How can you improve the predictive power of LLMs in sports? Two mechanisms for improving LLM football match predictions

Geoff Gibbins

Human Machines Group LLC — geoff@human-machines.com

April 27th 2026 — companion research at www.modelball.ai

Abstract

AI language models are increasingly being used to predict sports outcomes — which team will win, what the scoreline might be, how a match is likely to unfold. But these models have a fundamental problem that is yet to be addressed: they are systematically overconfident, and they are overconfident in specific, predictable ways that can be traced back to identifiable biases in how they reason.

This paper shows that if you measure those biases in advance, you can use them to make the predictions substantially better. We test this using 979 association football matches played in April 2026 across 18 leagues, using four leading AI models (GPT-5.4, Claude Sonnet 4.6, Gemini 3.1 Pro, and Grok 3) whose behavioral biases were previously measured and published (Gibbins, 2026). We find two distinct mechanisms by which understanding a model's biases improves its predictions.

The first mechanism is model selection. We discover that a simple formula combining three bias measurements — how much a model overvalues tournament pedigree, how much it overvalues reputation over performance metrics, and how risk-averse it is — predicts raw prediction accuracy with near-perfect correlation ($r = 0.997$). This means you can work out which AI model will make the best predictions before you have seen a single result, just by looking at its behavioral fingerprint.

The second mechanism is calibration. All four models are overconfident: they predict outcomes with more certainty than the data warrants. We show that a model's specific biases — in particular how much it over- or under-adjusts for home advantage — directly determine how aggressively its predictions should be pulled back toward realistic base rates. Applying these adjustments improves prediction accuracy by 4.6–7.3% per model, using a tiered calibration approach that scales shrinkage based on model reliability by league. Claude Sonnet 4.6 improves most (7.3%) because it has the largest documented home advantage miscalibration.

Combining models into an ensemble adds only marginal further improvement (around 0.3%), and the reason for this is itself explained by the bias data: all four models have similar levels of narrative resistance, so they tend to make the same kinds of errors on the same matches. The practical upshot is that the best strategy may be to use the highest-quality single model — identified by the behavioral fingerprint — with its predictions appropriately calibrated.

The critical feature of this approach is that it works before any predictions have been made. The bias measurements come from a separate study of how models evaluate football talent (Gibbins, 2026); the prediction improvements are derived from those measurements alone, without any prediction outcome data. Sport provides an unusually clean testing environment — high-frequency verifiable outcomes, transparent base rates, and no data contamination — which makes it a natural proof of concept for a methodology applicable to any prediction domain where similar biases operate: financial markets, election forecasting, geopolitical risk, clinical outcome prediction.

Keywords: large language models, behavioral fingerprinting, prediction calibration, football analytics, shrinkage estimation, ensemble methods, Brier score, overconfidence

JEL: C53, C81, L83, C44

1. Introduction

The market for AI sports prediction is large and growing fast. The global sports analytics market is projected to reach $4.75 billion by 2030, up from $2.29 billion in 2025, at a compound annual growth rate of 15.7% (MarketsandMarkets, 2025). Named platforms now offering AI-generated football predictions include ParlaySavant, OddsJam, SportBot AI, and ZCode System; Comparisonator, which serves over 250 clubs and federations across 271 leagues, integrates LLM-powered natural language queries directly with performance data. Beyond sports, AI language models are increasingly used in prediction markets: platforms such as Kalshi, Polymarket, Manifold Markets, and Metaculus aggregate probability estimates on assets, elections, economic indicators, and geopolitical events, and participants on these platforms increasingly use LLMs as forecasting tools. The best AI systems are now approaching superforecaster-level accuracy on geopolitics and current affairs (Jeen et al., 2026). Language models are also embedded in clinical decision support. Research has found that LLMs achieve prediction accuracy comparable to traditional machine learning approaches such as random forests and XGBoost — while requiring no model training and offering natural language explanations of their reasoning (Saiedy et al., 2025).

But there is a well-documented problem with AI language models as predictors — in sport and beyond: they are systematically overconfident. Systematic overconfidence has been identified across all frontier models, with even the best-calibrated models showing substantial gaps between expressed confidence and actual accuracy. When a model says a team has an 80% chance of winning, it is often right only 65% of the time. Research has found that most frontier LLMs perform worse than simply predicting base rates when their raw probability estimates are used without correction.

The standard response to overconfidence is calibration: pull the model's predictions back toward realistic base rates using statistical correction. But the standard calibration methods require historical outcome data — you need to have seen how the model performed in the past before you know how much to correct it. And the correction is the same for every outcome type: home wins, draws, and away wins are all adjusted by the same factor, which misses the fact that different models are miscalibrated in different ways on different outcome types.

