How can an understanding of LLM model bias be used to improve financial advice and investment decisions?

Confidence calibration, error detection and drift monitoring in AI-enabled financial advice

Geoff Gibbins

Human Machines Group LLC — geoff@human-machines.com — April 27th 2026

Learn about the project at www.verrix.ai

Abstract

Behavioral fingerprinting establishes that AI financial advisors carry stable, model-specific biases that systematically influence investment recommendations at scale (Gibbins, 2026). This paper addresses the practitioner question that follows: given a measured behavioral fingerprint, what can you actually do with it? We report findings from a validation study of 480 labeled responses across 24 financial advisory scenarios with dollar-impact ground truth ($383–$94,000 per error). Three things work. First, a calibrated confidence model (AUC = 0.876, Brier = 0.105) identifies which individual responses to trust: responses in the TRUST zone ($p \geq 0.6$; 19.8% of responses) are correct 100% of the time; those in the REVIEW zone ($p < 0.3$; 68.5%) are wrong 93% of the time. Second, a standalone calculation injection mechanism achieves 72% auto-approval and 16% error detection across a 111-response validation pilot with zero degradation events—detecting potential recommendation errors before they reach users, without knowing the correct answer. Third, behavioral fingerprints accurately track model behavioral change between versions, enabling drift detection before clients experience it. Three things do not work as practitioners might hope: ensemble weighting schemes do not improve binary accuracy due to a 42.4% joint failure rate from correlated model errors; quality metrics invert when applied across models rather than within them; and fingerprint h-values alone predict scenario-level failure worse than chance. We characterize why each fails and what to use instead.

Keywords: AI financial advisory, LLM behavioral bias, confidence calibration, response quality assurance, behavioral fingerprinting, drift detection, ensemble methods, investment advice governance

JEL: G41, C45, G24, C53

1. Introduction

AI financial advisors are not neutral. When a user asks ChatGPT, Gemini, Claude, or any other large language model for investment guidance, the answer reflects not only the financial facts but the behavioral tendencies embedded by training. These tendencies are stable, model-specific, and measurable. Behavioral fingerprinting—the systematic characterization of these tendencies across controlled vignette pairs — establishes the advice genome of each platform: its technology sector preference, brand recognition bias, momentum amplification, and regulatory compliance profile (Gibbins, 2026).

Knowing the fingerprint is the starting point. The deployment question is different: you have an AI advisor running in production. Thousands or millions of users are receiving responses. How do you know which of those responses to trust? What do you do about the ones you should not? And when your provider silently updates the model, how do you know whether the behavioral profile you validated has changed?

This paper provides empirical answers to these three questions. Section 3 presents the calibrated confidence framework: a model trained on extractable quality features that identifies which responses are reliably correct, genuinely uncertain, or almost certainly wrong. Section 4 presents the quality improvement loop: a mechanism that performs the analytical work a model skipped, checks the arithmetic against the recommendation without knowing the correct answer, and automatically flags potential errors before they reach clients. Section 5 shows what fingerprints do reliably predict — behavioral drift between model versions — and the specific consensus pattern that identifies near-certain errors without ground truth. Section 6 documents what we tested that does not work: why ensemble weighting fails, why quality metrics invert across models, and why fingerprints cannot predict instance-level failure. Section 7 describes the commercial tools these findings underpin.

The study was pre-registered at OSF (osf.io/2ngv7) on April 9, 2026. Four LLMs were tested: GPT-5.4 Thinking, GPT-5.3 Instant, Claude Sonnet (claude-sonnet-4-6), and Gemini Flash. The validation battery comprises 480 labeled responses across 24 scenarios with dollar-impact ground truth ranging from $383 to $94,000 per error.

2. Methods

2.1 Validation battery

Twenty-four scenarios spanning five categories (Table 1) were administered to four LLMs across five repetitions each, yielding 480 labeled responses. Each scenario has a pre-specified ground truth answer and a dollar-impact stake representing the financial harm from an incorrect recommendation. Stakes were set based on representative client scenarios; the $94,000 stake on HS_RET_001 reflects the cumulative difference between correct and incorrect pension election over a 20-year retirement horizon.

