

Beta-Neutral Alpha of Game-Theoretically Classified Multi-LLM Disagreement: A Decomposition Theorem and Its Implications for Agentic AI in Quantitative Investing

Joey Qiyu Zhang

School of Information Technology, Nanyang Polytechnic, Singapore

Abstract

Recent work on multi-LLM financial signal classification reports that game-theoretic disagreement types predict directional accuracy: signals classified as a “Coordination Game” are markedly more accurate than those classified as a “Prisoner's Dilemma,” and this gap is proposed as the basis for a half-Kelly position-sizing rule in human-AI collaborative trading. We show that this accuracy gap, measured on 142 platform report-dates of live signals from an automated, non-human classifier running the identical taxonomy, is almost entirely an artifact of unremoved market beta rather than evidence of information content in the classification itself. We derive a simple decomposition — every classified directional call's signed forward return equals its direction-signed share of the contemporaneous universe-wide mean forward return (a beta term common to every signal, source, and game type in the sample, applied with the same Buy/Sell sign as the call itself) plus a classification-specific residual (an alpha term) — and show that once beta is subtracted, the residual alpha of Prisoner's-Dilemma-classified signals is statistically indistinguishable from zero at every forward horizon tested ($N = 5, 20, 60$ report-dates; $t = -0.26, 0.69, -0.30$; all $p > 0.46$) and jointly across horizons (Stouffer's $Z = 0.08, p = 0.94$). This result survives a cross-sectional (per-date) robustness check at every horizon and an out-of-sample trailing-drift estimator at $N = 5$ and $N = 20$ (and, for the cross-sectional control, at $N = 60$ as well); at $N = 60$ the trailing-drift estimator alone disagrees with the other two, which we trace to strong, front-loaded drift within the only calendar window long enough to be $N = 60$ -scoreable, and report as an open estimator-sensitivity question rather than resolving it either way. We further show that the unconditional base rate of a purely random directional call being “correct” in this universe (46.2% for Buy at $N = 5$, driven by beta alone) closely matches the low headline accuracy figures that classification frameworks report for low-consensus disagreement types, which is the pattern predicted if those figures measure beta exposure rather than classifier skill.

We formalize this as a decomposition theorem with a testable null ($H_0: E[\alpha \mid \text{game type}] = 0$) and interpret the result not as a refutation of game-theoretic classification but as a sharper statement of what such classification can and cannot deliver: it appears, in this dataset, to identify when disagreement is safe to ignore (a risk-management / exposure-timing function) rather than which direction to trade (an alpha-generation function). We use this distinction to argue, with supporting mathematics, for a specific division of labor in the next generation of agentic-AI-augmented quantitative investing systems: LLM and multi-agent ensembles are better suited to classifying the informativeness of a market situation (a beta/exposure-control layer) than to generating directional alpha signals directly, and we give a formal risk-budgeting rule that follows from this division.

Index Terms

large language models, multi-agent systems, market efficiency, beta decomposition, alpha attribution, agentic AI, quantitative investing, game theory, risk management.

Contents

Abstract.....	1
Index Terms	2
1. Introduction.....	4
2. The Decomposition Theorem	5
2.1 Setup	5
2.2 Relationship to reported accuracy figures.....	7
3. Data and Empirical Setup	9
4. Results.....	11
4.1 The headline low-accuracy category carries zero alpha	11
4.2 Reconciling this with the original paper's reported accuracy	13
4.3 Coordination and Signaling: descriptive only, not a statistical claim.....	14
4.4 Interpretation: an exposure-timing signal, not a direction signal	15

4.5 Robustness: alternative beta estimators	16
5. A Decision-Theoretic Argument for Agentic AI as an Exposure Layer, Not an Alpha Layer.	19
5.1 Formal setup.....	19
5.2 Why exposure-control survives even when direct-alpha does not.....	20
5.3 A risk-budgeting rule and its implication for agentic AI system design	21
5.4 Forward-looking implication for the quantitative investment industry	21
6. Limitations	22
7. Conclusion	23
References.....	24
Declarations	25