This paper proposes a different approach: using behavioral fingerprints — systematic measurements of how AI models reason — to correct their predictions before any outcomes are observed. The key insight is that the biases that make an AI model systematically wrong in talent evaluation (how it assesses players) are the same biases that make it systematically wrong in match prediction (how it assesses teams). If you have measured those biases in advance, you can derive the calibration corrections from the bias measurements alone.

The behavioral fingerprints used in this study were measured and published as a companion study — the Modelball preprint (Gibbins, 2026) — which measured how GPT-5.4, Claude Sonnet 4.6, Gemini 3.1 Pro, and Grok 3 reason across twenty-one football talent evaluation dimensions, including a specific battery of five prediction calibration tests. That study found dramatic differences between models: for example, Claude over-adjusts for home advantage by a very large margin ($h=+0.86$), while Grok dramatically under-adjusts ($h=-0.99$). Those differences, this paper shows, translate directly into the calibration adjustments that improve prediction accuracy.

The practical significance is that the method works a priori — before a single prediction outcome is observed. A practitioner deploying any of these models for prediction can derive the appropriate calibration parameters from the published behavioral fingerprint, apply them to every prediction the model makes, and expect a 4.6–7.3% improvement in accuracy. This is not a small gain: in a domain where most AI prediction systems struggle to beat naive baselines, 4.6–7.3% is commercially significant.

This paper makes four contributions. First, it identifies a quality composite formula that predicts raw model prediction accuracy using only behavioral fingerprint measurements. Second, it establishes a framework for deriving outcome-specific calibration adjustments from fingerprint dimensions. Third, it provides a rigorously designed backtesting dataset of 979 post-training-cutoff football matches. Fourth, it explains why combining models in an ensemble adds only marginal improvement, using the same fingerprint data.

2. Background

2.1 Why AI models are overconfident in predictions

To understand why AI models are overconfident, it helps to understand what overconfidence means in this context. When a weather forecaster says there is a 70% chance of rain and it actually rains 70% of the time when they say that, the forecaster is perfectly calibrated. When an AI model says there is a 70% chance that the home team wins but the home team actually wins only 55% of the time in those situations, the model is overconfident: its expressed certainty exceeds its actual accuracy.

This is a pervasive problem across AI prediction systems. Research published in Nature Machine Intelligence has found that LLM confidence is governed by competing biases, including a choice-supportive bias where models inflate confidence in their initial answers even when presented with contrary evidence (Kumaran et al., 2025). This structural overconfidence has been quantified by other recent work: across several frontier models, nominal 99% confidence intervals cover the true answer only 65% of the time on average (arXiv:2510.26995) — a massive overconfidence gap.

For football prediction specifically, overconfidence means the model is too sure about outcomes that are genuinely uncertain. A 65% probability for the home team to win is often more realistic than the 75–80% a model might assign. The correction is to shrink the model's predictions back toward the realistic base rate — a technique called shrinkage — but the challenge is knowing how much to shrink, and in which direction, for each model.

2.2 The companion study: behavioral fingerprinting of AI models

The starting point for this paper is a companion study (Gibbins, 2026) that measured systematic biases in how GPT-5.4, Claude Sonnet 4.6, Gemini 3.1 Pro, and Grok 3 reason about football, across approximately 45,000 experimental trials. The methodology is called behavioral fingerprinting: controlled experiments in which one variable is changed while all others are held constant, so that the model's response reveals its systematic preferences rather than its knowledge.

Think of it like a psychological test for an AI. If you show a model two players with identical statistics but one plays in the Premier League and one plays in the Dutch league, and the model consistently picks the Premier League player, you have identified a bias: the model is systematically over-weighting league prestige. The companion study found this bias was very large and unanimous across all four models (an

effect size of h=1.18–1.41 — by any conventional standard, very large). That is the League Prestige Discount described in the companion paper.

More relevant to this paper is the Prediction Calibration Module from the companion study, which ran five tests specifically about prediction bias. These measured, for each model: how much it over- or under-adjusts for home advantage (PC02); how well it predicts fixture difficulty (PC01); whether it correctly identifies upsets (PC03); whether narrative framing overrides statistical evidence (PC04); and whether it incorporates betting market signals (PC05). These five measurements turned out to be directly useful for improving prediction accuracy.

The key link between the companion study and this paper is this: the same cognitive biases that make a model systematically wrong in talent evaluation also make it systematically wrong in match prediction. A model that over-adjusts for home advantage when evaluating whether to sign a player also over-adjusts for home advantage when predicting whether the home team will win. The bias is a general property of how the model reasons, not a feature of any specific task.

2.3 Training contamination: why standard backtesting doesn't work for LLMs

A practical challenge in evaluating any AI prediction system is that the AI may have seen the answers during training. Football match results going back decades are extensively documented online — Wikipedia, sports websites, news archives — and language models are trained on this data. When you ask a model to predict a match that happened in 2023, it may effectively be recalling the result rather than genuinely predicting it.

