Moneyball for LLMs: Behavioral fingerprinting of frontier AI models in football (soccer) talent evaluation and event prediction

Geoff Gibbins – Human Machines Group LLC – geoff@human-machines.com

April 21st, 2026 – learn more about the project at www.modelball.ai

Abstract

Football clubs, agents, and analysts increasingly rely on large language models (LLMs) for player valuation, transfer assessment, and squad selection — yet the systematic biases encoded in these models remain unmeasured. Unlike the architecturally inspectable algorithms dominating football's existing analytics stack, LLMs generate freeform advisory outputs whose biases emerge from training data and cannot be identified through model inspection, requiring behavioral measurement to characterize. We apply behavioral fingerprinting methodology to four frontier AI models (GPT-5.4, Claude Sonnet 4.6, Gemini 3.1 Pro, Grok 3) across twelve talent evaluation dimensions and a five-test Prediction Calibration Module, administering 11,325 unique stimuli per model across three evaluation framings (~45,000 trials total). The central finding inverts expectations: models demonstrate analytical sophistication on attribute and temporal dimensions — preferring current form over reputation, technical contribution over physical attributes, modern archetypes over traditional ones — while simultaneously exhibiting very large biases on institutional prestige dimensions. League Prestige Discount (Cohen's h = 1.18–1.41) is unanimous across all four models. Two of four models show very large demographic evaluation inconsistency (d = 1.14–1.17), comparable in magnitude to their prestige biases, with direct compliance implications under the EU AI Act and equivalent frameworks. Because LLM biases are shared across institutions using the same model, they generate correlated market-wide distortions that do not diversify away — and because model weights update silently and continuously, documented bias profiles require ongoing behavioral monitoring with no analog in the deterministic analytics stack. A fingerprint-weighted ensemble is validated prospectively across 104 FIFA World Cup 2026 matches. The study is pre-registered (OSF: https://osf.io/39gfm/) and establishes a reproducible audit methodology for organizations managing AI bias risk in talent evaluation contexts.

Keywords: *large language models; AI decision-making; behavioral fingerprinting; algorithmic bias; model bias measurement; information asymmetry; market distortion; AI risk; algorithmic auditing; football analytics; talent evaluation; sports economics*

JEL: D83, D91, C81, L83, O33, D47, J24

1. Introduction

Ask GPT-5.4 to choose between two players with identical statistics — same age, same price, same per-90 output — and it will choose the one playing in the more prestigious league. Ask it again with different positions, different leagues, different statistics. It will choose the prestigious-league player again. Do this across hundreds of trials, systematically varying every parameter. GPT-5.4 chooses the prestigious-league player 99% of the time. Not usually. Not often. Ninety-nine times out of a hundred, with no meaningful sensitivity to the magnitude of the quality gap between the leagues, the degree to which one player's statistics exceed the other's, or any other evidence in the prompt. The model's own description of its reasoning confirms this is not judgment but rule: 'If two players produce the exact same numbers, I will always back the one producing them in the stronger competitive environment.' The word 'always' is accurate.

The measured effect size is h=1.37 — by conventional standards, very large. And it is not GPT-5.4 alone. The same systematic preference, the same unconditional institutional hierarchy, appears in every one of the four frontier AI models this study tested: Claude Sonnet 4.6 (h=1.35), Grok 3 (h=1.18), Gemini 3.1 Pro (h=1.41). Four different architectures, four different training regimes, four different providers. One unanimous bias.

When Billy Beane's Oakland Athletics used on-base percentage and slugging percentage to identify players the market had mispriced, the central insight was not that statistics beat scouts. It was that specific, identifiable biases in human evaluation created systematic market inefficiencies that could be exploited before the market corrected. The bias documented above is structurally identical to the one Moneyball was designed to overcome — not a considered judgment that sometimes reaches the same conclusion as prestige, but a prior that prestige itself is the conclusion. The difference is that the bias now operates at scale, silently, inside the advisory tools that clubs, agents, and analysts are already using.

Football's adoption of data-driven recruitment has been substantial. Clubs like Brentford, Union Saint-Gilloise, Red Bull Salzburg, and Bodø/Glimt have demonstrated that signing players from lower leagues at modest fees and selling at significant premiums is a viable strategic alternative to outspending rivals. The post-Moneyball consensus on football analytics is well-documented: expected goals (xG), progressive carries, passing under pressure, pressing intensity, and positional play metrics — provided by StatsBomb, Opta, and Wyscout — have substantially replaced or supplemented traditional evaluation heuristics in elite recruitment departments.

Into this analytically sophisticated landscape, a new actor has arrived: the AI language model. Clubs are increasingly using tools like ChatGPT for player evaluation, budget allocation, and report generation (Playbook Sports, 2025). Agents use LLMs to benchmark contract valuations. Specialist tools such as Comparisonator, which serves over 250 clubs and federations across 271

leagues, now integrate LLM-powered natural language queries with performance data. AI systems are being embedded throughout the talent pipeline, from initial data scouting to transfer negotiation support.

The scale and opacity of this adoption creates not one but three interrelated commercial problems — and a methodology that addresses all three simultaneously.

The first is the market intelligence problem. Clubs and agents relying on AI advisory tools receive distorted valuations — and their counterparties receive them too. An agent who knows that a club's AI applies an LPD bias of h=1.37 to every non-Big 5 player can frame negotiations accordingly. This asymmetry is real and recurring: every silent model update resets every organization without monitoring capability to ignorance about the new fingerprint, while the organization with active behavioral monitoring immediately restores its edge. Without monitoring, the arbitrage is perishable; with monitoring, it is permanent.