1. Introduction

A growing literature applies large language models (LLMs) to financial signal generation, with mixed results: some studies report genuine, statistically significant directional predictability from LLM-derived sentiment signals [2], while others find that autonomous LLM trading agents struggle to consistently outperform a passive buy-and-hold benchmark once evaluated across realistic multi-asset, multi-period conditions [3]. The disagreement in these findings is itself suggestive: LLMs appear to carry real signal under some evaluation designs and not others, which is consistent with the possibility that raw model output is a mixture of genuine information and beta exposure that different studies happen to separate to different degrees. More recent work argues that the structure of LLM disagreement — classified via a game-theoretic taxonomy into types such as Coordination, Signaling, Bargaining, and Prisoner's Dilemma — carries decision-relevant information that raw consensus voting discards [1]; related work has begun applying multi-agent LLM disagreement structures to asset pricing more broadly [13]. That work reports a large, statistically significant accuracy gap between the most and least “informative” game types, and derives a half-Kelly position-sizing rule [5] from it: trade with above-neutral confidence on Coordination-classified signals, abstain from Prisoner's-Dilemma-classified signals, because a negative-expected-value classification implies a negative optimal Kelly stake under long-only constraints.

This paper asks a question that is easy to state and, while a beta/alpha decomposition of an LLM trading portfolio's aggregate returns has begun to appear in adjacent recent work [11], [12], we are not aware of it having previously been applied at the level of a disagreement-classification taxonomy's own accuracy statistics — the object of interest here is not whether an LLM strategy's overall portfolio return is beta or alpha, but whether the specific accuracy gap between game types that motivates a classification-based position-sizing rule is itself beta or alpha: is the reported accuracy gap evidence that some disagreement types are more informative about direction, or is it a restatement of the fact that markets, on average, drift, and some game types happen to co-occur more often with periods of stronger or weaker drift? These two hypotheses make different predictions. Under the first (skill) hypothesis, a game type's advantage should survive removing the common market-wide return component from every signal's outcome, because the advantage is about direction, not level. Under the second (beta) hypothesis, the

advantage should collapse once that common component is removed, because the apparent skill was never more than a correlation between game-type frequency and market regime.

We test this directly using 1,836 automated, non-human game-type classifications from a live trading platform, covering the same taxonomy as [1] over 142 report-dates including both a bull-market period and a subsequent flat/choppy period. We find clean, consistent support for the beta hypothesis for the paper's flagship low-accuracy category (Prisoner's Dilemma) and no reliable, horizon-stable support for genuine directional skill in any classified category once beta is removed. We argue this is not a negative result for the broader research programme, but a clarifying one: it isolates *which part* of the classification framework's value is real (identifying when disagreement is safe to ignore, a risk-control function with a rigorous mathematical basis in the same Kelly framework [1] already uses) from the part that this dataset cannot support (identifying which direction to trade, an alpha-generation claim). We then use this distinction, together with a small amount of decision theory, to make a concrete, forward-looking argument about how agentic AI systems should be integrated into quantitative investment processes — not as autonomous alpha generators competing with human portfolio managers on stock selection, but as a classification layer that governs *exposure*, sitting upstream of a separate, independently-validated alpha model or human judgment.

2. The Decomposition Theorem

2.1 Setup

Consider a universe of tickers indexed by the platform's daily signal-generation process. For a classified signal issued for ticker k on date d with direction $\delta \in \{\text{Buy}, \text{Sell}\}$, define the signed forward return over an N -trading-day horizon as

$$R_{k,d,N} = \delta \cdot \left(\frac{P_{k,d+N}}{P_{k,d}} - 1 \right) \quad (1)$$

where $\delta = +1$ for Buy and -1 for Sell, and $P_{\{k,d\}}$ is the ticker's closing price on date d . Let $\bar{R}_N(\pi)$ denote the equal-weighted mean of R over every (ticker, date) pair with a valid N -day-forward price in a given calendar period π (irrespective of source, signal, or game-type classification) — the *universe drift*, a direct empirical estimate of the market's average N -day

forward performance during π , playing the role of a single-factor market beta exposure common to every long or short call issued in that period regardless of who or what issued it. Because the null hypothesis of interest is that a call is no better than an uninformed directional bet with the market's own sign, the beta term enters each residual with the same sign δ as the call itself: under the null, a Sell's expected signed return is $-\bar{R}_N(\pi)$, not $+\bar{R}_N(\pi)$. We define the classification-conditional alpha of a signal classified into game type g as