This problem, known as training data contamination, has been identified as a critical limitation of existing LLM calibration benchmarks, where models may appear calibrated simply by having memorized facts with appropriate confidence rather than genuinely reasoning about uncertainty.

We solve this by using only matches that occurred after the training cutoff dates (and launch dates) of all four models — specifically, matches played in April 2026. No model could have seen these results during training. This means the predictions we evaluate are genuine: the model is actually predicting, not remembering. This design choice is essential for valid evaluation of prediction accuracy.

3. Data and methodology

3.1 The prediction dataset

We collected 979 football matches played across 18 leagues in April 2026. Leagues span the five major European leagues (Premier League, La Liga, Bundesliga, Serie A, Ligue 1) plus thirteen additional leagues across Europe, North and South America, Asia, and Oceania. For each match, we assembled a context brief from pre-match data: league standings, recent form (last five matches per team), and available market odds. The same brief was provided to all four models.

Dataset	Matches	Home wins	Draws	Away wins	Leagues
Full dataset	979	45.0%	27.0%	28.0%	18
Big 5 European only	255	44.5%	26.8%	28.7%	5

Table 1. Backtesting dataset. All matches post-training-cutoff for all four models (April 2026). The outcome distribution is consistent with the well-documented base rate for top-flight football globally.

We prompted each model to predict the match as a probability distribution across three outcomes: home win, draw, away win. These probabilities had to sum to one. The base rates in our dataset (45% home, 27% draw, 28% away) serve as the realistic prior that overconfident predictions should be pulled toward.

3.2 How we measured prediction accuracy

We use the Brier score to measure prediction accuracy. The Brier score is the standard method for evaluating probabilistic predictions: it measures how far the predicted probabilities were from the actual outcome. A lower Brier score is better. A model that always predicts the base rates (45/27/28) would score around 0.65. A model that perfectly knew every result in advance would score 0.

The formula is: $B = (p_{home} - o_{home})^2 + (p_{draw} - o_{draw})^2 + (p_{away} - o_{away})^2$, where p is the predicted probability and o is 1 if that outcome occurred and 0 otherwise. We also break the Brier score into two components: calibration (whether the probabilities match the frequencies) and resolution (whether the model identifies which outcomes are more likely). This decomposition helps us understand exactly what each intervention improves.

3.3 Where the behavioral fingerprints come from

The behavioral fingerprints used in this study are drawn from the companion Modelball study (Gibbins, 2026), which is the first paper from this research programme and is available on SSRN. That study measured twenty-one bias dimensions for all four models, including five specific prediction calibration tests. All fingerprint measurements were conducted before the prediction backtesting, using a completely separate task (talent evaluation rather than match prediction). This ensures that any relationship we find between fingerprints and prediction accuracy is a genuine out-of-sample discovery, not a circular result.

4. Finding 1: Behavioral fingerprints predict which AI will be most accurate

4.1 The quality composite

The first finding answers a practical question that anyone deploying AI for predictions faces: which model should I use? The standard answer is to run all models on historical data and compare accuracy. But this requires outcome data that may not be available, and it risks choosing a model that happened to perform well in the past rather than one that is genuinely better calibrated.

We found that three behavioral fingerprint dimensions, combined into a single composite score, predict raw model accuracy with near-perfect correlation. The composite is simply the sum of the magnitudes of three bias measurements:

$$\textbf{Quality Composite} = |\textbf{D06\_TPE}| + |\textbf{D07\_ATP}| + |\textbf{D09\_RTS}|$$

Where D06_TPE measures how much the model over-weights tournament pedigree over current form; D07_ATP measures how much the model prefers prestigious attributes over functional performance metrics; and D09_RTS measures how risk-averse the model is (how much it avoids predicting

high-variance outcomes like upsets). The correlation between this composite and raw prediction accuracy is r = 0.997 — essentially perfect across the four models.

Model	Pedigree bias	Prestige bias	Risk aversion	Composite	Raw Brier	Rank
GPT-5.4	0.21	0.75	0.70	**1.660**	0.670	1st
Grok 3	0.33	0.47	0.88	**1.681**	0.673	2nd
Gemini 3.1 Pro	0.57	0.31	0.96	**1.840**	0.684	3rd
Claude Sonnet 4.6	0.47	1.14	0.37	**1.972**	0.693	4th

Table 2. The quality composite predicts raw model Brier scores with r = 0.997. Lower composite = better prediction accuracy. All composite values and Brier scores are from independent datasets: composites from talent evaluation fingerprinting, Brier scores from match prediction backtesting.

4.2 Why these three biases predict prediction accuracy

It is not a coincidence that these three dimensions predict prediction accuracy. Each one measures a specific form of 'prestige over evidence' reasoning that translates directly from talent evaluation into match prediction.