Table 1. Validation Battery: Scenario Distribution by Category

Category	n Scenarios	Description	Dollar Stake	Responses
HS_	6	High-stakes retirement and debt decisions: pension election, withdrawal rate, Social Security timing, debt payoff sequencing	$765–$94,000	120
LS_	4	Consumer protection: suitability, payday lending, BNPL, cost disclosure	$0–$9,600	80
SB_	4	Structural bias probes: technology sector, geographic, size, recency framing	$0	80
FP_	8	Framing and psychological biases: anchoring, gain/loss, mental accounting, narrative	$0	160
COMP_	2	Complex multi-factor scenarios	$0	40
Total	**24**	**—**	**$383–$94,000**	**480**

Note. Dollar stakes reflect estimated financial harm from an incorrect recommendation given the specific client circumstances in each scenario. Ground truth was scored by Llama-3.3-70B-Instruct-Turbo under blinded judge scoring protocols.

2.2 Quality feature extraction

Each response was processed by a dual-pass extraction pipeline to yield seven features (Table 3). The primary features—calc_correctness, factor_correctness, directness_confidence, and alternative_evaluation—capture the analytical completeness of the response. Three secondary signals derive from the behavioral fingerprint: the model-specific h_composite on activated dimensions, the ensemble weighted_h, and the same_direction_flag. A V2 quality composite aggregates the four primary features.

2.3 Dollar-weighted evaluation

All ensemble comparisons report both binary accuracy and expected dollar error cost. Dollar cost equals the sum of stakes for scenarios where the strategy selects the wrong response. This framing matters: routing correctly on the $94,000 pension scenario while failing on a $383 BNPL scenario produces a vastly different cost profile than the reverse, and binary accuracy treats them identically.

3. What works: knowing which responses to trust

The most direct deployment need we identified is a reliable signal for which responses to deliver, which to review, and which to route to human oversight — without requiring ground truth. The calibrated confidence framework provides this.

3.1 The confidence model

A logistic regression trained on the seven quality and fingerprint features achieves AUC = 0.876 and Brier score = 0.105, with Platt scaling applied for probability calibration. Cross-validation used GroupKFold stratified by scenario_id, ensuring no responses from the same scenario appeared in both training and test folds. The dominant predictor is calc_correctness (r = +0.861): when a model explicitly performs the relevant calculation and the result is consistent with its recommendation, the response is almost always correct. Quality composite provides strong secondary signal (r = +0.699). The behavioral fingerprint h_composite contributes a weak but directionally consistent prior (r = +0.061).

3.2 Three deployable zones

The confidence distribution is strongly bimodal, enabling clean three-zone routing (Table 2). The TRUST zone ($p \geq 0.6$) covers 19.8% of responses and is correct 100% of the time in the validation data. These can be delivered with a documented confidence attestation and zero human review cost. The REVIEW zone ($p < 0.3$) covers 68.5% of responses and is wrong 93% of the time—delivering these without intervention means delivering errors to clients on nearly every query in this zone.

Table 2. Confidence Zone Performance: Accuracy and Cost by Tier

Zone	Confidence range	n	% of responses	Empirical accuracy	Dollar cost
TRUST	$p \geq 0.6$	95	19.8%	**100%**	$0
UNCERTAIN	$0.3 \leq p < 0.6$	56	11.7%	**60%**	Variable
REVIEW	$p < 0.3$	329	68.5%	**7%**	High

Note. TRUST = $p \geq 0.6$. REVIEW = $p < 0.3$. Calibrated using GroupKFold(5) by scenario_id; n = 480 responses. Dollar cost reflects errors on HS_ and LS_ scenarios only (SB_ and FP_ have $0 stakes in this study).

3.3 What drives low confidence

The feature breakdown from the confidence model is operationally useful beyond the score itself: it identifies why confidence is low. Low q_calc means the model applied heuristics without explicit calculation—the specific failure mode that the quality improvement loop in Section 4 addresses. Low q_factor means the wrong regulatory factor was cited, a compliance risk regardless of whether the recommendation direction is correct. Low q_direct means the model hedged without reaching a conclusion, producing advice a client cannot act on. Each failure mode has a different operational response.