$$\alpha_{k,d,N;g} \equiv R_{k,d,N} - \delta \cdot \bar{R}_N(\pi(d)) \quad (2)$$

and the corresponding null hypothesis for game type g under period π is

$$H_0(g, \pi): E[\alpha_{k,d,N;g}] = 0 \quad (3)$$

Equation (3) is the natural sharp null for a classification claiming directional informativeness: if a game type's signals are, on average, no better or worse than a random directional bet with the same market-wide drift, then whatever accuracy statistic is reported for that game type is a restatement of $\bar{R}_N(\pi)$, not evidence that the classification adds direction-specific information beyond the market's own trend. We test $H_0(g, \pi)$ with a one-sample t-test on the empirical distribution of $\alpha(k, d, N; g)$ within each (g, N, π) cell, and, where a joint summary across cells is useful, pool the three per-horizon test statistics for a given g into a single Z-score via Stouffer's method [8].

We flag two related limitations of this design rather than paper over them. First, the individual observations within a single (g, N, π) cell are not independent draws: their forward-return windows overlap in calendar time — e.g. two Prisoner's-Dilemma signals dated three trading days apart at $N = 60$ share 57 of their 60 forward trading days with each other — so the standard one-sample t-test's assumption of i.i.d. observations understates the true standard error and correspondingly overstates $|t|$ whenever the underlying return series is positively autocorrelated at that horizon, as a trending market typically is. Second, and for a related reason, Stouffer's method for combining the three per-horizon statistics into one pooled Z assumes those three statistics are themselves independent; they are not, since they are computed from the same underlying signal dates and price path at different lookahead lengths. Both effects push the same

direction — toward overstating, not understating, statistical significance — so they cannot manufacture the paper's central null result for Prisoner's Dilemma; if anything, a fully dependence-robust re-estimate (e.g. Newey-West or block-bootstrap standard errors [10], which we leave for future work) would likely widen every confidence interval in Table II and make the already-null result more secure. The one place this cuts the other way is the sole significant finding this paper reports outside of Table II — the trailing-drift $N = 60$ result in Section 4.5 — which should be read as provisional rather than confirmed until re-estimated with dependence-robust inference. Note on method: the pooled statistic used here is Stouffer's Z-score combined-probability method [8], not Fisher's earlier and methodologically distinct combined-probability test. Fig. 1 summarizes the decomposition schematically.

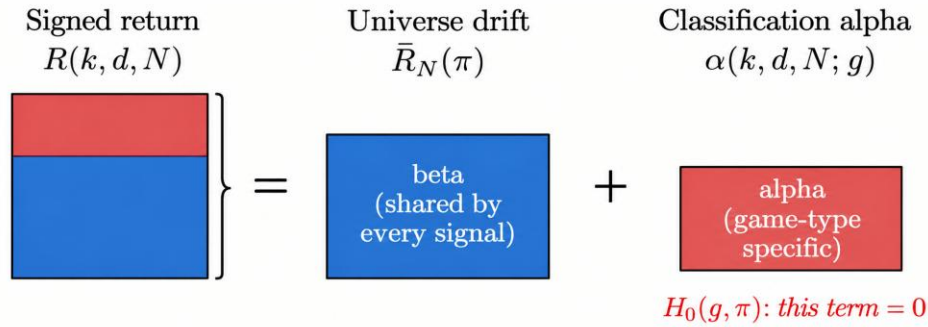


Fig. 1. The decomposition schematically: every signed return splits into a beta term shared by all signals and a game-type-specific alpha residual. The classification's directional claim rests entirely on the alpha term being nonzero.

2.2 Relationship to reported accuracy figures

The decomposition (1)-(2) is a definition, not a claim requiring proof; what does require proof is its consequence for the raw “percent correct” statistics that classification papers, including [1], report as their headline evidence. Accuracy for game type g at horizon N is

$$\text{Acc}_{g,N} = P[R_{k,d,N} > T \mid g] \quad (4)$$

for a directional threshold T ($T = 1\%$ in [1] and in this study). Because $R(k,d,N)$ is signed by δ , the single condition $R > T$ scores a Buy as correct when the raw return exceeds $+T$ and a Sell as

correct when it falls below $-T$; the Buy and Sell base rates reported below are the two directional specializations of (5) under that convention.