The second is the organizational risk problem, which affects every club regardless of counterparty behavior. A club whose AI scouting tools apply an LPD bias to every non-Big 5 player evaluation is building a systematically wrong picture of its transfer universe — in internal shortlisting, budget allocation, and squad planning.

The third is the compliance problem. Two major frontier LLMs show demographic evaluation biases (d = 1.14–1.17) comparable in magnitude to their league prestige biases — creating potential exposure under the EU AI Act, employment law in multiple European jurisdictions, and the regulatory trajectory of Local Law 144.

The methodology introduced in this paper addresses all three problems with the same infrastructure. We apply behavioral fingerprinting to four frontier AI language models across twelve football talent evaluation dimensions and a five-test Prediction Calibration Module.

1.1 Contributions

This paper makes six contributions to the literature at the intersection of AI decision science, sports economics, and computational social science.

- We introduce behavioral fingerprinting as a methodology for domain-specific LLM evaluation in the specific context of athletic talent evaluation, distinguishing it from the architecturally inspectable tools that currently dominate the football analytics stack.
- We provide the first large-scale empirical measurement of systematic biases across four frontier AI models on twelve football-specific evaluation dimensions, covering prestige anchoring, temporal weighting, attribute preferences, demographic consistency, tactical knowledge, and prediction calibration.
- We document a theoretically important pattern — the analytical-prestige divergence — in which AI models have encoded the post-Moneyball analytics consensus on how to

evaluate player attributes while simultaneously encoding the institutional prestige hierarchies that analytics-driven recruitment strategies are designed to overcome.

- We identify a model-specific split on demographic evaluation consistency — with Grok 3 and Gemini 3.1 Pro showing very large effects (d = 1.14–1.17) and GPT-5.4 and Claude Sonnet 4.6 showing negligible effects — establishing that demographic evaluation bias is large, model-specific, and creates potential compliance exposure.
- We establish a prospective public validation surface using 104 FIFA World Cup 2026 matches to test whether fingerprint-weighted ensemble prediction outperforms individual model prediction.
- We develop a solutions hierarchy — continuous behavioral monitoring, structured multi-model divergence flagging, prompt-level correction, and fingerprint-weighted ensembling — and a theoretical framework explaining why each solution addresses a different aspect of the LLM bias risk problem.

2. Theoretical framework

2.1 The football analytics ecosystem and the arrival of LLMs

Football analytics has matured into a sophisticated industry with distinct layers. At the data collection layer, providers including StatsBomb, Opta, Wyscout, and SkillCorner capture over 3,000–3,400 events per match and deliver metrics including expected goals, progressive carries, passing under pressure, pressing intensity, and physical tracking data — with SkillCorner using AI-powered computer vision to achieve up to 95% tracking accuracy across 150+ competitions (SkillCorner, 2025). Liverpool's collaboration with Google DeepMind produced TacticAI, a geometric deep learning system for corner kick optimization whose suggestions were preferred by expert coaches 90% of the time (Wang et al., 2024).

LLMs are fundamentally different from deterministic analytics tools in every relevant respect. They are generative, probabilistic systems trained on vast unstructured text corpora and they produce freeform advisory outputs rather than structured scores. Their biases are not explicit parameters that can be examined or adjusted; they are statistical patterns absorbed from training data that manifest as systematic tendencies across billions of parameters with no direct interpretability. This opacity is the source of their risk.

2.2 Cognitive biases in human talent evaluation

The academic literature on cognitive biases in talent evaluation is extensive. Anchoring and adjustment — where initial information serves as a reference point from which subsequent estimates insufficiently depart (Tversky and Kahneman, 1974) — has been documented in NFL draft decision-making, where top draft picks are significantly overvalued in a manner inconsistent with rational expectations (Thaler and Massey, 2013). In football, crowdsourced valuations from

platforms such as Transfermarkt are systematically influenced by factors that should not affect fundamental value, with evidence of biased estimates across leagues and positions (Coates and Parshakov, 2022; Müller et al., 2017). A recent review identifies anchoring, confirmation bias, availability heuristics, and linear bias as prevalent in how coaches and scouts evaluate players (Ramanayaka et al., 2025).

2.3 Market inefficiency in football transfer valuation

Empirical evidence confirms systematic league-level valuation biases in the human market: players competing in the Premier League are overvalued relative to identically performing players in other leagues (Müller et al., 2017), and crowdsourced valuations are biased predictors of actual fees, with bias magnitude differing between top and lesser leagues (Coates and Parshakov, 2022). These distortions represent the human baseline against which AI model biases should be assessed.

2.4 Cognitive biases in Large Language Models

The empirical study of cognitive biases in LLMs has accelerated rapidly. LLMs are systematically biased in moral decision-making contexts, with biases in some cases stronger than in humans (PNAS, 2025). Anchoring effects have been documented systematically: LLMs mimic the anchoring heuristic such that higher anchors reliably shift the distribution of predicted numeric values upward (Lou, 2025). Studies of LLMs in operations management contexts document a "dual-edged progression": models demonstrate higher accuracy on tasks with objective solutions while simultaneously exhibiting stronger human-like biases in preference-based tasks (Chen et al., 2025).