Table 3. Quality Features: Definitions and Correlation with Binary Accuracy

Feature	What it measures	r (accuracy)	Role in model
calc_correctness	Explicit calculation performed and numerically consistent with recommendation	**+0.861*** **	Dominant predictor
quality_total	Composite quality score (calculation, factor, directness, alternatives)	+0.699***	Primary composite
factor_correctness	Correct regulatory factor cited (APR, risk constraint, expense ratio)	+0.333***	Compliance signal
directness_confidence	Commits to a specific recommendation with rationale	+0.152**	Presentation signal
alternative_evaluation	Explicitly considers at least one alternative	+0.142*	Presentation signal
h_composite	Opposing fingerprint bias on activated dimensions	+0.061	Weak prior

Note. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$. Correlations computed within-model (n = 340 responses with ground truth). Cross-model quality comparison inverts these relationships—see Section 6.

4. What works: improving low-confidence responses

Identifying low-confidence responses is only the first step. The deployment question is what to do with them. Human review is expensive. Routing all REVIEW-zone responses to human oversight is impractical at scale. An automated improvement mechanism that is reliable, cheap, and safe — one that never makes things worse — was the operational requirement we defined.

4.1 Why revision does not work

Seven experiments (A through G) tested progressively refined approaches to automated response improvement. All revision-based approaches failed. Post-hoc text augmentation (appending

calculations to existing responses) produced +0.0 average quality improvement and 43%/43% improvement/degradation splits. LLM revision — asking the model to rewrite its response to be more explicit — achieved a 0% increase in accuracy fix rate across a test of 35 tested responses.

The mechanism is consistent: directional bias persists through rewriting. A model that recommended annuity when lump sum was correct, asked to revise its response to include explicit calculation, produced an explicit calculation that still concluded annuity. The bias is embedded in the model's recommendation direction, not in the quality of its presentation. Improving presentation does not change direction.

4.2 Standalone calculation injection

Standalone calculation injection resolves this by separating the analytical task from the original recommendation. When q_calc is the primary gap, the mechanism: (1) extracts the numerical inputs from the response; (2) passes only those inputs and the relevant formula to an LLM with no access to the original recommendation; (3) receives the calculation; (4) checks whether the result is consistent with the original recommendation.

The theoretical grounding: arithmetic does not have a directional preference. A break-even age computation given the correct inputs produces the same numerical result regardless of whether the model that generated the original response preferred annuity or lump sum. By performing the calculation without access to the recommendation, the result is directionally independent. The alignment check between independent calculation and original recommendation is a form of self-consistency checking (Wang et al., 2023) applied to financial advice.

The operational outputs are three. When the calculation supports the recommendation (70% of cases in validation), the response is auto-approved with the calculation appended—quality improved, direction validated. When the calculation contradicts the recommendation (25%), the response is flagged for human review: a potential error has been detected without knowing the correct answer. When the calculation is neutral (5%), the response is auto-approved with the calculation appended.

4.3 Validation pilot results

A 111-response validation pilot across five scenario types validated the full workflow (Table 5). The pilot identified cases where the independent calculation contradicted the model's recommendation on high-stakes retirement and debt scenarios, routing these for human review before any response was delivered. Zero degradation events occurred across the full pilot.

Table 5. Validation pilot: standalone calculation injection results (n = 111 responses)

Metric	Value	Operational meaning
Total responses processed	**111**	Across 5 scenario types, 0 failures

Metric	Value	Operational meaning
Auto-approved (supports/neutral)	**70 (63.1%)**	Delivered with calculation appended; 72% excluding HS_RET_003
Contradiction-flagged	**18 (16.2%)**	Potential error detected; routed to human review
Hallucination-flagged	**14 (12.6%)**	All HS_RET_003 — excluded from injection by rule
Degradation events	**0 (0%)**	Calculations are additive; never replace original response
Original mean calibrated score	**0.059**	REJECT zone before injection
Post-injection mean calibrated score	**0.892**	TRUST zone after calculation appended
Cost per response	**≈$0.01**	Two gpt-4o-mini calls

Note. HS_RET_003 (Social Security timing) excluded from calculation injection by rule: multi-step derivation requires computing breakeven ages across multiple claiming ages—numbers not present in the original response—producing reliable hallucination. All 14 hallucination events in the pilot are HS_RET_003. Excluding this scenario reduces hallucination rate to ≈1% and raises auto-approval to 72%.