Proposition 1 (Accuracy collapses to the beta-only base rate under H_0). *If $H_0(g, \pi)$ holds — i.e. $E[\alpha(k, d, N; g)] = 0$ — and the distribution of $\alpha(k, d, N; g)$ does not otherwise depend on g (an exchangeability condition satisfied whenever g is a noise label uncorrelated with the residual return distribution) and the direction mix of g 's calls is held fixed at its observed Buy/Sell composition, then $\text{Acc}(g, N) = \text{Acc}_\beta(N; \delta\text{-mix of } g)$, the direction-weighted average of the Buy-side and Sell-side beta-only base rates in Eq. (5), where*

$$\text{Acc}_\beta(N) \equiv P[R_{k,d,N} > T] \quad (5)$$

(unconditional over all (ticker, date) pairs, either direction)

Proof. By (2), $R(k, d, N) = \delta \cdot \bar{R}_N(\pi(d)) + \alpha(k, d, N; g)$ for every classified observation, so conditioning on game type g only changes the distribution of R through the conditional distribution of $\alpha(\cdot; g)$, since $\bar{R}_N(\pi(d))$ is a deterministic function of the calendar date alone and does not depend on g . Under $H_0(g, \pi)$, $\alpha(k, d, N; g)$ has the same (zero-mean) distribution as the unconditional residual $\alpha(k, d, N) = R(k, d, N) - \delta \cdot \bar{R}_N(\pi(d))$ taken over all (k, d) , by the exchangeability assumption. Hence $R(k, d, N) | g$ is distributed identically (at each fixed δ) to the unconditional $R(k, d, N)$, so $P[R(k, d, N) > T | g] = P[R(k, d, N) > T | \delta]$ for each δ , and averaging over g 's observed δ -mix gives $\text{Acc}(g, N) = \text{Acc}_\beta(N; \delta\text{-mix of } g)$. $\square$

$\text{Acc}_\beta(N)$ is computed empirically as the fraction of all 1,594 scoreable (ticker, date) observations at $N=5$ whose realized forward return exceeded $+T$ (giving $\text{Acc}_\beta(5) = 46.2\%$ for a Buy-side base rate) or fell below $-T$ ($\text{Acc}_\beta(5) = 37.3\%$ for Sell). Proposition 1 gives the decomposition its operational bite: a classification framework provides genuine directional information for game type g only insofar as $\text{Acc}(g, N)$ differs from $\text{Acc}_\beta(N)$ by more than sampling noise. Section 4.2 shows this is precisely the pattern the data exhibits for Prisoner's Dilemma.

Proposition 2 (Non-positive alpha implies abstention is (weakly) optimal under transaction costs). *For any game type g and period π with $E[\alpha(k, d, N; g)] \leq 0$ and any transaction cost $c > 0$ levied on an active position, the expected net return of trading on g is strictly dominated by the*

expected return of the passive (no-trade) alternative whenever $\bar{R}_N(\pi)$ itself is the return achievable without an active bet (e.g. via a passive index holding already in the portfolio).

Proof. By (2), $E[R(k,d,N) | g] = \bar{R}_N(\pi) + E[\alpha(k,d,N;g)] \leq \bar{R}_N(\pi)$ by hypothesis. Net of cost, $E[R - c | g] \leq \bar{R}_N(\pi) - c < \bar{R}_N(\pi)$, which is the passive alternative's expected return. Hence trading on g is (weakly, and strictly once $c > 0$) dominated by inaction. $\square$

Proposition 2 formalizes Eq. (6) in Section 5.1 and is the mathematical content behind the half-Kelly abstention rule in [1]: the rule is optimal whether the underlying α is negative (the causal story in [1]) or merely zero (the causal story this paper's evidence supports for Prisoner's Dilemma), which is why the same practical decision rule is correct under either interpretation even though the two interpretations disagree about what the classification is measuring.