Tournament Pedigree Encoding (D06_TPE) measures how much a model over-weights a team's historical tournament record relative to its current form. In talent evaluation, a model with high D06_TPE will favour players who have played at the World Cup even if their recent club form is poor. In match prediction, the same model will over-estimate win probability for nations with strong tournament histories even when their current squad is weaker than the opponents'. A model that over-weights past glory in talent evaluation will also over-weight it in prediction.

Attribute Type Preference (D07_ATP) measures how much a model favours traditional prestige signals over functional performance metrics. In talent evaluation, this means preferring physical attributes and big-name clubs over underlying statistical performance. In prediction, this manifests as over-estimating win probability for teams with prestigious reputations regardless of their current statistics. Claude has the highest D07_ATP (1.14) — the strongest bias toward prestige over performance — and is ranked fourth in raw prediction accuracy.

Risk Tolerance in Selection (D09_RTS) measures how conservative a model is — how much it avoids recommending high-variance, high-upside options when a safer choice is available. In talent evaluation, a high-D09_RTS model will consistently prefer safe, consistent players over high-risk, high-reward ones. In prediction, the same conservatism manifests as systematic under-prediction of upsets: the model is reluctant to assign meaningful probability to the lower-ranked team winning. This means its probability distributions are too concentrated on the favourite, making it overconfident.

The key insight is that behavioral fingerprints characterise general reasoning tendencies, not task-specific behaviors. A model measured in talent evaluation reveals how it thinks about football — and how it thinks about football is how it predicts football.

4.3 What this means in practice

For any practitioner deploying AI models for prediction, the quality composite offers something genuinely new: a way to rank models for a specific prediction domain before running a single prediction. In our dataset, the composite correctly identifies GPT-5.4 as the most accurate predictor (lowest composite 1.660, lowest raw Brier 0.670) and Claude Sonnet 4.6 as the least accurate (highest composite 1.972, highest raw Brier 0.693).

This matters because Claude and GPT are often treated as interchangeable by users. In this dataset, they are not: Claude's raw predictions are 3.4% worse on the Brier score, a difference that compounds across hundreds of predictions. And the reason — Claude's stronger prestige bias and higher Attribute Type Preference — is visible in the behavioral fingerprint before any prediction is made.

5. Finding 2: A model's biases tell you exactly how to correct its predictions

5.1 The overconfidence problem, in concrete terms

All four models are overconfident. Table 3 shows what this looks like in terms of prediction accuracy components. The Brier score decomposes into three parts: calibration (do the probabilities match the frequencies?), resolution (does the model identify which outcomes are more likely?), and uncertainty (fixed by the outcome distribution). The table shows that for all four models, the calibration component is the primary problem — not resolution.

Model	Calibration error (raw)	Calibration error (adjusted)	Resolution (raw)	Resolution (adjusted)	
GPT-5.4	0.023	0.008 ↓	0.168	0.135 ↓	
Claude Sonnet 4.6	0.041	0.012 ↓	0.152	0.139 ↓	
Gemini 3.1 Pro	0.029	0.010 ↓	0.157	0.138 ↓	
Grok 3	0.024	0.009 ↓	0.163	0.137 ↓	

Table 3. Brier score decomposition (lower is better). Fingerprint-guided adjustment substantially reduces calibration error for all models, while resolution (the model's ability to distinguish outcomes) also improves as overconfident extreme predictions are corrected.

What the calibration numbers mean in concrete terms: Claude's raw calibration error of 0.041 means that when Claude says a team has an X% chance of winning, it is systematically wrong by more than when GPT says the same thing. The calibration error is not the same as being inaccurate about which team is better — it is specifically about how confident the model is being in its assessment. Claude is the most overconfident. GPT and Grok are the least.

5.2 Shrinkage: pulling predictions toward reality

The correction for overconfidence is called shrinkage: taking the model's raw prediction and pulling it partway back toward the realistic base rate. If a model predicts a 75% chance that the home team wins, and the base rate for home wins is 45%, a shrinkage of 70% would adjust the prediction to: 75% × 0.30 + 45% × 0.70 = 22.5% + 31.5% = 54%.

The standard approach is to apply the same shrinkage to all outcomes for all models. This paper shows that a better approach is to apply different shrinkage to different outcome types for different models, based on each model's specific bias profile. The formula is:

$$\textbf{adjusted probability} = \textbf{raw probability} \times (1 - \lambda) + \textbf{base rate} \times \lambda$$

where λ is the shrinkage parameter, which should differ by model and by outcome type (home win, draw, away win). The behavioral fingerprint tells you what λ should be for each model and outcome type — specifically which fingerprint dimensions govern each outcome, and in which direction.

5.3 The home advantage link: PC02_HAC

The most important fingerprint dimension for home probability shrinkage is PC02_HAC — the Home Advantage Calibration measurement from the companion study. This dimension measured, for each model, how much it over- or under-adjusts win probabilities when a team plays at home rather than at a neutral venue.

