Forecasting Football Careers: A Foundation Model for Player Trajectories, with Transparent Backtests and a Pre-Registered Forward Evaluation

Oraca Research

11 June 2026

CONTENTS

Forecasting Football Careers: A Foundation Model for Player Trajectories, with Transparent Backtests and a Pre-Registered Forward Evaluation

Oraca Research · oraca.ai · research@oraca.ai · Version 1.0 · 11 June 2026

Artifacts: predictions ledger and pre-registration hash (§9) at oraca.ai/research

---

Abstract

A recruitment decision in professional football is a forecast. The player a club signs today is signed for what he will be in two or four years, yet the decision is usually made from current form and market price. We model a player's career as a sequence of monthly observations and train a decoder-only transformer, autoregressively, over 53,353 careers (2.73 million player-months across 30 leagues) to forecast future development. Evaluation is strictly as-of: the model sees only data visible at a historical cutoff, and outcomes are measured years later. Given only information available on 30 June 2021, the model ranked 3,418 under-24 players by the probability of reaching the top 5% of their position within four seasons. 199 of them did (a base rate of 5.8%), and the model's top 50 contained 37, against 32 for the strongest baseline at the same cutoff and roughly three expected at random. A second vintage, trained independently with a 2017 cutoff and scored on 2020 outcomes, reproduced the result with AUC 0.91. We also evaluate a companion undervaluation screen over nine annual as-of anchors (2016–2024): flagged young players subsequently moved up a league tier at 1.95 times the cohort base rate (444 observed against 228 expected), and the lift cleared 1.3 in every individual year. We release a complete predictions ledger containing every prediction in this paper joined to its realized outcome, and we pre-register a forward evaluation, publishing the cryptographic hash of a frozen prediction set to be scored publicly after two transfer windows, whatever the result shows. The paper also reports what our own validation killed: a non-replicated internal estimate and a contaminated historical backtest we rejected.

---

1. Introduction

A sporting director at a promotion-chasing club faces a search problem with bad geometry. The eligible universe runs to tens of thousands of players; a scouting department can properly watch a few hundred a season; a transfer budget punishes mistakes for years. Within that squeeze, the question that matters is rarely "how good is this player now?" Current ability is comparatively easy to estimate and is in any case priced in. The valuable question is what the player will be in two or three years, and whether that future is worth more than today's price.

That is a forecasting problem, and it has structure. Careers unfold as sequences: minutes precede starts, a move up in tier rewards or punishes development, and a given level of performance carries different information at 19 than at 23. Most quantitative tools in football answer the snapshot question instead — ratings, percentile profiles, action values. This paper describes a system built for the trajectory question. We represent a career as a sequence of monthly timesteps carrying playing time, production, club and competition context, valuation, and development signals, and we train a generative sequence model, of the same architectural family that underlies modern language models, to predict each next month given everything before it. Sampling continuations of a career then yields a distribution over futures: the probability a player reaches an elite level, the minutes he is likely to play, the risk he drifts out of covered professional football altogether.

Claims about forecasting are cheap. The football analytics industry has a particular weakness for historical validations that quietly use information from the future, and we see no reason our claims should be exempt from the suspicion this has earned. We therefore hold the paper to the evidence standard we would want applied to any vendor:

1. Every reported result comes from a model given only data visible at a historical cutoff and scored against outcomes realized afterward. The leakage controls are enumerated in §6 and asserted in code.
2. Baselines are strong and decision-relevant, including the forecast a club implicitly uses already: ranking by market value.
3. The back data is released. Our predictions ledger contains every prediction underlying every table here, joined to its realized outcome, so a skeptical reader can recompute the headline numbers without trusting us.
4. The next test is pre-registered. A frozen prediction set is hashed in print and will be scored publicly after two transfer windows, win or lose.
5. Negative results are in the paper, not in a drawer.

Our contributions: a foundation model over player-career sequences for development forecasting; a leakage-controlled, walk-forward evaluation protocol for career forecasting, with results on two independent clean vintages; a nine-anchor as-of evaluation of a market-facing undervaluation screen; and the public ledger and pre-registered forward test described above.

2. Related Work

Rating the present is a mature field. Elo-family ratings aggregate results into a strength scale [1, 2]; possession-value frameworks such as VAEP credit individual actions [3]; plus-minus methods isolate player contributions from lineup outcomes [4]. Our team-strength backbone (§5.1) belongs to this family. The career model does not: it forecasts what a player becomes, not what he is.

On development itself, the age-curve literature establishes broad position-dependent rise and decline patterns [5], and transfer-valuation work models prices. We are not aware of a published system that forecasts individual development distributionally from full career sequences under an as-of protocol, and we would welcome correction on this point.