The prior work most directly relevant to the present study is Pei et al. (2025), who introduce a behavioral fingerprinting framework using a diagnostic prompt suite and automated LLM-as-judge evaluation pipeline, analyzing eighteen models across capability tiers. The present study extends Pei et al. to a specific high-stakes applied domain, introduces a multi-framing instantiation design, and provides prospective real-world outcome validation.

2.5 LLM bias in talent evaluation contexts: evidence from hiring

The closest existing literature is the growing body of work on LLM bias in hiring and resume screening. Frontier LLMs show significant demographic evaluation inconsistency in hiring contexts: models favor white-associated names 85% of the time (Wilson and Caliskan, 2025); 22 models consistently favor female-named candidates across 70 professions (PeerJ, 2026); general-purpose LLMs achieve race-wise impact ratios of 0.809 against a 0.8 fairness benchmark (Eightfold AI, 2025); and intersectional gender-racial biases persist across model versions despite debiasing efforts (PMC, 2025).

2.6 The dual-edged progression hypothesis

Prior work documents a "dual-edged progression": models demonstrate higher accuracy on tasks with objective solutions while exhibiting stronger human-like biases in preference-based tasks. In football talent evaluation, this maps onto the difference between analytically grounded evaluation (preferring current form, technical metrics) and institutionally anchored evaluation (premium for prestigious leagues, clubs, and tournaments). The present study tests whether frontier AI models show this dual-edged progression in the football domain.

2.7 The opacity problem: why LLM biases require behavioral measurement

Because LLM biases cannot be identified by model inspection, they can only be characterized through behavioral measurement — systematic administration of controlled stimuli and observation of response patterns across many trials. With a deterministic tool — an xG model, a Comparisonator AI points score — a sophisticated user can in principle understand the model's logic through feature importance scores and systematic input variation. With an LLM, no equivalent inspection path exists. A sporting director receiving a valuation from GPT-5.4 has no mechanism for determining whether the output reflects a genuine performance assessment or a league prestige discount of h=1.37.

A second critical difference concerns update transparency. Deterministic xG models are versioned and announced; LLMs are updated continuously and without disclosure. A bias of h=1.35 documented today may be different after a silent update next month — creating an ongoing behavioral monitoring requirement with no analog in the deterministic analytics stack.

3. Methodology

3.1 Behavioral fingerprinting design

We apply a within-subjects controlled experiment to four frontier AI language models: GPT-5.4 (OpenAI), Claude Sonnet 4.6 (Anthropic), Gemini 3.1 Pro (Google), and Grok 3 (xAI). For each of twelve talent evaluation dimensions, matched-pair stimuli are generated in which one evaluative variable is systematically manipulated while all other candidate attributes are held constant. Each dimension is instantiated across three functionally distinct evaluation framings: personnel selection, competitive market participation, and secondary market valuation.

A judge model (Claude Opus 4.6) scores all responses against pre-specified dimension rubrics, blind to study model identity. The judge receives only the study model's response and the rubric and returns a structured score of 1 (bias demonstrated), 0 (no bias), or −1 (ambiguous/excluded). Effect sizes are computed as Cohen's h for binary forced-choice outcomes and Cohen's d for continuous valuation outcomes, with 95% bootstrap confidence intervals from 10,000 resamples. Benjamini-Hochberg FDR correction is applied within dimension clusters.

All stimuli were generated from random seed 20260503, enabling full reproducibility. The design is pre-registered at OSF (https://osf.io/39gfm/). 11,325 unique stimuli were administered to each of the four study models, generating approximately 45,000 model query trials in total.

3.2 Dimension taxonomy

Twelve dimensions organized in four theoretical clusters:

Market perception cluster:

- *League Prestige Discount (D01)* — systematic preference for Big 5 league players over identically performing non-Big 5 players.
- *Club Prestige Halo (D02)* — valuation premium for players at or formerly at prestigious clubs.
- *Demographic Evaluation Consistency (D03)* — consistency of valuations across name and nationality signals with identical performance data.

Temporal cluster:

- *Age Curve Encoding (D04)* — early application of age-based discounts relative to sports science evidence.
- *Temporal Weighting (D05)* — weighting of historical pedigree vs. current form.
- *Tournament Pedigree Encoding (D06)* — weighting of international tournament history vs. current club form.

Attribute cluster:

- *Attribute Type Preference (D07)* — preference for physical vs. technical attribute profiles with equivalent composite match impact.
- *Role Value Encoding (D08)* — positional label premium and full-back archetype encoding.
- *Risk Tolerance in Selection (D09)* — preference for consistent vs. high-variance performers.

Contextual cluster:

- *Media Narrative Anchoring (D10)* — valuation deviation from a neutral baseline under positive or negative narrative framing.
- *Tactical Knowledge Index (D11)* — breadth and recency of football tactical knowledge, scored on a 0–10 composite scale.
- *Tactical Context Adjustment (D12)* — application of tactical knowledge to player valuations when explicitly instructed.

3.3 Prediction calibration module

Five prediction calibration tests assess each model's tendency toward prestige anchoring in match outcome prediction:

- Fixture Difficulty Calibration (PC01)
- Home Advantage Calibration (PC02)
- Upset Identification (PC03)
- Team Narrative Override (PC04)
- Odds Integration (PC05)

PC02 uses a paired scoring design comparing home and neutral condition responses for the same match to compute home advantage adjustment relative to the empirically documented benchmark range of 5–15 percentage points.

3.4 Tactical knowledge index

Twenty-five structured questions across five domains (definitional knowledge, player profile matching, system-player compatibility, manager philosophy, temporal evolution) assessed against pre-specified answer keys on a 0–3 scale per question, producing a 0–10 composite. The temporal evolution domain specifically tests whether models have updated their understanding of the game's development between 2010 and 2025.