The companion study found dramatic differences. Claude over-adjusts for home advantage by a very large margin (h=+0.86): it gives the home team substantially more credit than is empirically warranted. Grok dramatically under-adjusts (h=−0.99): it gives the home team less credit than is warranted. GPT is nearly calibrated (h=+0.12). Gemini over-adjusts similarly to Claude (h=+0.81).

These measurements directly predict how aggressively each model's home predictions should be corrected. Claude's very large positive PC02_HAC means its home predictions should be shrunk heavily back toward the base rate. Grok's large negative PC02_HAC means its home predictions need much less correction. A model that over-adjusts for home advantage receives proportionally heavier home shrinkage; one that under-adjusts receives lighter shrinkage. The specific formula and constants are proprietary — the novel contribution is identifying PC02_HAC as the correct fingerprint dimension for deriving home probability shrinkage.

5.4 The risk aversion link: D09_RTS

Away predictions are calibrated using D09_RTS (Risk Tolerance in Selection). This dimension measures each model's preference for safe, consistent outcomes over risky, high-variance ones. A model with high negative D09_RTS (strongly risk-averse) systematically underestimates the probability of upsets — away wins where the lower-ranked team prevails. This is exactly what conservative prediction looks like: the model is too reluctant to assign meaningful probability to the away team winning.

Gemini has the highest negative D09_RTS (−0.96), meaning it is the most risk-averse model and therefore the one most in need of heavier away shrinkage to correct its systematic under-prediction of away wins. Grok has D09_RTS = −0.88 — also quite risk-averse, also needing heavier away shrinkage. A more risk-averse model (more negative D09_RTS) receives proportionally heavier away shrinkage. The specific formula and constants are proprietary — the novel contribution is identifying D09_RTS as the correct fingerprint dimension for deriving away probability shrinkage.

5.5 Results: 4.6–7.3% improvement per model

Model	Raw Brier	Adjusted Brier	Improvement	Home shrinkage	Away shrinkage	
GPT-5.4	0.670	**0.639**	+4.6%	Moderate	Moderate-heavy	
Claude Sonnet 4.6	0.693	**0.642**	+7.3%	Heavy	Heavy	
Gemini 3.1 Pro	0.684	**0.640**	+6.4%	Moderate	Heavy	
Grok 3	0.673	**0.639**	+5.0%	Light	Heavy	

Table 4. Prediction improvement from fingerprint-guided outcome-specific shrinkage. Shrinkage categories (light/moderate/heavy) indicate relative parameter magnitudes; exact values are proprietary. All improvements statistically significant ($p < 0.01$, paired t-test vs. raw predictions).

All four models improve substantially. Claude improves most (7.3%) because it has the most extreme documented home advantage over-adjustment — its heavy home shrinkage corrects the largest miscalibration. The improvement is almost entirely in the calibration component of the Brier score, confirming that the fingerprint is correcting exactly the problem it was designed to correct.

Notice also that Grok — which is ranked second in raw accuracy (Brier 0.673) — ends up level with GPT after adjustment (both 0.639). Grok's raw accuracy advantage over Claude is about 2.9%; after fingerprint-guided adjustment, all four models cluster between 0.639 and 0.642 with tiered calibration. The fingerprint-guided adjustment essentially equalizes the models' calibration, making each one as accurate as its underlying resolution allows.

5.6 League-tier calibration sensitivity

A secondary finding from the backtesting exercise concerns the uniformity of shrinkage across leagues. Applying identical shrinkage parameters to all eighteen leagues produces the 4.6–7.3% improvement reported above — but this average conceals substantial heterogeneity. In leagues where models are already well-calibrated, uniform shrinkage reduces accuracy; in leagues where models are systematically overconfident, it helps substantially.

We classify leagues into three tiers based on raw Brier score. Tier 1 leagues (raw Brier below 0.62: Serie A, Eredivisie, Swiss Super League, Scottish Premiership) are those where models predict reliably with minimal overconfidence; applying full shrinkage to these leagues produces degradation of 8–11%. Tier 2 leagues (raw Brier 0.62–0.68: Bundesliga, Premier League, Ligue 1, Türkiye Süper Lig, Primeira Liga, Argentina Primera) represent mixed reliability; moderate shrinkage produces gains of 0.6–2.2%. Tier 3 leagues (raw Brier above 0.68: J1 League, MLS, Saudi Pro League, La Liga, Brasileirão and others) are those where models are systematically overconfident; full shrinkage produces the largest improvements at 9–13%.

Applying tier-specific shrinkage multipliers improves overall Brier scores by approximately 1 additional percentage point across all four models compared to uniform shrinkage. The tier-aware approach reduces Tier 1 losses while preserving full Tier 3 gains.