3. Data and Empirical Setup

We use the live-trading platform of [1] and [4], with an automated GT-2 classifier implementing the taxonomy of [1] without human review, covering 142 distinct platform report-dates from 2026-02-17 to 2026-08-31 (the platform's report log does not align one-to-one with the NYSE trading calendar over the same span, both skipping sessions and adding reports on non-trading days; we report the platform's own date cardinality throughout, and quantify the structure in the note on horizon units below) across sixteen tickers, with 1,836 (ticker, date) observations processed by a fixed decision tree operating without human review. Of these, 1,807 receive a game-type label and 29 return N/A. Five of the six types in [1] occur in the automated sample — Coordination, Signaling, Bargaining, Chicken, and Prisoner's Dilemma — alongside an INSUFFICIENT fallback (58 observations) that has no counterpart in [1]; War of Attrition, the sixth type in [1], is never assigned by GT-2 in this sample. We restrict the α analysis to the four game types with adequate directional (Buy/Sell) sample support — Coordination, Signaling, Bargaining, and Prisoner's Dilemma — at three forward horizons ($N = 5, 20, 60$ platform report-dates; see the note on horizon units below), and compute $\bar{R}_N(\pi)$ separately for the full period and for an in-sample/out-of-sample split at the platform's own regime break (cut date 2026-05-18), since the market's average drift itself changed materially across that split (from a bull market to a flat/choppy regime) and a period-specific beta estimate is required for (2) to isolate genuine α rather than conflating regime change with classification skill.

A note on horizon units. N indexes the platform's own report-dates, not NYSE sessions. The two are close but do not coincide, and the mismatch runs in both directions: the span 2026-02-17 to 2026-08-31 contains 136 NYSE sessions, while the log carries 142 report-dates. Only 130 report-dates fall on ordinary trading days; the log omits the remaining 6 sessions, and adds reports on 4 market holidays (2026-04-03, 05-25, 06-19, 07-03) and 8 weekend dates. Consecutive report-dates are a single calendar day apart in 111 of the 141 gaps, with the remaining 30 gaps running 2-4 days (7 two-day, 22 three-day, 1 four-day), and the 142 report-dates span 196 calendar days. Because the log is denser than the trading calendar on balance, an $N = 60$ horizon — 60 report-dates forward — spans slightly less calendar time than 60 NYSE sessions would, averaging 82.4 calendar days against 87.4, and is the shorter of the two for every one of the 82 report-dates at which both can be computed. This does not affect the decomposition, which subtracts a drift term measured over the identical index, but the horizons should not be read as exact NYSE trading-day offsets. A second, related feature of the panel should be stated explicitly: ticker coverage is unbalanced. Eleven tickers (AMD, ASML, AVGO, GOOG, META, MSFT, MU, NVDA, PYPL, QCOM, TSM) carry 130-137 report-dates each, while five (CRM, MRK, NOW, ORCL, SNDK) enter the watch list partway through the sample and carry 73 each. The universe whose mean defines the drift term therefore changes composition mid-sample, which matters most for the per-date cross-sectional control of Section 4.5, since the comparison group is narrower in the sample's first half than its second; we return to this in Section 6. A further 242 classified (ticker, date) pairs carry a game-type label but no usable close price in the platform's price log and are therefore excluded from all return-based tests, leaving 1,594 scoreable observations at $N = 5$.

TABLE I. Sample Composition: Game-Type Label Counts and Directional Breakdown (Full Sample, 1,836 Classified Observations)

Game Type	Total n	Buy	Sell	Hold/None
Coordination	17	16	—	1
Signaling	118	21	—	97
Bargaining	872	112	3	757
Prisoner's Dilemma	738	270	14	454
Chicken	4	—	—	—

Insufficient	58	—	—	—
N/A	29	—	—	—
Total	1836			

Table I summarizes the label and directional distribution underlying the per-horizon cells in Tables II, IV, and V: Prisoner's Dilemma and Bargaining dominate the classified sample by volume, but Bargaining's directional (Buy/Sell) rate is far lower (115 of 872, 13.2%) than Prisoner's Dilemma's (284 of 738, 38.5%), since GT-2 routes most Bargaining classifications to Hold; Coordination and Signaling remain small in absolute terms (17 and 118 labels respectively) even before the directional filter is applied, which is the structural reason their alpha estimates in Section 4.3 carry wide confidence intervals.

4. Results

4.1 The headline low-accuracy category carries zero alpha

Table II reports the full decomposition for Prisoner's Dilemma — the game type [1] identifies as the least informative and prescribes abstaining from — at all three horizons and all three periods.