3.5 Pilot study and methodology corrections

A technical validation pilot of 12,596 trials across three dimensions (D01, D05, D10) identified three issues resolved before full collection. D10 had a scoring architecture flaw — the judge was asked to compare valuations to a neutral baseline it could not access; stimuli were redesigned to embed neutral-condition reference valuations in positive and negative narrative prompts. D12 required passing the correct_answer field to the judge, which had been omitted from the initial scoring pipeline; all 4,800 D12 trials were rescored with minimal effect on estimates ($|\Delta h| < 0.06$ for all models). D11 required recollection of Gemini trials truncated by an insufficient token limit parameter. All three corrections are documented in an OSF amendment filed prior to analysis.

A post-hoc audit of the full scored dataset identified four additional findings reported here for transparency. Study-wide Gemini truncation — resolved: raw response files confirmed that the TKI truncation was study-wide; a full recollection was conducted with corrected max_tokens settings; pre-registered hypothesis H13 (Gemini elevated non-committal response rate) is not confirmed as it was a collection artifact. D08 additional collection: an initial high ambiguity rate for Gemini on D08 was resolved by targeted recollection. Final Gemini D08 n_total=433, ambiguity=0%. Non-Gemini models show negligible D08 ambiguity (<1%). The n_ambiguous_excluded pipeline reporting bug was a reporting error only; all n_total values are

correct post-exclusion. D06 tournament framing was reinterpreted as context-sensitivity rather than a fixed bias. PC05_OI (Odds Integration) retains elevated ambiguity for Gemini only (48.3%, n=15 scoreable trials), driven by safety refusals on odds-based prediction tasks — this is the only remaining dimension with a sample size caveat.

4. Results

4.1 Overview: the analytical-prestige divergence

Table 1 presents behavioral fingerprint scores across all twelve dimensions and four models. The clearest pattern is a systematic divergence between dimension clusters: models show large to very large negative effects on the Attribute, Role, and Temporal (form sub-condition) dimensions, indicating analytical sophistication aligned with the post-Moneyball consensus, while showing large to very large positive effects on the Market Perception dimensions, indicating strong institutional prestige anchoring.

Table 1. Behavioral Fingerprint Scores (Cohen's h or d)

Dimension	GPT-5.4	Claude 4.6	Grok 3	Gemini 3.1	Cross-Model Pattern
D01 League Prestige Discount	+1.37	+1.35	+1.18	+1.41	Unanimous very large positive
D02 Club Prestige Halo	+0.65	+0.19	+0.62	+0.04	3/4 positive
D03 Demographic Evaluation Consistency	−0.11	+0.09	+1.17	+1.14	Split — Grok/Gemini very large positive
D04 Age Curve Encoding	+0.75	+1.06	+0.87	+0.75	Unanimous positive
D05 Temporal Weighting	−0.61	−0.69	−0.61	−0.47	Unanimous negative
D06 Tournament Pedigree Encoding	−0.21	−0.47	−0.33	−0.57	Unanimous negative
D07 Attribute Type Preference	−0.75	−1.14	−0.47	−0.31	Unanimous negative
D08 Role Value Encoding	−0.58	−0.24	0.00	+1.27	Mixed: GPT large −, Claude small −, Grok negligible, Gemini very large + (uniquely positive)
D09 Risk Tolerance in Selection	−0.70	−0.37	−0.88	−0.96	Unanimous negative

D10 Media Narrative Anchoring	+0.03	−0.16	+0.12	−0.00	Mixed: GPT/Grok small +, Claude small −, Gemini negligible
D12 Tactical Context Adjustment	−1.30	−1.11	−1.44	−0.70	GPT, Claude, Grok: very large −; Gemini: large −

Note: All Gemini effect sizes based on recollected data (see Section 3.5). Effect sizes with |h| or |d| > 0.2 and 95% bootstrap CI excluding zero are statistically significant after Benjamini-Hochberg FDR correction within clusters.

4.2 League Prestige Discount (D01)

All four models show a positive League Prestige Discount effect, unanimous in direction and magnitude. After Gemini recollection, all four models show very large effects: GPT-5.4 (h=1.37), Claude Sonnet 4.6 (h=1.35), Grok 3 (h=1.18), and Gemini 3.1 Pro (h=1.41). All four models systematically select Big 5 league players over identically performing non-Big 5 players in 96–99% of trials. GPT articulates the underlying heuristic directly across hundreds of responses:

> *"Ligue 1 is much closer to Premier League level than MLS in terms of speed and physical intensity, tactical structure, quality of opposition, and defensive responsibility in transition."*

The model has a fully elaborated four-tier league hierarchy that it applies as a near-automatic tiebreaker whenever player statistics are identical. The unanimity of this finding across four models from four different providers (h = 1.18–1.41) is the strongest single result in the study.

This finding maps onto documented market-level phenomena. Research confirms that players competing in the Premier League are systematically overvalued relative to identically performing players in other leagues (Müller et al., 2017). The AI models studied here encode a bias of comparable direction but substantially larger magnitude. Platforms such as Comparisonator explicitly address this distortion through "virtual transfer" simulations — the present study shows that LLMs do not perform this adjustment; they apply a prestige premium on top of performance data.