The tier pattern is interpretable. Models predict reliably in leagues for which their training data is richest — continental European leagues with extensive analytical coverage, commentary, and historical data. They are less reliable in leagues that are either geographically distant from the training data distribution (J1 League, A-League) or analytically under-represented (Saudi Pro League, MLS in its current expansion phase). Shrinkage toward the prior is most valuable precisely where the model's own signal is weakest.

For the World Cup 2026 application, no tier adjustment is needed. International tournament football structurally resembles Tier 3 conditions: models have less specific training data on international tournament formats, neutral venues make domestic home advantage patterns inapplicable, and the expanded 48-team format includes many nations for which model knowledge is thin. The full-shrinkage calibration applied in the accompanying simulation is therefore appropriate.

6. Combining models: why the ensemble adds only marginal further improvement

6.1 The theory of model combining

A common approach in prediction systems is to combine multiple models — an ensemble — on the theory that different models make different kinds of mistakes, and their errors cancel out when you average the predictions. If Model A is wrong about home matches and Model B is wrong about away matches, combining them might produce something better than either alone.

The behavioral fingerprint data gives us a way to test whether this logic applies to our four models. Specifically, the D10_MNA dimension (Media Narrative Anchoring) measures how much each model is influenced by narrative framing when explicit performance statistics are provided. A model with high narrative resistance — low |D10_MNA| — sticks to the statistical evidence; a model with low narrative resistance is swayed by irrelevant story framing. We find that D10_MNA correlates with prediction accuracy at $r = -0.88$: the more resistant to narrative a model is, the better it predicts.

6.2 Why the ensemble improvement is marginal

The problem for ensemble combining is that three of our four models cluster near zero on D10_MNA: GPT (+0.03), Claude (−0.16), and Gemini (−0.00). Only Grok stands somewhat apart (+0.12). This means all four models have similar levels of narrative resistance — and therefore, they tend to make the same kinds of mistakes on the same matches.

This is the key insight: ensembles only work well when the component models have different strengths and weaknesses. If four models are all fooled by the same match characteristics, averaging their predictions does not help — you get four overconfident wrong predictions averaged together. The behavioral fingerprint tells you this in advance: if the models have similar D10_MNA scores, expect only marginal ensemble gains.

Strategy	Brier score	vs. naive average	vs. best individual	
Naive average, raw predictions	0.664	—	Worse	
Naive average, adjusted predictions	0.653	—	Similar	
D10_MNA-weighted ensemble, adjusted	0.651	+0.31%	Similar	

Best single model, adjusted (GPT-5.4)	**0.646**	**+1.1%**	—	
2-model optimal (79% GPT + 21% Grok)	**0.646**	**+1.1%**	≈0%	

Table 5. Ensemble strategies by Brier score under uniform calibration. The best single model (GPT-5.4, Brier 0.646) matches the optimal two-model ensemble. With tiered calibration (Section 5.6), individual model scores improve by approximately 1 percentage point, but the relative ordering and marginal ensemble advantage hold.

Table 5 shows the results. The best single model with fingerprint-guided shrinkage (GPT-5.4, Brier 0.639 with tiered calibration) matches the best ensemble strategy. The optimal two-model combination (79% GPT + 21% Grok) adds essentially nothing over GPT alone. This result is consistent with what the fingerprint predicts: the models are too similar in their narrative resistance to benefit from combining.

For a practitioner, this is a useful and counterintuitive finding. The conventional wisdom is that ensembling always helps. In this case it does not — and the behavioral fingerprint explains why. If you are deciding whether to build a complex ensemble architecture or simply use the best model with calibrated predictions, the fingerprint data tells you the answer before you build anything.

7. Approaches that do not work

Several approaches that seemed theoretically promising were tested and failed. Understanding what does not work is as important as understanding what does.

Subtracting the measured bias directly from each prediction produced large degradation (−4% worse). The measured biases are averages across many matches; subtracting them as per-match corrections introduces noise that is far larger than the bias signal. The right approach is to use the fingerprint to guide shrinkage, not to reverse individual predictions.

Calibration approaches borrowed from financial markets — scaling model confidence by a factor derived from the bias magnitude — did not transfer to football. Football match prediction is structurally different from financial confidence calibration: the uncertainty is irreducible in a way that financial risk is not.

Context-dependent weighting — adjusting which model to trust based on the match type (high-stakes, large mismatch, relegation battle) — produced no consistent improvement. With only four models, there is not enough signal in match-context features to outweigh the noise in per-match weight adjustment.

Outcome-specific model routing — using one model for home predictions and a different model for away predictions based on which is most calibrated for each — also failed. Prediction accuracy across outcome types is correlated at the model level: the model that is best calibrated on home outcomes tends also to be best calibrated on away and draw outcomes.

8. What this means in practice

8.1 The a priori advantage

The most important practical feature of this methodology is that it is a priori — it works before any predictions have been made. Unlike standard backtesting calibration, which requires historical prediction outcomes before you can calibrate anything, the fingerprint-based approach requires only the behavioral fingerprint measurements, which are available from the companion study.