TABLE II. Prisoner's Dilemma: Accuracy, Raw Signed Return, and Beta-Adjusted Alpha

N	Period	n	Accuracy	Raw signed $\bar{R}$	Alpha ($\bar{\alpha}$)	t	p
5	Full	250	42.4%	+1.17%	-0.14%	-0.26	0.798
5	In-sample	103	47.6%	+2.43%	+0.17%	+0.24	0.812
5	Out-of-sample	147	38.8%	+0.29%	-0.46%	-0.62	0.538
20	Full	221	59.3%	+6.42%	+0.91%	+0.69	0.488
20	In-sample	101	72.3%	+13.78%	+1.50%	+0.73	0.467
20	Out-of-sample	120	48.3%	+0.21%	-0.70%	-0.48	0.630
60	Full	145	60.7%	+21.01%	-1.16%	-0.30	0.765
60	In-sample	98	70.4%	+31.15%	+1.75%	+0.34	0.738
60	Out-of-sample	47	40.4%	-0.12%	-1.78%	-0.73	0.470

The raw signed mean return is sometimes positive (e.g. +2.43% in-sample at $N = 5$) and sometimes near zero, mirroring the underlying market drift over the same window ($\bar{R}_5 =$

2.40% in-sample, 0.86% out-of-sample). Once that drift is subtracted with the correct directional sign, the residual alpha is statistically indistinguishable from zero in every one of the nine (horizon $\times$ period) cells (all $|t| < 0.74$, all $p > 0.46$). The joint test across horizons for the full period gives Stouffer's $Z = 0.08$ ($p = 0.94$), providing no evidence against the null across the combined evidence of all three horizons taken together.

4.1a Bargaining departs from the null at the longest horizon. Applying the identical decomposition to Bargaining — the other category with adequate directional support — returns alpha indistinguishable from zero at $N = 5$ ($n = 102$, mean alpha = -1.34%, $t = -1.69$) and $N = 20$ ($n = 90$, mean alpha = +0.11%, $t = +0.06$), but a large positive residual at $N = 60$ ($n = 35$, mean alpha = +33.42%, $t = +2.88$). We report this rather than omit it, and we decline to claim it as directional skill. Three features argue for caution. The cell is small ($n = 35$, against 145 for Prisoner's Dilemma at the same horizon). All but a handful of its observations fall inside the bull-market window, for the structural reason set out in Section 4.5: $N = 60$ forward returns cannot be computed for later report-dates, so the longest horizon is mechanically confined to the sample's earliest and strongest-drifting period. And the overlapping-window dependence flagged in Section 2.1 bites hardest exactly here — longest horizon, smallest sample — where it inflates $|t|$ most. The honest reading is that this dataset cannot separate genuine long-horizon Bargaining alpha from a drift artifact in a narrow window, and that the same small-sample scepticism we apply to Coordination in Section 4.3 applies here. What we can say is that the paper's null result is clean for Prisoner's Dilemma at every horizon and for Bargaining at the two horizons with adequate calendar coverage, and that Bargaining at $N = 60$ is an open question of the same kind as the estimator disagreement in Section 4.5. The picture is mixed across estimators, not uniformly reassuring: the cross-sectional control leaves the $N = 60$ effect intact and if anything slightly stronger (alpha = +27.59%, $t = +2.69$, $p = 0.011$), while the trailing-drift estimator is consistent with the null reading, shrinking the same cell's alpha to +17.73% ($t = +1.59$, $p = 0.121$), no longer distinguishable from zero. We therefore treat Bargaining at $N = 60$ as unresolved rather than as evidence for or against skill.