4.3 Attribute Type Preference — the analytical sophistication finding (D07)

Contrary to the pre-registration hypothesis that models would show a physical attribute preference, all four models show negative effects — preferring the technically proficient player over the physically dominant player when composite match impact is held constant. Claude Sonnet 4.6 shows the largest effect (h=−1.14), followed by GPT-5.4 (h=−0.75), Grok 3 (h=−0.47), and Gemini 3.1 Pro (h=−0.31). The raw responses reveal distinct analytic frameworks:

> *GPT-5.4: "I want the player with the more scalable, possession-dominant profile... Elite chance creation matters most here."*
>
> *Claude Sonnet 4.6: "Player B's data is incomplete in the ways that matter most. The physical metrics tell me Player B can do physical things. They tell me almost nothing about what he does with the ball."*

Grok 3: "Player A's exceptional physical profile offers a unique advantage in the Premier League, where physicality and intensity often define key moments."

4.4 Demographic Evaluation Consistency — the split finding (D03)

The most commercially and ethically significant pattern in the dataset is the cross-model split on demographic evaluation consistency. Grok 3 (d=1.17) and Gemini 3.1 Pro (d=1.14, confirmed after recollection) show very large effects, with framing-consistent patterns indicating robust behavioral tendencies. GPT-5.4 (d=−0.11) and Claude Sonnet 4.6 (d=+0.09) show near-equal treatment across demographic signals.

The Claude result is particularly arresting when viewed through the raw responses, because Claude's near-zero aggregate effect coexists with individual responses that explicitly invoke nationality as a decision criterion:

"Player A. If the football data is truly identical, I take the English international every time for a Premier League club. Homegrown value under registration rules, market premium for resale, lower adaptation risk for language, culture, and league familiarity."

For Grok and Gemini, the effects are not explainable by legitimate market reasoning — the magnitude (d=1.14–1.17) is too large and too framing-consistent to reflect selective application of contextually appropriate signals. Wilson and Caliskan (2025) found that three frontier LLMs favored white-associated names 85% of the time in resume ranking tasks. The present study finds that demographic bias in talent evaluation is model-specific rather than universal.

4.5 Five unanimous negative dimensions — the analytical sophistication cluster

Five dimensions show unanimous negative effects across all four models, indicating that all frontier AI models prefer the analytically sophisticated option over the traditional scouting heuristic.

Temporal Weighting (D05): All models strongly prefer current form over historical pedigree (h ranging from −0.61 to −0.69), directly contradicting hypothesis H5. GPT states the principle directly: *"Pedigree, awards, and past greatness do not score fantasy points this season."*

Tournament Pedigree Encoding (D06): All models prefer players with strong current club form over those with strong international tournament histories (h ranging from −0.21 to −0.57), contradicting hypothesis H6. A representative response: *"Player B's current club output matters more than World Cup pedigree in a £12m decision. Tournament résumé is historical and not strongly predictive of immediate club impact."*

Role Value Encoding (D08):

Risk Tolerance in Selection (D09): All models prefer the high-variance high-upside performer over the consistent lower-ceiling performer (h = −0.37 to −0.96).

Tactical Context Adjustment (D12): All models apply explicit tactical system requirements with near-perfect fidelity when instructed (h ranging from −0.70 to −1.44). Claude's D12 responses produced some of the most analytically sophisticated reasoning in the dataset:

> *"Player B's pressing numbers are not marginally better — they are transformatively different: 14.8 vs 4.2 pressures per 90 — Player B applies 3.5x more pressure. In a high-press 4-3-3, the forwards trigger the entire pressing structure. When the striker presses the goalkeeper, it forces rushed distribution, compresses the opposition's build-up, and activates the midfield press. Player A's 4.2 pressures per 90 means the press never starts."*
>
> *"At £48m for a 23-year-old, you are not simply buying a footballer — you are buying a system component."*

4.6 Tactical Knowledge Index

Table 2. TKI Scores by Model and Domain (0–10 scale)

Model	Overall	Definitional	Player Profile	Sys. Compat.	Manager Phil.	Temporal Evol.
Claude 4.6	**8.53**	10.0	8.67	6.0	8.67	9.33
GPT-5.4	**8.40**	10.0	8.0	7.33	8.67	8.0
Grok 3	**8.13**	10.0	6.67	6.0	8.0	10.0
Gemini 3.1	**6.40**	9.33	5.33	6.67	8.0	**2.67**

All models score at or near ceiling on definitional tactical knowledge (9.33–10.0). The critical differentiating domain is temporal evolution — knowledge of how football has changed between 2010 and 2025. Gemini scores 2.67/10 on temporal evolution against 8.0–10.0 for the other three models. GPT-5.4's temporal evolution responses track the tactical shift with historical specificity:

> *"In 2010, many top teams still built around a classic No. 9: occupy centre-backs, attack crosses, finish moves in the box, play with back to goal, provide a focal point for direct play. By 2025, the striker is often expected to: press aggressively from the front, link play between lines, create space for wide forwards and attacking midfielders, contribute in transitions, be tactically flexible, still score but not only through traditional poaching."*

4.7 Prediction Calibration Module

Table 3. PCM Effect Sizes by Model

Dimension	GPT-5.4	Claude 4.6	Grok 3	Gemini 3.1	Note
PC01 FDC	−0.75 (large)	−1.08 (v. large)	**0.00 (negligible)**	−0.04 (neg.)	Grok only calibrated model
PC02 HAC	+0.12 (small)	+0.86 (v. large)	−0.99 (v. large)	+0.81 (v. large)	Maximum divergence — core ensemble signal

PC03 USI	−1.57 (v. large)	−1.57 (v. large)	−1.57 (v. large)	−1.57 (v. large)	All three models: no upset bias
PC04 TNO	−1.57 (v. large)	−1.57 (v. large)	−0.03 (negligible)	−1.57 (v. large)	Grok: 42.5% narrative bias rate vs 0% for GPT/Claude
PC05 OI	−1.05 (v. large)	−1.57 (v. large)	**+0.64 (large)**	−1.57 (v. large)	Grok only model integrating market signals

Note: All Gemini PCM dimensions are from the recollected dataset (max_tokens=1,500). Sample sizes: PC01 n=50, PC02 n=44 (paired scoring post-recollection), PC03 n=30, PC04 n=34, PC05 n=15. All four models are included in PCM comparative analysis.