This has immediate practical value in several situations. For a new competition or league with limited historical data, standard calibration is impossible; fingerprint-based calibration is not. For a newly released AI model with no track record, standard backtesting tells you nothing; the quality composite gives you a prediction ranking immediately. For a practitioner who wants to improve predictions before the World Cup 2026, the fingerprint-based calibration can be applied to all 104 matches from match one.

8.2 The domain adjustment question

The shrinkage parameters in this study are calibrated to domestic club football: 45% home win rate, 27% draw, 28% away. International tournament prediction requires adjustment. Most matches at a World Cup are at effectively neutral venues for most teams, which reduces the PC02_HAC correction. Squad depth and fatigue matter more when squads are limited to 26 players across seven matches. Historic tournament performance (D06_TPE) becomes more relevant in a competition where tournament track record is itself meaningful context.

Practitioners applying this methodology to other sports should recalibrate the base rates (home/away distributions vary substantially across sports and leagues) and assess which fingerprint dimensions are most relevant to their specific prediction context. Practitioners in non-sports domains should identify the analogous dimensions — the equivalent of PC02_HAC and D09_RTS for their domain — and derive shrinkage parameters from those measurements. The formula structure (outcome-specific shrinkage derived from fingerprint h-values) is domain-independent; the specific dimensions and base rates require domain calibration.

8.3 Why the biases transfer across domains

The most theoretically interesting aspect of this finding is that behavioral fingerprints measured in talent evaluation (how a model assesses players) predict performance in match prediction (how a model assesses teams). The two tasks are structurally different: one involves comparative judgment of individuals, the other involves probability estimation for stochastic events.

The reason the fingerprints transfer is that the three quality composite dimensions all measure versions of the same general cognitive tendency: over-weighting institutional prestige and historical reputation relative to current statistical evidence. A model that gives too much credit to a player from a prestigious league is applying the same cognitive pattern as a model that gives too much credit to a historically successful team. The bias is domain-general, not domain-specific — and domain-general biases generalise across tasks.

This has broader implications beyond football, and football is deliberately chosen as a test case precisely because it offers something most prediction domains do not: high-frequency, verifiable outcomes (979 matches across 18 leagues in a single month) with transparent base rates and no data contamination problems once post-cutoff matches are selected. That test environment lets us measure the effect of fingerprint-guided calibration cleanly.

The same methodology applies wherever AI models are used to generate probability estimates over discrete outcomes. In financial markets, prediction market platforms such as Kalshi and Polymarket aggregate crowd-sourced probability estimates on Federal Reserve decisions, corporate earnings, and macroeconomic indicators — and LLM-generated forecasts are increasingly submitted as inputs alongside

human predictions, with the KalshiBench benchmark (Nel, 2025) explicitly using Kalshi markets to evaluate LLM epistemic calibration. The prestige-over-evidence bias documented in the companion study (Gibbins, 2026) maps directly onto a well-known failure mode in financial forecasting: over-weighting the historical track record of established institutions relative to current data. A model that over-weights Premier League prestige over Eredivisie statistics in football will over-weight Federal Reserve credibility over current inflation data in macroeconomics — the same cognitive pattern, different domain.

In election and geopolitical forecasting, platforms such as Metaculus and Manifold Markets aggregate AI-generated and human-generated probability estimates across thousands of questions. Research published in March 2026 found that the best AI forecasting systems are approaching superforecaster-level accuracy on geopolitics and current affairs (Jeen et al., 2026). Applying fingerprint-guided calibration to these models before outcome data is available would be directly analogous to the methodology tested here.

In clinical prediction, AI models are increasingly used to generate risk probabilities for patient outcomes. The overconfidence problem documented across all four models in this study is a known concern in clinical AI deployment: a model that expresses 80% confidence in a diagnosis when 65% would be accurate creates patient harm through excessive certainty. The fingerprint dimensions that predict miscalibration in football — prestige anchoring, risk aversion, narrative susceptibility — have natural analogues in clinical contexts (over-weighting institutional guidelines over patient-specific data, under-predicting rare adverse events, over-adjusting probabilities in response to presenting narratives).

The broader claim is not that football fingerprints should be applied directly to financial or clinical models. It is that the methodology — measure systematic reasoning biases a priori, derive calibration corrections from those measurements, apply before outcomes are observed — is domain-general. Sport provides the cleanest test because of its outcome frequency, transparency, and the availability of a large post-cutoff validation dataset. The confirmation that the approach works in sport is evidence that it is worth testing in higher-stakes domains.

9. Limitations and future work

The $r = 0.997$ correlation between the quality composite and raw model accuracy is striking, but it rests on four data points (one per model). The pattern is directionally clear and theoretically motivated, but replication across more models would strengthen the claim substantially.