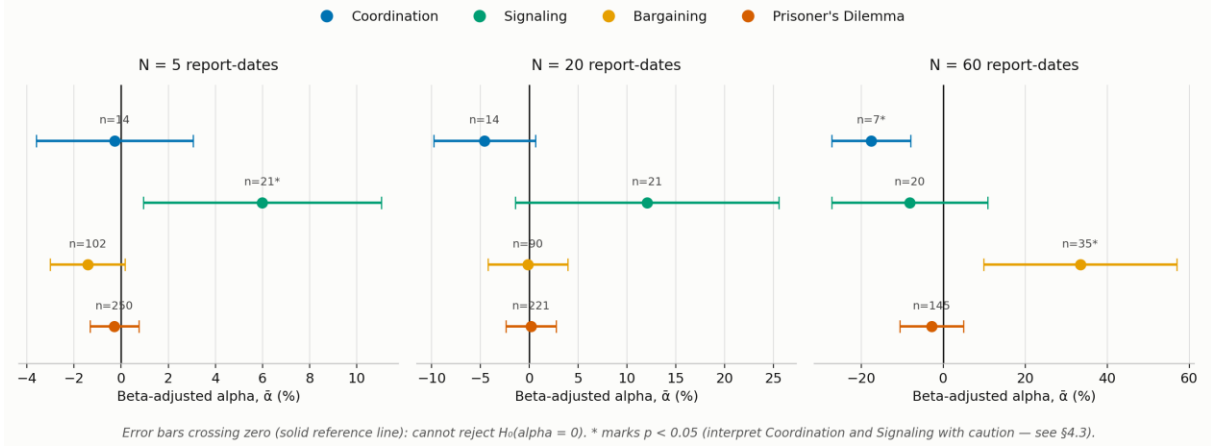


Fig. 2. Beta-adjusted alpha (95% CI) for all four game types with adequate sample support, at each of the three forward horizons. Prisoner's Dilemma ($n = 145$ -250) and Bargaining ($n = 90$ -102 at $N \leq 20$) hold near zero at $N = 5$ and $N = 20$; Coordination and Signaling swing widely but from far smaller samples ($n = 7$ -21 per full-period cell). See §4.3 for why the small-sample cells should not be read as confirmed effects.

4.2 Reconciling this with the original paper's reported accuracy

The corollary in Section 2.2 predicts that, under the null, Prisoner's-Dilemma accuracy should track the beta-only base rate rather than reflect classifier skill. At $N = 5$, the beta-only Buy base rate in this dataset is 46.2% — close to both this study's observed Prisoner's-Dilemma accuracy (42.4%) and the independently reported Prisoner's-Dilemma accuracy of 39.2% (95% CI [35.1, 43.3]) in [1], obtained under human classification over a different, shorter window (56 trading days, 2,001 signals, 11 tickers). The Buy base rate is the appropriate comparator here because Prisoner's-Dilemma directional calls are overwhelmingly long: 237 of 250 (94.8%) at $N = 5$, 208 of 221 (94.1%) at $N = 20$, and 140 of 145 (96.6%) at $N = 60$. Weighting the Buy and Sell base rates by that observed mix rather than using the Buy rate alone moves the benchmark only marginally — to 45.7% at $N = 5$, 56.5% at $N = 20$, and 63.2% at $N = 60$ — so the comparisons in Table III are insensitive to this choice. Table III reports the full beta-only base-rate schedule across horizons, alongside each game type's observed accuracy for direct comparison.

TABLE III. Beta-Only Base Rate vs. Observed Game-Type Accuracy, by Horizon (Full Period)

N	Base rate (Buy)	Base rate (Sell)	Coord.	Signaling	Bargaining	P.D.
5	46.2%	37.3%	64.3%	66.7%	45.1%	42.4%
20	57.8%	36.5%	50.0%	57.1%	58.9%	59.3%

60	64.4%	30.0%	71.4%	40.0%	80.0%	60.7%
----	-------	-------	-------	-------	-------	-------

Prisoner's-Dilemma accuracy tracks the beta-only base rate closely at every horizon (within a few points, comfortably inside sampling noise for these cell sizes), consistent with the zero-alpha finding in Table II. Bargaining tracks it at $N = 5$ and $N = 20$ but not at $N = 60$, where its 80.0% accuracy sits well above the 64.4% base rate; Section 4.1a takes up that exception. Coordination and Signaling deviate further from the base rate, but from much smaller samples ($n = 7$ -21 per full-period cell); Section 4.3 examines whether that deviation reflects real information or residual small-sample noise. Fig. 3 plots each game type's observed accuracy against the beta-only base rate at every horizon.