Three findings are particularly notable. On fixture difficulty calibration (PC01), Grok 3 is the only model achieving calibrated prediction; Claude shows the largest miscalibration (h=−1.08). On home advantage calibration (PC02), all four models diverge substantially. GPT is near-calibrated (h=+0.12). Claude (h=+0.86, very large) and Gemini (h=+0.81, very large) both dramatically over-adjust for host nation home advantage in the same direction. Grok dramatically under-adjusts (h=−0.99, very large). Three of the four models are substantially wrong, but in two different directions — making this the dimension with the highest ensemble correction potential. On odds integration (PC05), Grok 3 is the only model showing meaningful positive odds integration (h=+0.64, large) — the only model that meaningfully incorporates betting market signals.

5. Discussion

5.1 The commercial edge in understanding LLM biases

The three commercial problems introduced in Section 1 are not independent. Each is addressed by the behavioral fingerprinting methodology, and two of the three are amplified by the same products.

The market intelligence problem — arbitrage — is real and recurring. An agent who knows that a buying club's AI applies an LPD bias of h=1.37 can frame the opening discussion with Premier League comparable transfers, activate the prestige anchor deliberately, and negotiate against a buyer whose AI has inflated the reference point. The Bournemouth Paradox illustrates this: GPT-5.4 is capable of inverting its club prestige halo in the right framing:

> *"Sign Player B from Bournemouth. If age, league, output, underlying numbers, and price are all identical, I prefer the striker producing those numbers in a weaker team context... That suggests more self-sufficiency, less inflation from elite teammates, and greater likelihood the production continues in a new environment."*

The model is capable of this reasoning. The behavioral fingerprint reveals that it applies it inconsistently, defaulting to prestige anchoring in most contexts while applying the inverse logic in specific framings. With monitoring, the edge is permanent — every update is a reset event that reinstates a clean information gap. The monitoring product is not layered on top of the arbitrage strategy; it is the mechanism that converts a perishable edge into a structural one.

The compliance problem is the most legally immediate. The DEC finding — Grok d=1.17, Gemini d=1.14 (confirmed after recollection) — documents that two major frontier LLMs show demographic evaluation biases comparable in magnitude to their league prestige biases, with player valuations influenced by name and nationality signals independent of performance data.

5.2 Second-order effects of legible AI biases

The legibility of LLM bias profiles creates second-order effects that extend beyond straightforward first-mover arbitrage. We identify five distinct mechanisms by which this might show up:

- **Adversarial framing:** once the bias profile of a counterparty's AI tool is known, sophisticated actors can strategically present information to exploit it.
- **Correlated market-wide bias:** LLMs are shared infrastructure. Multiple clubs querying GPT-5.4 for transfer valuations receive the same LPD bias of h=1.37. The Financial Stability Board (2024) warns that AI-driven herding may "streamline modeling approaches leading to increased market correlation" in ways that amplify fragility.
- **Bias reinforcement through feedback loops:** if AI-influenced negotiations result in lower fees for non-Big 5 players, those fees enter the training data for future models, reinforcing the LPD in the next training generation.
- **The two-tier information market:** publication of behavioral fingerprint findings creates an immediate asymmetry between actors who have acted on them and those who have not — the original Moneyball dynamic.
- **Silent update instability:** bias profiles documented today may shift after silent model updates with no announcement. Glickman and Sharot (2025) note that AI outputs carry an appearance of objectivity that may reduce organizational resistance to AI-influenced decisions.

5.3 The post-Moneyball paradox

The most theoretically interesting finding is the divergence between analytical sophistication dimensions and prestige anchoring dimensions. Models have encoded the post-Moneyball consensus on how to evaluate players — technical metrics over physical, current form over reputation, modern archetypes over traditional ones — but have not encoded the corresponding lesson about institutional prestige. The contrast is sharpest when individual responses from opposite ends of the dimension taxonomy are placed alongside each other. On the analytical sophistication dimensions, Claude explicitly rejects historical reputation:

> *"Historical reputation is not a current performance predictor. Two consecutive group-stage exits invalidate the 'always reaches semi-finals' narrative for predictive purposes."*

On the prestige anchoring dimensions, GPT explicitly invokes reputation as the decisive factor:

> *"If two players produce the exact same numbers, I will always back the one producing them in the stronger competitive environment."*

A plausible explanation is training data composition: the post-Moneyball analytics discourse is extensively documented in public writing while the parallel critique of institutional prestige inflation is less developed in text form.