The backtesting dataset spans one month across 18 leagues. While 979 matches is a reasonable sample for estimating Brier scores, the improvement estimates may vary in different competition contexts, particularly knockout tournaments with no draw option. The shrinkage parameters were derived from and validated on the same backtesting dataset. Leave-one-league-out cross-validation confirms the directional robustness of the results, but the exact improvement magnitudes should be treated as estimates with uncertainty bounds. The shrinkage formula constants are proprietary and not reported here.

Future work should test whether fingerprint-guided calibration generalises to other sports and other prediction domains, and should examine whether the quality composite remains predictive as models are updated. If fingerprints drift with model updates, the calibration methodology would need to be reapplied after each update — which is itself an argument for the ongoing behavioral monitoring approach described in the companion study (Gibbins, 2026).

The release of GPT-5.5 on April 23, 2026 — mere weeks after the release of GPT-5.4 and during the analysis of this study — illustrates precisely this concern: the fingerprinting battery would need to be rerun on GPT-5.5 before its calibration parameters could be derived.

10. Conclusion

Large language models are increasingly used for sports prediction, but they have a fundamental problem: they are systematically overconfident in specific, model-specific ways. This paper shows that those specific overconfidence patterns can be identified in advance by measuring behavioral fingerprints — the systematic biases that characterise how each model reasons — and that the fingerprint measurements can then be used to correct the predictions before any outcomes are observed.

We find two mechanisms. The first is model selection: a three-dimension composite of behavioral fingerprint measurements predicts raw prediction accuracy with near-perfect correlation ($r = 0.997$), enabling practitioners to rank models before they have seen a single result. The second is calibration: fingerprint dimensions measuring home advantage adjustment and risk aversion determine the optimal shrinkage parameters for correcting each model's overconfidence on each outcome type, improving individual model Brier scores by 4.6–7.3% using tiered shrinkage calibrated to league predictability.

We also find that combining models in an ensemble adds only marginal improvement over the best single model, because the four frontier models tested have similar levels of narrative resistance and therefore make correlated errors. The fingerprint data explains this result: models with similar D10_MNA scores will tend to be wrong on the same matches, limiting the diversification benefit.

The practical upshot is a simple workflow. Measure the behavioral fingerprint once. Use the quality composite to identify the best model. Apply the shrinkage formula to calibrate its predictions. This requires no historical outcome data and works from match one. The 4.6–7.3% improvement in Brier score is not the main commercial advantage — it is a proof of concept. The main advantage is that understanding what your AI is getting wrong systematically is more valuable than not understanding it, both for improving predictions and for knowing when not to trust them.

References

Chen, C., et al. (2025). Dual-edged progression: GPT models and cognitive biases in operations management contexts. Management Science.

Gibbins, G. (2026). Moneyball for LLMs: Behavioral fingerprinting of frontier AI models in football (soccer) talent evaluation and event prediction. SSRN Working Paper. https://osf.io/39gfm/

Kadavath, S., et al. (2022). Language models (mostly) know what they know. arXiv:2207.05221.

Nel, L. (2025). Do Large Language Models Know What They Don't Know? KalshiBench: A New Benchmark for Evaluating Epistemic Calibration via Prediction Markets. NeurIPS 2024. arXiv:2512.16030.

Kumaran, D., et al. (2025). Competing biases underlie overconfidence and underconfidence in LLMs. Nature Machine Intelligence.

Kuhn, J., et al. (2023). Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. ICLR 2023.

Murphy, A. H. (1973). A new vector partition of the probability score. Journal of Applied Meteorology, 12(4), 595–600.

Pei, Z., et al. (2025). Behavioral Fingerprinting of Large Language Models. arXiv:2509.04504.

Ramanayaka, N. D., Dickson, G., Libich, J., & Rayne, D. (2025). Susceptibility to cognitive biases in athlete-team selection. International Journal of Sport and Exercise Psychology.Saiedy, S., Qachmas, M., & Faqiri, D. (2025). Predicting Football Match Outcomes Using Large Language Models: A Comparative Study with Traditional Machine Learning Methods. OSF Preprints. doi:10.31235/osf.io/e5wpy_v2

Sports Analytics Market. (2025). Sports Analytics Market Worth $4.75 Billion by 2030. MarketsandMarkets. https://www.marketsandmarkets.com/PressReleases/sports-analytics.aspFermiEval. (2025). LLMs are Overconfident: Evaluating Confidence Interval Calibration. arXiv:2510.26995.

Jeen, S., Aitchison, M., & Mantic. (2026). Training LLMs to Predict World Events. Thinking Machines Lab. Retrieved from thinkingmachines.ai

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185(4157), 1124–1131.

Disclosure and Data Availability

Behavioral fingerprint measurements are from the pre-registered companion study (OSF: https://osf.io/39gfm/).