largest forecast probability is realized. The scores were computed as follows:

$$\hat{y} = \arg\max_{j \in \{H,D,A\}} p_j.$$

$$\text{Brier} = \sum_{j \in \{H,D,A\}} (p_j - y_j)^2.$$

$$\text{LogLoss} = -\ln(p_y).$$

$$\mathrm{RPS} = \frac{1}{2}\sum_{r=1}^{2}\left(\sum_{j=1}^{r} p_j - \sum_{j=1}^{r} y_j\right)^2.$$

The RPS uses the ordered outcome categories H-D-A and averages the squared errors of the first two cumulative probabilities. Brier score, log loss and RPS are minimized, so lower values indicate better probabilistic forecasts.

3.4 Statistical analysis

Accuracy intervals use the Wilson method. Comparisons are paired by match. Exact two-sided McNemar tests use discordant correct/incorrect counts. Paired percentile bootstrap intervals resample whole matches with replacement, retaining both forecasts and the realized outcome within each draw. We used 20,000 replicates and deterministic seed 20260714. No formal equivalence margin was preregistered, so a non-significant difference must not be interpreted as proof of equivalence.

4. Results

4.1 Champions League primary result

Forecast	n	Correct	Accuracy	Brier	Log loss	RPS
Ranko	64	40	62.5%	0.5570	1.0433	0.2121
BWIN day-before	64	36	56.3%	0.5539	0.9432	0.2124

Ranko selected 40 of 64 CL outcomes correctly (62.5%; Wilson 95% interval 50.3%-73.3%). BWIN selected 36 (56.3%; Wilson interval 44.1%-67.7%). The point estimate therefore favors Ranko by four matches or 6.25 percentage points. Brier and RPS point estimates are virtually identical. BWIN's lower log loss indicates better protection against confident errors.

4.2 Paired uncertainty

Ranko only	BWIN only	Difference	Paired 95% interval	McNemar p
12	8	+6.25 pp	-7.81 to +20.31 pp	0.503

The paired interval crosses zero and the exact test is non-significant. The observed advantage is therefore compatible with a meaningful Ranko benefit, no difference, or a modest BWIN benefit. It supports continued evaluation but not a claim of established superiority.

4.3 Replication by season

Season	n	Ranko correct	Ranko acc.	BWIN correct	BWIN acc.	Difference
2024/25	30	21	70.0%	19	63.3%	+6.67 pp
2025/26	34	19	55.9%	17	50.0%	+5.88 pp

The count relationship is exactly replicated: Ranko made two more correct CL picks than BWIN in each season. Absolute accuracy declined in 2025/26 for both methods. Replication of the direction is encouraging, but neither season-specific paired interval excludes zero.

4.4 Probability scores by season

Season	Score	Ranko	BWIN	Ranko - BWIN
2024/25	Brier	0.4836	0.5257	-0.0421
2024/25	Log loss	0.8853	0.9033	-0.0180
2024/25	RPS	0.1941	0.2084	-0.0143
2025/26	Brier	0.6217	0.5789	+0.0428
2025/26	Log loss	1.1827	0.9784	+0.2043
2025/26	RPS	0.2280	0.2160	+0.0120

Negative score differences favor Ranko. The favorable 2024/25 probability scores did not replicate in 2025/26. Pooling produces an almost exact Brier/RPS tie but a BWIN advantage in log loss. This pattern is consistent with occasional confident Ranko errors.

4.5 Matchday-window sensitivity (CL)

To test dependence on the final league-phase matchday, we repeated the paired CL comparison in three inclusive windows ending at matchday 7. Matchday 8 was excluded because qualified teams may rotate. The nested windows are sensitivity summaries, not independent tests.

Window	n	Ranko	BWIN	Difference	Paired 95% interval	McNemar p
MD2-7	59	36 (61.0%)	33 (55.9%)	+5.08 pp	-8.47 to +20.34 pp	0.648
MD3-7	56	35 (62.5%)	33 (58.9%)	+3.57 pp	-10.71 to +17.86 pp	0.815
MD4-7	50	31 (62.0%)	28 (56.0%)	+6.00 pp	-10.00 to +22.00 pp	0.629

Ranko's observed top-pick accuracy was 3.57 to 6.00 percentage points higher in all three windows. All paired intervals included zero and all exact McNemar tests were non-significant. Thus the favorable point estimate is not attributable only to matchday 8; the overlapping subsets do not establish superiority.

4.6 Closing-market sensitivity subset

BetExplorer closing-average odds were matched for 60 CL and 13 EL rows. Because closing odds are later than day-before BWIN and are not available for every ledger row, this is a secondary timing-mismatched sensitivity analysis, not the primary cohort definition.

CL forecast	n	Correct	Accuracy	Brier	Log loss
Ranko	60	38	63.3%	0.5536	1.0388
BWIN day-before	60	34	56.7%	0.5552	0.9448
Closing average	60	38	63.3%	0.5467	0.9310

Ranko tied the closing average's top-pick count in the matched CL subset and exceeded BWIN by four picks. The closing average achieved lower Brier and log loss. Matching a top-pick count does not establish probability equivalence; the paired Ranko-versus-closing accuracy interval was -11.7 to +11.7 percentage points.

4.7 Europa League exploratory result

Forecast	n	Correct	Accuracy	Brier	Log loss	RPS
Ranko	14	5	35.7%	0.7976	1.4341	0.2998
BWIN day-before	14	9	64.3%	0.6020	1.0087	0.2132

Ranko trailed BWIN by four EL top picks. The paired accuracy difference was -28.6 percentage points (bootstrap interval -57.1 to 0.0; exact McNemar p=0.219). The sample is small, but it provides no basis for a positive EL claim. Rotation is a plausible hypothesis, not a tested explanation.