5.4 The demographic evaluation split and fairness implications

The finding that Grok 3 and Gemini 3.1 Pro show very large demographic evaluation inconsistency effects while GPT-5.4 and Claude Sonnet 4.6 show negligible effects reveals that model-specific variation in demographic evaluation bias is substantial. If AI scouting tools influence which players are assessed and which valuations are offered in negotiations, a very large demographic evaluation bias directly affects career and income outcomes for players from affected demographic groups. The finding that this bias is model-specific rather than universal indicates it is an encoding that can and should be identified, reported, and corrected.

5.5 Grok as the market-responsive model

Grok 3 stands out in the Prediction Calibration Module as the only model achieving calibrated fixture difficulty prediction and the only model showing positive odds integration. Grok's training data composition — which includes substantially more social media sports commentary and real-time betting market discussion — appears to produce a more market-responsive prediction profile. However, Grok's very large demographic evaluation inconsistency (d=1.17) represents a significant countervailing liability for talent evaluation use cases.

5.6 Solutions that emerge from these findings

A number of solutions could be deployed to overcome the biases found in this analysis, including:

- **Continuous behavioral** monitoring – because LLM model weights are updated continuously and without disclosure, a bias profile documented at one point in time may be materially different after the next model update.
- **Multi-model divergence flagging** is the most practically accessible near-term intervention. Rather than a full ensemble with bias-corrected weights, this approach queries two or three frontier models with the same evaluation prompt and interprets divergence through the lens of documented fingerprint differences.
- **Prompt-level correction and structured evaluation formats** partially suppress documented biases without requiring model access. A system prompt instructing the model to evaluate players without reference to league label can reduce the prestige anchor.

- **Fingerprint-weighted ensembling** is the most sophisticated solution and the one most directly validated by the World Cup prediction component of this study, in which prediction quality is improved by understanding the relative fingerprints of the models (to evaluate which differences between models are signal, and which are noise).
- **Regulatory disclosure and certification** – the EU AI Act creates a high-risk classification framework with relevance to AI systems influencing employment-adjacent decisions. The present study's pre-registered, publicly documented methodology is structurally positioned to contribute to an audit standard.

6. Limitations and future directions

Several limitations warrant acknowledgment. Gemini 3.1 Pro underwent full study-wide recollection after a truncation issue was identified in post-hoc audit (but was rectified fully). The study models represent a snapshot of four frontier models at a specific point in time; model updates may shift fingerprints materially.

The D08 (Role Value Encoding) dimension required targeted recollection for Gemini following an initial high ambiguity rate. After recollection, Gemini D08 has n_total=433 with 0% ambiguity, yielding a very large positive effect (d=+1.27) — the dimension on which Gemini most distinctively diverges from the other three models in direction. Non-Gemini models show negligible D08 ambiguity. Future work should further investigate the positional re-labelling stimulus design. The D03 DEC measurement captures an aggregate effect across name and nationality signal variations; the direction of the DEC effect for specific demographic groups requires more granular analysis. The TKI-TCA gap hypothesis was disconfirmed by a TCA ceiling effect; future work should design stimuli with more subtle tactical manipulations. A direct comparison between LLM bias profiles and the biases of deterministic football analytics tools would clarify whether LLMs introduce additional distortions.

7. FIFA World Cup 2026 validation exercise

7.1 How will the ensemble prediction perform in predicting what happens?

The Prediction Calibration Module findings establish distinct and non-overlapping miscalibration profiles for each model. On fixture difficulty calibration, Grok is the only model achieving near-zero bias; Claude over-weights historical prestige at h = −1.08. On home advantage calibration, GPT is near-calibrated (h = +0.12), Claude massively over-adjusts (h = +0.86), and Grok dramatically under-adjusts (h = −0.99). On odds integration, Grok partially incorporates betting market signals (+0.64), GPT largely ignores them (−1.05), and Claude ignores them entirely (−1.57). On upset identification and narrative override, all three models perform without meaningful bias.

No individual model is well-calibrated across all five prediction dimensions simultaneously. The ensemble weighting formula is: weight(m) = normalize(1 / (|bias_score(m,d)| + ε)) for PCM dimensions d active for each match context (ε = 0.1 prevents extreme weight concentration). The ensemble superiority claim is not a generic "wisdom of crowds" argument but a specific prediction that documented miscalibrations cancel at the match types where individual models diverge most — host-nation fixtures, form-vs-prestige mismatches, and odds-divergence markets. The goal is to see whether the ensemble prediction outperforms individual models, and a naïve average of the predictions from different models. In other words – does knowing about the biases help us to predict the outcome of the World Cup?

7.2 Implementation and prospective validation

Pre-event predictions for all 104 FIFA World Cup 2026 matches will be generated from a standardized real-time context brief, queried from all four study models — GPT-5.4, Claude Sonnet 4.6, Grok 3, and Gemini 3.1 Pro — plus the fingerprint-weighted ensemble and a naive average of the same four models. The full PCM recollection documented in Section 3.5 provides complete calibration data for all four models, enabling Gemini's inclusion in the prospective validation. All predictions will be timestamped and publicly logged at www.modelball.ai before each match commences, using an append-only archival mechanism that prevents modification after the event starts.

The pre-specified criterion for ensemble superiority (H19) is that the fingerprint-weighted ensemble outperforms every individual model on at least two of the three primary accuracy metrics across the full 104-match tournament. Four secondary hypotheses are pre-registered. H19a: ensemble gains are largest on host-nation fixtures, where PC02 miscalibration diverges most severely across models. H19b: ensemble gains are largest on odds-divergence matches, where PC05 differences are most salient. H19c: the fingerprint-weighted ensemble achieves a lower mean Brier score than the naive average of the same four models — the direct test that fingerprint weighting adds value beyond simple model averaging. H19d: the Brier score advantage of the fingerprint ensemble over the naive average is larger on host-nation and odds-divergence match types than on standard matches, testing whether the fingerprint mechanism specifically improves predictions where PCM calibration signals are most divergent. Results will be reported in a post-tournament addendum.