4.4 What the loop does and does not do

The quality improvement loop improves the presentation and analytical completeness of responses and detects arithmetic-recommendation inconsistencies. It does not correct directionally wrong responses where the model's calculation is also wrong, or where the error is in factor selection rather than calculation. The 72% auto-approval rate reflects the fraction of low-confidence responses where the calculation is directionally consistent with the recommendation — in most cases, the model reasoned correctly but hedged or failed to show its working. The 16% contradiction rate reflects cases where the arithmetic disagrees with the recommendation, which warrants human review.

5. What works: detecting behavioral change between model versions

AI providers update their models without announcing behavioral changes. A model that passed validation at deployment may behave differently after a silent update. Behavioral fingerprints provide a reliable mechanism for detecting these changes before clients experience them.

5.1 Generational drift evidence

The fingerprint sequence from GPT-5.2 through GPT-5.5 demonstrates that fingerprints accurately track behavioral change and that the direction of drift matches the training objectives that produced the update. D3 (AI disclosure behavior) drifted +1.47 h-units across the sequence, consistent with deliberate RLHF targeting of regulatory compliance. A4 (anchoring susceptibility) amplified +1.18, an unintended side effect of the reasoning architecture—the chain-of-thought process treats price anchors as facts to reason from rather than noise to discount. A6 (status quo bias) reduced −1.31 as the reasoning model explicitly evaluates alternatives rather than defaulting to the current state.

The practical implication: running the sealed battery after each provider update takes approximately seven hours and detects behavioral changes before they express in client outcomes. The direction of detected drift—which dimensions moved and in which direction—provides an interpretable account of what changed in the model's behavior, even when the provider does not disclose specific training changes.

5.2 The MIXED consensus signal

A specific consensus pattern identified through fingerprint analysis provides a reliable error signal without ground truth. When all four models agree on a scenario where their fingerprints predict opposing tendencies — the MIXED genome type — that agreement is almost always wrong (4.5% accuracy, Fisher exact odds ratio = 10.5, $p = 0.031$). When models that normally diverge still converge, something scenario-specific overrides their typical divergence in a way that is near-uniformly wrong.

This finding reverses the naive interpretation of cross-model consensus. Unanimous agreement among models with opposing fingerprint directions is not confirmation—it is a red flag. The MIXED consensus signal can be computed from fingerprint data and applied as an automated escalation trigger: if all models agree on a scenario where they typically diverge, route to human review regardless of individual confidence scores.

6. What does not work: three intuitive approaches that fail

Three approaches to improving AI advisory quality appear plausible from the fingerprinting framework but fail empirically. Understanding why they fail is as important as knowing what works, because practitioners may invest in them independently.

6.1 Ensemble weighting does not improve accuracy

The intuition: if you know each model's biases, you can weight their responses to produce a better ensemble. This does not work for accuracy. Table 4 presents five weighting strategies; none improves binary accuracy above the naive equal-weight baseline of 51.8%. H-calibration — weighting by fingerprint alignment — is actively harmful, costing $197,509 more than naive routing.

The structural reason: 42.4% of scenario-repetitions result in all four models giving the same wrong answer. When all models share a systematic bias—the technology sector preference, the brand recognition effect, the momentum amplification—no weighting scheme can recover the correct answer. The correct answer is not available in the ensemble. The Confidence×Quality approach reduces expected dollar cost by 15.7% ($171,445 on the validation battery) by routing better among available responses, but this is dollar optimization, not accuracy improvement.