Architecturally we follow the decoder-only autoregressive transformer [6, 7], with numeric binning and channel-embedding choices related to recent time-series transformers [8]. “A career is a language” is rhetoric, but the engineering consequence is real: next-token prediction over careers admits the same training recipe, and the same evaluation traps, as any generative model.

None of our evaluation choices is novel in isolation. Temporal holdout, pre-specified metrics, bootstrap uncertainty [9], calibration analysis [10], released artifacts [11, 12]: each is standard somewhere. Their joint application to football scouting models is, as far as we know, not.

3. Problem Formulation

Let a player’s career be a sequence (x_{1..t}) of calendar-month observations (§5.2). Tasks are defined relative to an information date (T), the cutoff; models may use only data visible at (T).

T1 (development forecast). For each player, forecast the development of his performance level over horizons up to 48 months, as a distribution over futures.

T2 (shortlist). At cutoff (T), rank the cohort by the probability of reaching the top 5% of position by outcome season (S). The cohort is every player who at (T) is under 24 years old, has at least 600 career minutes in covered competitions, and is active, meaning non-zero recent covered minutes. A player realizes the outcome if at season (S) he has a valid observed rating anchor with at least 600 season minutes and his level ranks in the top 5% of his position among all such players (Appendix A). The primary evaluation sets (T) = 2021-06-30 and (S) = the 2025/26 season, four full seasons of outcome runway.

T3 (undervaluation screen). At each annual anchor (T), among regular players under 26 with goalkeepers excluded, flag the top 5% by an undervaluation signal: the residual of a player’s estimated level relative to what his club’s strength, league tier, and age would predict (§5.4). The primary outcome is that the player is in a better league tier in season (T{+}1). A stricter secondary outcome requires that he changed clubs to a stronger club.

All outcome definitions were fixed before evaluation. No post-hoc variants are reported. Exact specifications, including tie and edge-case handling, are in Appendix A and implemented verbatim in the released evaluation code.

4. Data Types and Coverage

The system is built from the following data types: match- and appearance-level performance records (lineups, minutes, goals, and detailed in-match statistics where available); competition results and standings; transfer and career-history records; market-valuation histories; international appearance records; biographical attributes; and competition structure (league tiers, season calendars). Input data is drawn from commercial and public sources under terms that do not permit redistribution; we therefore release model outputs and evaluation results rather than inputs (§10).

The corpus spans 30 domestic leagues plus four UEFA club competitions: 6.02 million appearance records covering 71,571 players, from 2005 in the largest leagues and roughly 2010 elsewhere. The 2021-cutoff training corpus contains 53,353 players and 2.73 million player-months. Coverage is uneven, and the unevenness is itself information the model receives:

Era	Coverage character
2005–2009	Largest leagues only: lineups, minutes, goals
2010–2014	~25 leagues: lineups, minutes, goals; detailed statistics largely absent
2015–2019	Top leagues gain detailed in-match statistics
2019–2022	Detailed statistics across nearly all 30 leagues
2023–	Detailed statistics ~98%+ everywhere

Every input channel carries an explicit missing bin. The absence of a statistic encodes league tier and era, and the model is told about it rather than having it silently imputed. The per-competition table (first covered season, first detailed-statistics season, appearance counts) is Appendix B.

Lower-tier and youth coverage is thinner and arrives later in the historical record, and international-appearance coverage varies by confederation and era. The competition list in Appendix B defines the system's horizon.

5. Model

We describe the model family, the input representation, and the training objective. Exact feature recipes, architecture sizes, and training schedules are proprietary. Nothing in the evaluation asks the reader to trust an undisclosed component, because every claim is backed by the released ledger.

5.1 TEAM-STRENGTH BACKBONE

An Elo-family rating over club results across all covered competitions provides club- and league-strength context. Clubs meet across leagues, in continental competition, cups, and through promotion, so the scale is connected internationally; it is validated for cross-league comparability at the team level. Player performance is contextualized against opponent strength within league. The player-level opponent adjustment does not itself transfer across leagues. This is a design limitation, and we return to it in §11.

5.2 CAREER-SEQUENCE REPRESENTATION

One player is one sequence of calendar-month timesteps, from first covered appearance to the information date. Zero-minute months are real timesteps, since absence is signal, and a trailing window of post-career months teaches the model how careers end. Each timestep carries roughly 25 channels: playing time (minutes, starts, appearances, squad presence), per-90 production, club strength and league tier, market valuation (sparse), per-season development anchors on a common scale, injury status, international caps, club-change events, and position. Continuous channels are quantile-binned, with bin 0 reserved for missing in every channel. The input embedding is the sum of per-channel embeddings, a sinusoidal age encoding (age, not sequence index, is the positional signal), a calendar-month embedding, and static player conditioning.