5. Discussion

5.1 What the positive result supports

The strongest favorable observation is narrow and domain-specific. Across two CL seasons, a results-only cross-league model made four more correct top picks than day-before BWIN, with the same +2 count in each season. The favorable accuracy direction also persisted in each nested MD2-7, MD3-7 and MD4-7 window after excluding matchday 8. Its pooled Brier and RPS were essentially indistinguishable from BWIN at the point estimate. In the smaller closing subset, Ranko matched the closing average's top-pick count. For a model deliberately limited to match results, this is credible evidence of competitive observed CL performance.

The result must not be inflated. The paired interval is wide, log loss favors BWIN, and no equivalence margin was specified. 'Competitive in this recorded CL cohort' is supported; 'beats bookmakers' is not.

Seasonal network calibration. Domestic results primarily refresh the within-league regions of the rating network; the first UEFA matches add current cross-league links and can recalibrate their relative position for later forecasts. This structural updating is distinct from calibration of H/D/A probabilities. Because the nested MD2-7, MD3-7 and MD4-7 cohorts neither isolate the effect nor show a monotonic gain, it remains a mechanism for prospective testing with a predefined early-season calibration window.

5.2 Interpretation of the mixed scoring profile

Accuracy rewards only the most likely category. Brier, log loss and RPS evaluate the full vector. The near-equal pooled Brier/RPS suggests that Ranko's average probability geometry is comparable to BWIN in CL, while the worse log loss shows greater exposure to high-confidence mistakes. Calibration shrinkage of extreme forecasts is therefore a more defensible development priority than optimizing top-pick accuracy alone.

5.3 Why EL should not be combined into the headline

Pooling CL and EL yields equal total top-pick counts (45/78 each) because Ranko's favorable CL difference is offset by its unfavorable EL difference. The competitions should not be collapsed into one marketing statistic: CL is the prespecified primary domain, whereas EL has only 14 matches and may differ structurally. The negative EL result remains part of the scientific record.

5.4 Limitations

- The ledger is a recorded subset, not a complete tournament census; 2024/25 tracking begins at matchday 4.
- The analysis was specified after data collection and was not preregistered before the two seasons.
- BWIN is a day-before reference, while BetExplorer is a closing average; those references answer different timing questions.
- The EL sample is only 14 matches; proposed rotation mechanisms are untested.

5.5 Next prospective test

A stronger test should freeze and publish every eligible forecast before kickoff with an immutable identifier; specify eligibility, competition and matchday cutoffs before results; acquire same-time market references where licensing permits; score every included match; publish exclusions; and report paired intervals, proper scores and calibration diagnostics. A club- or matchweek-clustered sensitivity analysis should be added once the archive is large enough.

6. Reporting and publication policy

The complete Ranko-versus-BWIN ledger cohort is distinct from the smaller closing-market matched subset. No Pinnacle superiority claim is made because a complete prospectively recorded Pinnacle series is not available for the 64-match CL cohort.

Publication rule. Headline claims must state the competition, sample size, reference timing and uncertainty. A favorable point estimate may be described as observed evidence; superiority requires an appropriate paired interval or test, and equivalence requires a prospectively defined margin.

7. Data and code availability

The public Google Sheet contains the complete 78-match evaluation dataset, row-level formulas, summaries, closing-market sensitivity and data dictionary. The XLSX download is its versioned archival snapshot (14 July 2026). Forecast generation remains proprietary; evaluation data and calculations are public.

- Public Google Sheets validation dataset
- Versioned dataset snapshot (XLSX)
- Study overview article

8. Conclusion

Ranko was created to estimate relative strength across domestic leagues. In the 64-match CL cohort, it selected 40 outcomes correctly versus 36 for day-before BWIN, with a +2 difference in each season. Brier and RPS were almost identical, while BWIN had better log loss. These findings support a carefully limited statement of competitive observed CL performance and justify a larger preregistered prospective test. They do not establish a bookmaker-beating edge or formal equivalence.

References

[1] UEFA. New format for Champions League post-2024.

[2] Brier GW. 1950. Verification of Forecasts Expressed in Terms of Probability.

[3] Good IJ. 1952. Rational Decisions. Journal of the Royal Statistical Society, Series B 14(1):107-114.

[4] Gneiting T, Raftery AE. 2007. Strictly Proper Scoring Rules, Prediction, and Estimation.

[5] Wilson EB. 1927. Probable Inference, the Law of Succession, and Statistical Inference.

[6] McNemar Q. 1947. Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages.

[7] Rank-O-Football. Public model overview.

[8] BetExplorer. Champions League results and 1X2 odds.

Appendix A. Worked scoring example

For Bologna-Monaco on 5 November 2024, the Ranko odds were 15.10, 6.01 and 1.30. Their reciprocal values were normalized to H=0.066103, D=0.166083 and A=0.767814. Monaco won 1-0 in 90 minutes, so Ranko's top pick was correct. The row-level Brier contribution was 0.085864, log loss 0.264208 and RPS 0.029140. The Excel Calculations sheet exposes the formulas for every match.

Appendix B. Evaluation reproducibility checklist

Audit item	Workbook location
Fixed study identities and source references	Raw_Matches, 78 rows
Symmetric odds normalization	Calculations N:P and Y:AA
Row-level accuracy and scores	Calculations Q:AJ
Full and stratified summaries	Results
Closing matched subset	Closing_Sensitivity
Definitions and sign conventions	Data_Dictionary
Bootstrap specification	20,000 paired draws; seed 20260714
Forecast generation	Proprietary; fixed outputs supplied