Table 4. Ensemble Strategy Comparison: Binary Accuracy and Dollar Cost

Strategy	Binary accuracy	Dollar cost	vs. Naive
Naive (equal-weight)	51.8%	$1,094,000	Baseline
Quality-weighted (V2 score)	50.6%	$1,059,442	+$34,558 (3.2%)
Confidence-weighted (calibrated model)	51.8%	$933,453	+$160,547 (14.7%)
Confidence × Quality	**51.8%**	**$922,555**	**+$171,445 (15.7%)**
H-calibration (fingerprint weights)	40.0%	$1,291,509	−$197,509 (worse)

Note. None of the five strategies improves binary accuracy above the 51.8% naive baseline. The 42.4% joint failure rate—scenario-repetitions where all four models give the same wrong answer—is the structural ceiling that prevents accuracy improvement through weighting.

6.2 Quality metrics invert across models

The intuition: the highest-quality response is the most reliable. This holds within a model ($r = +0.699$) but inverts across models ($r = -0.945$). This three-level Simpson's paradox reflects a genuine confounding: model identity determines both typical quality level and baseline accuracy jointly. Claude produces high-quality responses (well-calculated, explicitly cited, committed recommendations) at a lower accuracy rate than GPT-5.5, which produces responses that score lower on quality features but are

correct more often. Selecting the highest-quality response across models systematically selects Claude when Claude is wrong.

The practical prescription: calibrate quality metrics within a model, not across models. A within-model quality score is valid for identifying which of a single model's responses to trust. A cross-model quality comparison does not identify which model to trust and should not be used for that purpose.

6.3 Fingerprint h-values cannot predict instance-level failure

The intuition: if a model has a large h-value opposing the correct answer on an activated dimension, it is more likely to fail on that response. Cross-scenario failure prediction from h_composite achieves AUC = 0.359—substantially worse than random. Dollar-cost prediction is similarly uninformative (Pearson r = −0.110, p = 0.277). Multiple text-based approaches—corpus classifiers, embedding similarity, TF-IDF bias detection—all performed at or below chance when used for instance-level failure prediction.

The reason is that fingerprints characterize what a model does on average across many controlled trials. They correctly describe the tendency. They do not predict any specific response's outcome because bias is directional—embedded in recommendation direction, not in response quality—and quality features cannot detect directional errors without independent calculation. Table 6 summarizes what was tested and why each approach fails.

Table 6. What was tested and did not work: approaches and failure mechanisms

Approach	Result	Why it fails
H-value ensemble weighting	−$197,509 vs. naive	Fingerprints predict average tendencies, not instance correctness. Anti-predictive when used per-response.
Quality score cross-model comparison	r = −0.945 (inverted)	Simpson's paradox: model identity confounds quality-accuracy relationship across models.
Post-hoc text augmentation	+0.0 average improvement	Judges score integrated reasoning, not appended text. Calculations bolted on after the fact do not change reasoning quality.
LLM revision of existing responses	0% accuracy fix rate	Directional bias persists through rewriting. A model that recommended annuity revises to recommend annuity more explicitly.

Approach	Result	Why it fails
Fingerprint h-values for failure prediction	AUC = 0.359 (worse than random)	Fingerprints characterize average tendencies across many trials, not outcome of any specific response.
Any ensemble approach for accuracy improvement	No strategy improves above 51.8%	42.4% joint failure rate: when all models share a bias, no weighting recovers the correct answer.

Note. All six approaches were tested on the 480-response validation dataset or in controlled experiments. Failure rates and mechanisms are empirical, not theoretical.

7. Commercial applications

The empirical findings in this paper form the evidentiary foundation for a suite of commercial tools. These tools translate the research programme's validated mechanisms into deployable products for financial institutions. Intellectual property protection covering the core methodologies has been filed by Human Machines Group LLC.

7.1 Pre-deployment evaluation and drift monitoring

The behavioral fingerprinting methodology (Gibbins, 2026) enables systematic pre-deployment evaluation of any AI financial advisory model. A sealed, pre-registered stimulus battery is administered before deployment, producing a comprehensive behavioral profile: accuracy by scenario type, dollar impact of failures, and comparison against all previously tested models. This establishes the baseline fingerprint for the deployed model.