5.3 ARCHITECTURE AND TRAINING

A decoder-only transformer is trained autoregressively, with one softmax head per channel predicting the next month's bin given the career so far. Two points of training discipline matter for what follows. First, for backtest models the cutoff is enforced at the frame level before sequence assembly, and every normalization (quantile bins, thresholds, position labels) is fit only on data visible at the cutoff. Second, hyperparameters and model selection were decided within-cutoff. Forecasts are produced by sampling continuations (Monte Carlo rollouts) under structured decoding: channels that exist only at certain calendar points, such as season-end development anchors, are generated only at those points, mirroring the data-generating process. The probability that a player reaches the top 5% of his position is read off the sampled futures.

5.4 FROM FORECASTS TO DECISION SURFACES

The shortlist ranks the cohort by the rollout-derived probability. Reported probabilities additionally pass through isotonic calibration [10] fitted on an earlier cohort (2017) against outcomes that predate the primary cutoff; the calibrator never sees a 2021-cohort outcome. The undervaluation screen is deliberately simple and fully disclosed. A player's estimated level is regressed, by least squares refit as-of each anchor, on club rating, league tier, and age over regular players; the screen flags the top 5% of residuals, the players whose level most exceeds what their context implies. It is a selection device and is evaluated as one (§7.2).

6. Evaluation Methodology

The results in §7 are worth exactly as much as the separation between what the model saw and what it is scored on. This section specifies that separation.

6.1 AS-OF PROTOCOL AND LEAKAGE CONTROLS

For each evaluation we train with data strictly before the cutoff, freeze all derived quantities as-of the cutoff, score the cohort, and measure outcomes strictly after, by the fixed definitions of §3. The controls, all asserted in the evaluation harness's code:

1. The cutoff is enforced at the frame level, before sequence assembly.
2. All normalizations are fit only on cutoff-visible data.
3. Development anchors attach to season-end months, so cutoff-season anchors are excluded by construction.
4. Cohort membership is computed as-of the cutoff. There is no survivorship: players who later vanished remain in the cohort and count against the model wherever it predicted they would succeed.

5. Hyperparameters and model selection were decided within-cutoff.
6. Outcome labels are computed only from post-cutoff data, by pre-fixed definition.
7. A 5% player-level holdout guards against memorization of individual careers on the training side.
8. Nine boundary unit tests assert the cutoff in code.

One residual risk survives all of this, and we disclose it rather than wave at it. Today's database reconstructs the past; later data corrections and identity merges mean the 2021 snapshot is rebuilt, not archived. We control this in-pipeline, but only forward evaluation eliminates it. That is what §9 is for.

6.2 BASELINES

Baselines are chosen to be what a club could do without this system, plus the strongest simple statistical competitor. Baseline (a) ranks by current-ability percentile within position — sign the best young player today. Baseline (b) is a supervised logistic model over as-of features (age, current level, minutes, league strength, recent trajectory), trained on an earlier cohort against pre-cutoff outcomes; it is the strongest fair statistical baseline. Baseline (c) ranks by market valuation, which is both the market's own forecast and, implicitly, the rule a budget-constrained club follows. Random selection appears as the base rate.

One detail of baseline (b) deserves a sentence, because it says something about evaluation hygiene in general. Its league-strength feature originally carried a chronological-ordering bug: players on loan were attributed to stale clubs, because a last-row lookup ran over unsorted match records. The bug was caught when a reader of a published list noticed the stale clubs. The corrected baseline is stronger than the one we first measured, and the corrected version is what appears in every table here. All baselines see the identical cohort, cutoff, and outcome definitions.

6.3 METRICS

Primary metrics, fixed in advance: precision@50 and precision@100 for the shortlist, and realized move-up lift of the flagged top-5% for the screen. Fifty is roughly the number of players a club can genuinely watch in a season; it is the operating point this system exists for. AUC is reported as a cohort-wide summary, with percentile-bootstrap 95% intervals over 1,000 cohort resamples [9], and calibration is examined by score band. No metric, threshold, or horizon was chosen after seeing results.

6.4 SLICES

Results are sliced by position, by age band, and by pre-cutoff league strength (terciles of the cohort by the league-strength rating of each player's modal pre-cutoff competition). The league-strength slice exists to interrogate the known weak point, thin lower-tier coverage. It was not chosen to flatter the headline, and as §7.4 shows, it does not.