8. Conclusion

Beane's 'Moneyball' methodology kept producing new insights as the rest of the market kept using the same uncritical heuristics. The behavioral fingerprinting methodology established here attempts to do the equivalent for AI tools: measuring what they systematically get wrong, before counterparties discover it through the decisions of others.

The findings document a coherent pattern across four frontier models and twelve evaluation dimensions. On the analytical sophistication dimensions — attribute preferences, temporal weighting, role encoding — models have absorbed the post-Moneyball consensus. On the institutional prestige dimensions — league label, club affiliation, age — they encode the biases that analytics-driven recruitment was designed to overcome, at very large effect sizes unanimous across all models. On demographic evaluation, two of four models show biases comparable in magnitude to their prestige biases, creating compliance exposure that is model-specific, measurable, and avoidable.

The three commercial problems are not symmetrically distributed. The arbitrage opportunity accrues to organizations that act on the fingerprint; the organizational risk accrues to every organization using AI advisory tools regardless of counterparty behavior; the compliance exposure is immediate for organizations using Grok or Gemini in talent evaluation contexts. Monitoring converts the arbitrage from perishable intelligence into a structural permanent advantage, because every silent model update is a reset event that reinstates a clean information gap between organizations with and without monitoring capability.

The behavioral fingerprinting methodology established here — pre-registered stimulus battery, judge-blind scoring, bootstrap confidence intervals, multi-framing instantiation — addresses all three problems with the same infrastructure. As AI advisory adoption expands and regulatory direction toward mandatory bias disclosure clarifies, the organizations that have characterized what their tools are getting wrong will hold the informational edge: in negotiations, in their own decision quality, and in regulatory readiness. The Moneyball analogy fully applies — it just turns out the thing worth measuring was never only what the other scouts were missing.

References

Beane, B., & Lewis, M. (2003). Moneyball: The Art of Winning an Unfair Game. *Norton.*

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences, 120(6), e2218523120.*

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

Coates, D., & Parshakov, P. (2022). The wisdom of crowds and transfer market values. *European Journal of Operational Research.*

Eightfold AI. (2025). Evaluating the Promise and Pitfalls of LLMs in Hiring Decisions. *Eightfold AI White Paper.*

Financial Stability Board. (2024). Financial Stability Implications of Artificial Intelligence. *FSB Report.*

Franceschi, M., et al. (2024). Determinants of football players' valuation: A systematic review. *Journal of Economic Surveys, 38(2).*

Glickman, M., & Sharot, T. (2025). Human-AI feedback loops amplify judgment errors. *Nature Human Behaviour.*

Hagendorff, T., Fabi, S., & Kosinski, M. (2023). Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nature Computational Science, 3, 833–838.*

IBM. (2024). Avoiding Bias in AI: Why Addressing Bias Is Critical to AI Success. *IBM Institute for Business Value.*

IMF. (2025). Regulatory considerations regarding accelerated use of AI in securities markets. *IMF Technical Notes and Manuals, 2025/016.*

Jones, E., & Steinhardt, J. (2022). Capturing failures of large language models via human cognitive biases. *NeurIPS 2022.*

Kahneman, D. (2011). Thinking, Fast and Slow. *Farrar, Straus and Giroux.*

Lou, J. (2025). Anchoring bias in large language models: An experimental study. *Journal of Computational Social Science.*

McHale, I., & Holmes, B. (2023). Estimating transfer fees of professional footballers. *European Journal of Operational Research.*

Müller, O., Simons, A., & Weinmann, M. (2017). Beyond crowd judgments: Data-driven estimation of market value in association football. *European Journal of Operational Research, 263(2), 611–624.*

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

PeerJ. (2026). Gender and positional biases in LLM-based hiring decisions: evidence from comparative CV evaluations. *PeerJ Computer Science.*

PMC. (2025). Measuring gender and racial biases in large language models: Intersectional evidence from automated resume evaluation. *PMC.*

PNAS. (2025). Large language models show amplified cognitive biases in moral decision-making. *Proceedings of the National Academy of Sciences, 122(25).*

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.*

SkillCorner. (2025). AI and Computer Vision in Football Analytics. *SkillCorner Technical Overview.*

Thaler, R., & Massey, C. (2013). The loser's curse: Decision making and market efficiency in the NFL draft. *Management Science, 59(7), 1479–1495.*

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

Wang, S., & Redelmeier, D. (2024). Cognitive biases in large language models: Implications for research and practice. *NEJM AI, 1, AIe2400961.*

Wang, Z., Veličković, P., Hennes, D., et al. (2024). TacticAI: an AI assistant for football tactics. *Nature Communications, 15, 1906.*

Wilson, K., & Caliskan, A. (2025). AI tools show biases in ranking job applicants' names according to perceived race and gender. *AAAI/ACM Conference on AI, Ethics, and Society.*

Note: Stimulus files are maintained as proprietary. Scored trial data and analysis code available upon paper submission per OSF data sharing policy. Pre-registration: https://osf.io/39gfm/

The author is principal of Human Machines Group LLC, which has filed provisional patent applications covering aspects of the methodology described.