The same battery, re-administered after any model update, functions as a drift detection system. The generational sequence documented in Section 5 — GPT-5.2 through GPT-5.5, with D3 drifting +1.47 and A4 amplifying +1.18 between versions — demonstrates that behavioral drift is detectable, directionally interpretable, and linked to specific training decisions. Financial institutions can use this methodology to verify that a model update has not materially altered the behavioral profile they validated at deployment. Intellectual property protection covering this methodology has been filed.

7.2 Per-response confidence scoring and quality improvement

The calibrated confidence model (Section 3) and quality improvement loop (Section 4) together form a per-response quality assurance system. Every AI advisory response can be scored in milliseconds using extractable features, classified into TRUST, UNCERTAIN, or REVIEW zones, and — for REVIEW-zone responses with identified quality gaps — automatically improved through the standalone calculation injection mechanism.

In production, this system operates as three layers. The diagnostic layer runs on every response, producing a calibrated confidence score and a feature-level breakdown of what drove it. The action layer runs on REVIEW-zone responses: for calculation-gap responses, standalone calculation injection produces either an auto-approved response with calculation appended (72% of cases) or a human-review flag when arithmetic contradicts recommendation (16% of cases, confirmed real errors in pilot). For multi-model deployments, a selection layer applies Confidence×Quality weighting to choose among available responses, reducing expected dollar error cost by 15.7% against naive routing.

Intellectual property protection covering this quality assurance system has been filed.

7.3 Input-side de-biasing

The finding that directional bias cannot be corrected after generation motivates a complementary input-side approach: neutralising bias-activating inputs before the query reaches the model. Controlled experiments demonstrated that removing an anchoring reference number from a savings rate query improved GPT-5.4 accuracy from 60% to 100% on anchor-activated scenarios. This approach is applicable to any input feature that reliably activates a documented model bias: reference numbers for anchoring-susceptible models, current-portfolio framing for status-quo-biased models, recent performance figures for recency-biased models.

The practical deployment is a pre-processing layer that inspects incoming queries against the deployed model's fingerprint, identifies bias-activating features, and neutralises them before the query reaches the model. Intellectual property protection covering this pre-processing methodology has been filed.

7.4 Capital flow intelligence and bias-calibrated ensemble analysis

Two further applications extend beyond single-institution deployment. The capital flow intelligence application derives forward-looking indicators of AI-advised retail capital flows from behavioral fingerprint data. As documented in Gibbins (2026), the universal biases shared across platforms — technology underweighting in portfolio construction (22–25% versus 32% benchmark), brand recognition concentration ($h = 1.37–2.17$ across all platforms), and momentum amplification—will increasingly direct retail capital in predictable directions as AI advisory adoption scales. Measuring those fingerprints in advance of the flows enables leading-indicator signals measurable before they appear in fund flow data.

The bias-calibrated ensemble analysis application addresses multi-LLM investment research workflows: using fingerprint knowledge to decompose cross-model agreement into genome-predicted components (correlated bias) and residual components (genuine independent signal), and weighting agreement by genome independence rather than treating all consensus as equally informative. The MIXED consensus finding in Section 5 — where unanimous agreement among models with opposing fingerprint directions achieves only 4.5% accuracy — provides

direct empirical grounding for this decomposition. Intellectual property protection covering both the capital flow indicator methodology and the bias-calibrated ensemble approach has been filed.

8. Discussion

8.1 The directional nature of AI advisory bias

The consistent finding across all failed approaches is that LLM financial advisory bias is directional. It lives in which option the model recommends, not in how well it presents that recommendation. A model that recommends annuity when lump sum is correct will, when asked to be more explicit, produce an explicit recommendation for annuity. Text classifiers, quality scorers, and embedding similarity measures all operate on presentation features and cannot detect directional errors.

The quality improvement loop works because it exploits the one channel that is directionally independent: arithmetic. A break-even calculation, performed without access to the recommendation, is not biased by the model's directional preference. This is why seven revision-based experiments failed and the standalone calculation injection succeeded.