7. Results

7.1 SHORTLIST TASK (PRIMARY, CUTOFF 2021-06-30 → SEASON 2025/26)

The cohort contains 3,418 under-24 players, of whom 199 realized the top-5% outcome, a base rate of 5.8%.

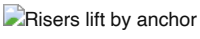
System	p@50	p@100	AUC [95% CI]	lift@100
Career forecaster (calibrated)	**0.74**	**0.58**	0.861	**10.0×**
Career forecaster (raw ranking)	**0.74**	0.55 [0.45, 0.65]	**0.866** [0.835, 0.893]	9.4×
(b) cutoff-trained logistic	0.64	0.55 [0.43, 0.64]	0.856 [0.827, 0.882]	9.4×
(c) market-value ranking	0.62	0.52 [0.42, 0.63]	0.721 [0.670, 0.769]	8.9×
(a) current-ability percentile	0.46	0.41 [0.33, 0.53]	0.808 [0.775, 0.840]	7.1×
random	0.058	0.058	0.500	1.0×

Watch the model's top 50 and 37 become top-5%-of-position players within four seasons. The strongest baseline delivers 32; the market delivers 31; random delivers about three. The market-value comparison is the instructive one. The market is a good forecaster in aggregate (lift 8.9 at depth 100), and the model's advantage concentrates at the head of the list, which happens to be where a club spends its watching budget. The current-ability baseline confirms the premise of the paper from the other direction: knowing who is best today is a much worse guide to who will be best in four years (p@50 of 0.46 against 0.74) than modeling the trajectory.

Could one cutoff be a fluke? A separately trained model with cutoff 2017-06-30, evaluated on the 2017 cohort (n = 3,406, base rate 4.8%) against season-2020 outcomes, reached AUC 0.913, p@50 0.66, and lift@100 of 11.6. Two clean vintages, four years apart, both clear of every baseline. The 2017→2020 era is genuinely easier to predict than 2021→2025; we suspect post-COVID transfer-market disruption accounts for part of the difference, though we cannot demonstrate it.

7.2 UNDERVALUATION SCREEN (NINE AS-OF ANCHORS, 2016–2024)

At each annual anchor the screen flags the top 5% of under-26 regulars by the disclosed residual of §5.4, and the outcome is a better league tier the following season. Across nine consecutive anchors, 2,477 player-seasons were flagged. 444 moved up, where base rates predict 228: a pooled lift of 1.95, with every individual anchor clearing 1.3 (range 1.32 to 2.39, median 2.00):



Under the stricter secondary outcome, in which the player must have changed clubs to a stronger club, isolating individual moves from club promotion, the pooled lift is 1.51 (range 1.22 to 1.90). Per-anchor tables for both outcome definitions are in Appendix C.

This is a selection claim: the flagged cohort realizes materially elevated move-up rates. It is not a discrimination claim, and the distinction is not pedantry: an earlier internal estimate of the signal's discriminative value did not survive this paper's protocol. See §8.

7.3 CALIBRATION

Calibration is evaluated the hard way. The isotonic calibrator was fit on the 2017 cohort against pre-cutoff outcomes and applied untouched to the 2021 cohort, a four-year transfer across eras.

Predicted band	mean predicted	realized	n
< 0.01	0.004	0.011	2,262
0.01–0.05	0.03	0.05	552
0.05–0.20	0.10	0.12	266
0.20–0.60	0.33	0.22	199
0.60–0.80	0.64	0.38	80
≥ 0.80	0.95	0.66	59

The ranking is faithful: realized rates rise monotonically across bands, and mid-range probabilities track well. At the top, transferred probabilities overstate 2021-era frequencies: a predicted 0.95 realized at 0.66. Era shift compressed what was near-certainty in 2017-era data. We report ranking metrics as primary for this reason, and we state the bias rather than refitting the calibrator after the fact. A user should read top-band probabilities as "very likely by historical standards," not as literal frequencies.

7.4 SLICES

By position, AUC spans 0.78 to 0.92 with positive selection lift everywhere (Appendix C). By age, the under-19 slice has the highest base rate at 19.5% (simply being visible in covered football at 17 is already a strong signal) and the highest head precision, 0.75 on the top 5% of the slice. The slice that matters commercially is league strength:

Pre-cutoff league strength	n	base rate	AUC	top-5%-of-slice precision	lift
Top tercile	1,019	12.8%	0.854	0.71	5.5×
Middle tercile	1,210	4.0%	0.814	0.27	6.6×
Lower tercile	1,134	1.2%	0.681	0.05	4.3×

The model keeps useful selection lift in the lower tercile, 4.3 times a 1.2% base rate, but discrimination degrades badly there: AUC 0.68, where the data is thinnest and where our target clubs operate. We regard this as the most important number in the paper after the headline, because it measures the gap between what the system is and what its core market needs it to be. It is the direct motivation for the coverage expansion discussed in §11.

7.5 WHAT THE SEQUENCE MODEL ADDS

The forecaster's margin over baseline (b), a supervised model over the same as-of context, is the cleanest internal measure of what sequence modeling buys: ten points of precision at the head of the list (0.74 against 0.64 at p@50) at equal AUC. Rollout-based trajectory modeling earns its complexity at the head; cohort-wide discrimination is largely available from static features. We report this single-model comparison rather than an ensemble variant. An ensemble improved metrics during development, but its exact combining recipe could not be reproduced at verification time, and by the rules of this paper that excludes it (§8).

7.6 CASE STUDIES

These are illustrations, not evidence; the evidence is §7.1–7.4 and the ledger.

The model's 2021 top-25, with full scores in the released ledger, contains 22 players who realized the top-5% outcome. The list is headed by Pedri, then 18, who realized the 99.6th percentile of his position by 2025. It includes Musiala (ranked 4th, then 18 at Bayern), Mbappé (20th), Saka (22nd), and Bellingham (28th, then 18 at Borussia Dortmund). The three misses are instructive. Two of them, Cristian Romero and Ansu Fati, landed at the 91st and 95th percentile of their positions, near-misses against a hard 95th-percentile line. The third, João Félix, played no qualifying 2025 season under the 600-minute rule and counts as a miss by construction. Among the 199 players who did realize the outcome, the model ranked Haaland 43rd on the strength of his Molde and early Dortmund months, and Gravenberch 33rd while at Jong Ajax.

The other half of the product is bust avoidance. For a typical fringe profile (age 23, no development anchors, weak-club context) the model forecasts a top-5% probability of zero and an 80% probability of leaving covered football within the horizon. No snapshot rating expresses that forecast.

Historical rating trajectories displayed in our product or in illustrations are reconstructed retrospectively and labeled as such. No quantitative claim in this paper rests on reconstructed history.

8. Negative Results and Non-Replications

A validation regime that never kills a claim is not a validation regime. Ours killed three things during the preparation of this paper, and they are reported here rather than omitted.

A non-replicated discrimination estimate. An earlier internal document reported that the undervaluation signal added a material increment of within-tier discrimination (AUC) over a consensus model of club strength, tier, and age. Under the rolling as-of protocol of §7.2, the measured increment across all nine anchors is an order of magnitude smaller, between +0.005 and +0.029 per anchor. The original estimate came from a per-season panel with less careful temporal isolation. This paper therefore makes no discrimination claim for the screen. The selection claim is what replicated, on every anchor, and the selection claim is what we publish.

A contaminated backtest, rejected. A second validation was available cheaply: score the 2021-trained model, backdated, on the 2017 cohort against 2020 outcomes. It produced AUC 0.919. We rejected it, because the 2017–2020 outcome window lies inside that model's training data; a model evaluated on its own training period recalls rather than predicts. The clean replacement, a separate model trained with a 2017 cutoff, scored 0.913 (§7.1). In this instance the contamination inflated the result by only 0.006. But the only way to know that was to train the clean model. We include the episode as a worked example of the question any historical validation should face, ours included: was the model trained before, or after, the period it is being validated on?

Development findings that constrain our claims. Better teacher-forced validation loss did not predict better forecast quality; three architecture variants improved validation loss and degraded the backtest, which leaves the backtest as the only honest model-selection signal for a generative forecaster. Across random seeds, single-run differences of ±0.015 AUC and ±0.04 p@50 are noise at this corpus size, so we avoid claims resting on deltas of that magnitude. And as noted in §7.5, a development ensemble whose recipe could not be reproduced at verification time was dropped from the paper entirely.

9. Pre-Registered Forward Evaluation

Backtests are reconstructions, however carefully controlled (§6.1). The central claim of this paper is therefore converted into a falsifiable public test, as follows.

1. On 11 June 2026 we froze a prediction file: the production model's scored current cohort of 4,021 players under the §3 cohort rules, at information date 1 June 2026 with data through May 2026, carrying each player's ranked probability of reaching the top 5% of his position and his probability of leaving covered football.
2. The SHA-256 digest of the frozen file is `3103caa67e6b64245182778dec31ffd3dbdbb7cfdd6c4c74a2a0e750cd160fcf`. The file itself is held privately until scoring, since publishing the predictions could perturb the market behavior they predict.