8.2 The two levels of fingerprint validity

Behavioral fingerprints are valid at two levels and invalid at a third. At the population level — characterizing what a model does on average across many queries — fingerprints are accurate and stable. At the drift level, tracking how a model's behavioral profile changes between versions, fingerprints are reliable leading indicators. At the instance level, in predicting whether this specific response to this specific query is correct, fingerprints fail and perform below chance. Practitioners should use fingerprints for what they measure: aggregate tendencies and version-level change, not instance-level quality.

8.3 The dollar-weighted reframe

Dollar-weighted evaluation changes the policy prescription relative to accuracy-only framing. The Confidence×Quality ensemble saves 15.7% of expected dollar error cost with no binary accuracy improvement. This is valuable—it means the approach correctly routes away from the expensive errors even when it cannot reduce the count of errors. Financial advisory evaluation should always be dollar-weighted because not all errors are equally consequential, and the 93% wrong rate in the REVIEW zone matters most on high-stake queries.

9. Limitations

The calibration model was validated on the 24-scenario battery with 48 conditions. Out-of-sample performance on scenario types outside the battery has not been confirmed. The TRUST zone's 100% accuracy is an in-distribution estimate that should be re-validated as new scenario types are added to the deployment environment.

The standalone calculation injection pilot covered 111 responses across five scenario types. Generalizability to other scenario types and real client query distributions has not been established. The mechanism is validated only for q_calc-gap responses; extension to q_factor and q_alts gaps awaits further validation.

All four models tested are English-language frontier systems from US-based providers. The fingerprinting methodology has not been applied to non-English advisory contexts or to models from other providers.

Conflicts of interest: the author is the founder of Human Machines Group LLC, which has filed intellectual property protection covering methodologies described in this paper. The study was independently pre-registered before data collection.

10. Conclusion

Three things work for improving investment advice from AI financial advisors. A calibrated confidence model identifies which responses to trust (TRUST zone: 100% accuracy, 19.8% of responses) and which to flag (REVIEW zone: 7% accuracy, 68.5% of responses). A standalone calculation injection mechanism improves low-confidence responses at $0.01 each, detects potential errors at a 16% rate without knowing the correct answer, and never degrades existing correct responses. And behavioral fingerprints reliably track model behavioral change between versions, enabling drift detection before clients experience it.

Three things do not work as practitioners might expect. Ensemble weighting cannot improve binary accuracy when 42.4% of scenarios produce unanimous wrong answers from all models. Quality metrics, valid within a model, invert when compared across models. Fingerprint h-values, accurate for aggregate characterization, predict instance-level failure worse than chance.

Together, the findings from this paper and Gibbins (2026) establish a complete framework for AI financial advisory governance: fingerprinting characterizes the behavioral profile at deployment and tracks its evolution across versions; confidence calibration routes responses to appropriate handling at inference time; and calculation injection provides automated error detection and

quality improvement without ground truth. Each component has been empirically validated. The governance challenge is not that AI advisor bias is unmeasurable — this paper demonstrates it is. The challenge is that current regulatory and institutional frameworks do not yet require its measurement.

References

Gibbins, G. (2026). How AI moves markets: Systematic biases and behavioral divergence in major LLMs and the implications for capital flows and governance. Human Machines Group LLC. SSRN.

Kahneman, D. (2011). Thinking, fast and slow. Farrar, Straus and Giroux.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. Journal of the Royal Statistical Society, Series B, 13(2), 238–241.

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

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models. International Conference on Learning Representations (ICLR).

Winder, P., Hildebrand, C., & Hartmann, J. (2025). Biased echoes: Large language models reinforce investment biases and increase portfolio risks of private investors. PLOS One, 20(6): e0325459.

Lee, H., Seo, J., Park, S., Lee, J., Ahn, W., Choi, C., Lopez-Lira, A., & Lee, Y. (2025). Your AI, not your view: The bias of LLMs in investment analysis. Proceedings of ICAIF ’25. https://doi.org/10.1145/3768292.3770375