3. The first scoring takes place once the 2026/27 season is complete, on 30 June 2027 — a window spanning the two transfer windows that follow publication — using the §3 outcome definitions without modification, and the cohort is re-scored annually thereafter through the 2029/30 season, the full horizon evaluated in §7.1.
4. Within 30 days of each scoring date we publish, at oraca.ai/research: the original frozen file, verifiable against the printed hash; the realized outcomes; and the same metrics as §6.3. This commitment is unconditional. It does not depend on the result.
5. The frozen file cannot be amended. Players with no covered appearance and no rating anchor during a scoring window are reported as exclusions with counts; no other exclusions are permitted.

We commit to this because it is the one evaluation design immune to every concern raised in §6.1, and because a vendor unwilling to be scored forward is asking clubs to take reconstructions on faith.

10. Reproducibility and Data Availability

Three files are released at oraca.ai/research:

career_2021_predictions.csv — all 3,418 players scored at the 2021 cutoff: identity, position and age at cutoff, raw and calibrated probability, model rank, realized within-position percentile at 2025 (null where no qualifying season exists), and the realized outcome label.

career_2017_vintage_predictions.csv — the same for the independently trained 2017 vintage (n = 3,406, outcomes at 2020).

risers_asof_screen.csv — 49,523 player-seasons across the nine screen anchors: the screen's selection flag and both realized move-up outcomes, with players identified by stable UUID. The continuous undervaluation signal itself is proprietary and is not released; the flag and outcomes are sufficient to recompute every published screen result.

Every number in §7 can be recomputed from these files; schemas are in Appendix D. Input data is drawn from commercial and public sources under terms that do not permit redistribution, which is why we release model outputs and evaluation results rather than inputs.

The evaluation protocol of §3, §6, and Appendix A is specified fully, and a party with equivalent data types could implement the harness independently. Before publication, the headline numbers in §7.1–7.3 were also re-derived from the raw stored artifacts by an independent re-implementation of the label and metric code, a second code path written against a different dataframe library. The agreement was exact, system by system.

11. Limitations

Cross-league player comparability is only partially earned. Team strength is globally connected and validated cross-league, but the player-level opponent adjustment operates within league, so cross-league claims at player level rest on the global team scale and the development anchors. A dedicated validation of player-level cross-league transfer has not been completed. Until it is, we scope player-level claims accordingly.

The edge narrows where the clubs are. Discrimination in the lowest league-strength tercile is AUC 0.68 against 0.85 at the top (§7.4). This is the binding constraint for the system's core market and the first target of coverage expansion; our scaling experiments were consistent on the remedy, which is more leagues rather than more parameters, because the corpus rather than model capacity binds.

Coverage defines the model's horizon. A player outside covered football does not exist for the model until he enters it, and visibility is selective: players appear in the data only through covered appearances, so "leaves covered football" conflates retirement, serious injury, and moves to uncovered leagues.

The past is reconstructed. Backtests read today's database, which rebuilds rather than archives history. This is controlled in-pipeline and eliminated only by the forward evaluation of §9. Relatedly, identity resolution across sources is imperfect; a small fraction of player histories may merge or split records. We believe the effect on cohort-level metrics is negligible but have not bounded it formally.

Calibrated probabilities shift across eras. Top-band probabilities transferred from the 2017 era overstate 2021-era frequencies (§7.3). Ranking claims are unaffected; users of the probabilities should apply the stated correction.

Finally, scope. The system models performance development and selection tasks. It uses no wage data and models no contract economics, and its outcome labels (position-relative level, league-tier movement) are proxies for transfer success, not measurements of it.

12. Responsible Use

These are predictions about people's careers. In the product, forecasts are decision support for professional recruitment staff: every recommendation carries its evidence and its uncertainty, and the system is designed to widen the set of players a club seriously considers rather than to automate exclusion. Forecasts are surfaced for professional players within recruitment-relevant age ranges. We do not publish individual fore-

casts outside the professional recruitment context, and the released ledger contains historical evaluation cohorts only.

The coverage asymmetries of §4 imply a systematic risk worth stating in plain terms: players from sparsely covered competitions can be under-ranked relative to their true ability. Human users should weight this, and coverage expansion is the mitigation.

13. Conclusion

Scouting is forecasting, and forecasting can be held to a standard: train before the cutoff, score after it, publish the predictions, and come back when the future arrives. Under that standard, a foundation model over career sequences turned four years of foresight into 37 hits in a top-50 list where the best alternative managed 32 and chance would manage three; a fully disclosed undervaluation screen selected players who moved up at twice the base rate, nine years running. We have said where the system is weak, reported what our own validation killed, and hashed our next set of predictions. This, rather than accuracy theater, is what we think trust in football analytics should look like, and the standard is open to anyone — including our competitors.

References

[1] Elo, A. (1978). *The Rating of Chessplayers, Past and Present*. Arco. [2] Hvattum, L. M., & Arntzen, H. (2010). Using ELO ratings for match result prediction in association football. *International Journal of Forecasting*, 26(3), 460–470. [3] Decroos, T., Bransen, L., Van Haaren, J., & Davis, J. (2019). Actions speak louder than goals: Valuing player actions in soccer. *KDD 2019*, 1851–1861. [4] Hvattum, L. M. (2019). A comprehensive review of plus-minus ratings for evaluating individual players in team sports. *International Journal of Computer Science in Sport*, 18(1), 1–23. [5] Dendir, S. (2016). When do soccer players peak? A note. *Journal of Sports Analytics*, 2(2), 89–105. [6] Vaswani, A., et al. (2017). Attention is all you need. *NeurIPS 2017*, 5998–6008. [7] Radford, A., et al. (2018). Improving language understanding by generative pre-training. OpenAI. [8] Nie, Y., Nguyen, N. H., Sinthong, P., & Kalagnanam, J. (2023). A time series is worth 64 words: Long-term forecasting with transformers. *ICLR 2023*. [9] Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1), 1–26. [10] Zadrozny, B., & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. *KDD 2002*, 694–699. [11] Gebru, T., et al. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86–92. [12] Mitchell, M., et al. (2019). Model cards for model reporting. *FAT* 2019*, 220–229.

Appendix A — Exact Definitions

Cohort (T2), at cutoff (T): every player with (i) age < 24.0 years at (T), date-of-birth based; (ii) at least 600 career minutes in covered competitions at (T); (iii) non-zero covered minutes in the recent activity window at (T). Primary cutoff 2021-06-30: n = 3,418.

Outcome label (T2), at season (S): among all players with an observed development anchor at (S) and season minutes ≥ 600 (season minutes being the cumulative-minutes difference against the player's previous anchor), rank the level estimate within position, using the position at (S), average rank for ties, and the count of validly rated players as denominator. The label is a within-position percentile of at least 0.95. Cohort players without a qualifying anchor at (S) are labeled negative — a conservative choice that counts disappearance against the model.

Position assignment: the season-specific position if observed at (S); otherwise the most recent observed position at or before (S); otherwise a static position bucket. Players with no mappable position form their own "UNK" ranking group (18% of the 2021 cohort; see the note in Appendix C).

Undervaluation signal (T3), at anchor (T): `under = level − E[level | club rating, league tier, age]`, where the expectation is a least-squares fit over regular players at (T), regularity being a recent-minutes threshold at (T), and all inputs as-of (T). The cohort is players under 26, regular, non-goalkeeper, with observable context at (T) and (T{+}1). The flag is the top 5% of `under` within the cohort.

Move-up outcomes (T3): *tier* — the league tier at (T{+}1) is strictly better than at (T); *transfer* — the club at (T{+}1) differs from the club at (T), and is better by tier or higher by club rating.

Metrics: precision@k is the fraction of label-positives among the top k by score under a stable sort; lift@k is precision@k divided by the cohort base rate; AUC is the usual rank statistic over the full cohort; intervals are percentile bootstrap over 1,000 cohort resamples.

Appendix B — Coverage by Competition

34 competitions: 30 domestic leagues and four UEFA club competitions. Columns: first covered season, first season with detailed in-match statistics, appearance records, distinct players. Source: the released `coverage_by_league.csv`.

Competition	Country	First season	Detail from	Appearances	Players
Serie A	Italy	2005	2010	333,422	4,959
Championship	England	2010	2013	324,973	4,901
League Two	England	2009	2017	310,797	5,437
League One	England	2008	2016	310,328	5,317
Premier League	England	2004	2006	299,456	4,370
La Liga 2	Spain	2008	2013	297,300	4,961
La Liga	Spain	2005	2012	291,434	4,351
Serie B	Italy	2008	2018	278,437	4,731
Süper Lig	Türkiye	2005	2014	257,368	5,009
Ligue 1	France	2006	2012	244,972	4,300
Bundesliga	Germany	2005	2012	238,048	3,734
Major League Soccer	United States	2009	2013	222,856	3,686
Liga Portugal 2	Portugal	2010	2023	203,378	4,426
Eredivisie	Netherlands	2009	2012	201,521	3,581
Ligue 2	France	2010	2014	187,259	3,859
J1 League	Japan	2009	2014	185,341	2,414
2. Bundesliga	Germany	2010	2016	183,764	3,543
Eerste Divisie	Netherlands	2010	2016	178,138	3,663
Pro League	Belgium	2009	2020	175,491	3,458
Liga Portugal	Portugal	2009	2014	163,484	3,711
Eliteserien	Norway	2010	2014	135,807	2,626
Allsvenskan	Sweden	2010	2016	130,407	2,636
K League 1	South Korea	2011	2020	126,544	2,177
Superliga	Denmark	2010	2017	120,985	2,231
Veikkausliiga	Finland	2010	2023	103,727	2,285
Europa League	Europe	2008	2012	92,850	13,361
Champions League	Europe	2005	2006	92,441	9,231
Premier Division	Rep. of Ireland	2013	2022	90,688	1,922
A-League Men	Australia	2009	2012	77,421	1,577
Championship	Scotland	2014	2024	72,727	1,717
Challenger Pro League	Belgium	2017	2022	56,253	1,902
Europa Conference League	Europe	2021	2021	22,013	3,768
Premiership	Scotland	2025	2025	9,111	439
UEFA Super Cup	Europe	2010	2017	640	471

Totals: 6,019,381 appearance records and 71,571 distinct players in the full corpus through May 2026. The 2021-cutoff training corpus: 53,353 players and 2,733,107 player-months.

Appendix C — Per-Anchor Screen Results and Slices

Undervaluation screen, tier outcome (primary):

Anchor	Cohort	Flagged	Base rate	Flagged rate	Lift	ΔAUC of signal*
2016→2017	4,919	246	0.078	0.159	2.02×	+0.005
2017→2018	5,156	258	0.085	0.112	1.32×	+0.014
2018→2019	5,158	258	0.091	0.198	2.17×	+0.017
2019→2020	5,036	252	0.117	0.234	2.00×	+0.005
2020→2021	5,537	277	0.073	0.126	1.74×	+0.006
2021→2022	5,745	287	0.089	0.167	1.89×	+0.018
2022→2023	5,823	291	0.095	0.227	2.38×	+0.021
2023→2024	5,992	300	0.100	0.147	1.47×	+0.006
2024→2025	6,157	308	0.099	0.237	2.39×	+0.027
Pooled	**49,523**	**2,477**	—	—	**1.95×**	—

* The incremental AUC of adding the signal to a club-rating/tier/age consensus model, reported for transparency of the §8 non-replication. The paper claims no discrimination effect. The 2024→2025 outcome season was approximately 95% complete at evaluation. The transfer-mode table is in the released ledger.

Shortlist slices (2021 cohort, raw ranking). By position: AUC from 0.78 (wingers) to 0.92 (defensive midfielders) over the eight mapped positions, with top-5%-of-slice precision between 0.33 and 0.52 on base rates of 5–9%. The UNK group (n = 616, base rate 1.9%) is included in the ledger, but its within-UNK ranking denominator makes its slice metrics hard to interpret, and we do not lean on them. By age band: under-19, AUC 0.747 with slice-head precision 0.75 on a 19.5% base rate; 19–21, 0.843 and 0.50 on 7.8%; 21–24, 0.875 and 0.41 on 4.8%. League-strength terciles are in §7.4.

Appendix D — Released File Schemas

`career_2021_predictions.csv` (3,418 rows) and `career_2017_vintage_predictions.csv` (3,406 rows): `player_id` (stable UUID), `name`, `position_at_cutoff`, `age_at_cutoff`, `p_top5_raw`, `p_top5_calibrated` (2021 file only), `model_rank`, `realized_pct_{2025|2020}` (within-position percentile; empty where no qualifying season exists), `realized_top5_{2025|2020}` (boolean).

`risers_asof_screen.csv` (49,523 rows): `anchor_season`, `player_id` (stable UUID), `pos`, `age`, `tier` (league tier at anchor), `flagged_top5` (boolean: in the top 5% by the undervaluation signal), `moveup_tier`, `moveup_transfer` (0/1 outcomes).

Appendix E — Model Card (Summary)

Model. Decoder-only transformer over monthly player-career sequences; multi-channel next-month prediction; Monte Carlo rollout forecasts; isotonic-calibrated probabilities. **Intended use.** Decision support for professional football recruitment — shortlist generation, development forecasting, undervaluation screening, bust-risk assessment — alongside human scouting, not in place of it. **Out of scope.** Automated player exclusion; wage or contract decisions; players outside covered competitions; non-professional contexts. **Training data.** 30 domestic leagues and four UEFA competitions, 2005–present (Appendix B), with the coverage asymmetries described in §4 and §11. **Evaluation.** Two clean as-of vintages (2017, 2021), a nine-anchor screen evaluation, and the pre-registered forward test of §9. **Known biases and failure modes.** Underranking of players from sparsely covered competitions; degraded discrimination in the lowest-covered tiers (AUC 0.68 against 0.85); top-band probability overconfidence under era shift; residual identity-resolution error.