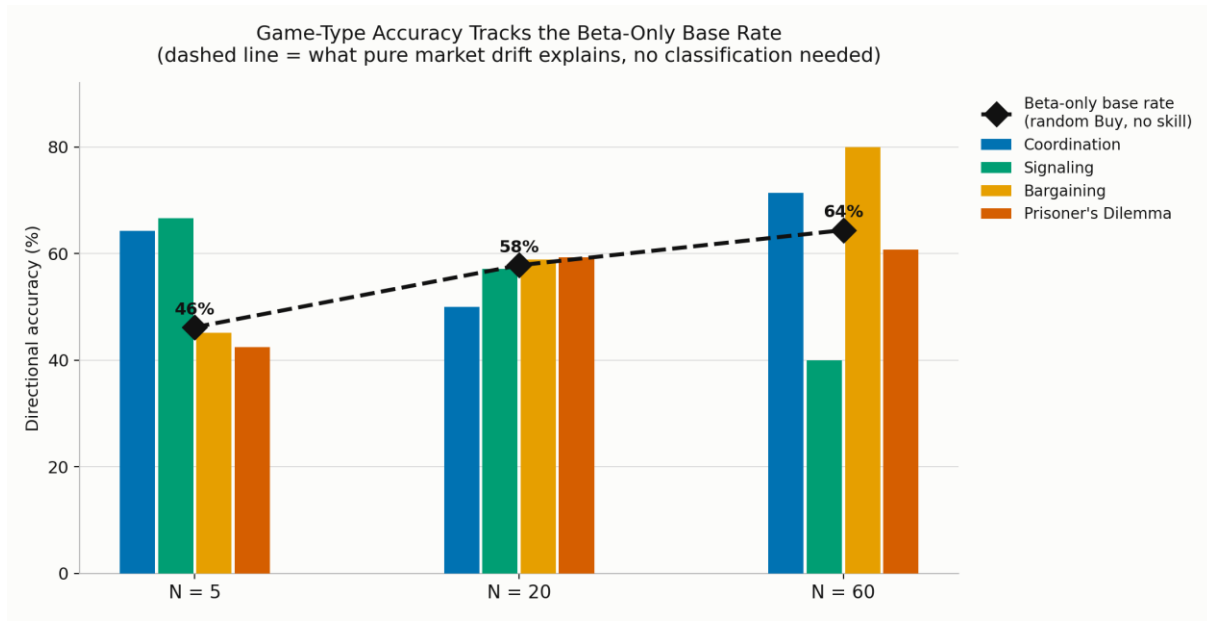


Fig. 3. Observed accuracy by game type (bars) against the beta-only base rate (dashed line with labeled markers) — the accuracy a purely random directional call would achieve from market drift alone, with no classification. Bargaining and Prisoner's Dilemma sit close to the dashed line at every horizon; the gap the original paper attributes to classifier skill is largely already explained before any classification is applied.

4.3 Coordination and Signaling: descriptive only, not a statistical claim

TABLE IV. Stouffer's Pooled Statistic Across Horizons, Full Period (descriptive; see caveat below) Nominal p -values are unadjusted for the overlapping-window serial correlation described in Section 2.1 and are therefore anti-conservative; Coordination's $p = 0.006$ in particular should be read as an upper bound on significance, not as a confirmed effect.

Game Type	Horizons pooled	Stouffer's Z	p
Coordination	3	-2.77	0.006
Signaling	3	+1.83	0.067
Bargaining	3	+0.62	0.532
Prisoner's Dilemma	3	+0.08	0.936

We report Coordination's pooled statistic in Table IV for completeness, but explicitly decline to treat it as a valid hypothesis test: the three full-period cells it pools contain only 14, 14, and 7 observations at $N = 5, 20$, and 60 respectively, and Stouffer's method assumes each per-horizon test statistic is itself a reasonably well-behaved estimate of a Z-score, an assumption that is fragile for a one-sample t-test on cells this small — particularly the $N = 60$ cell ($n = 7$), where a single outlier can move both the point estimate and its implied p-value by an order of magnitude. The sign of the per-horizon alpha estimate is also unstable: Fig. 2 shows it swinging from negative to strongly positive and back to negative across $N = 5, 20$, and 60 for Coordination, with confidence intervals wide enough to overlap both zero and each other at every horizon (the wide confidence interval on Coordination accuracy reported in [1], 95% CI [55.2, 61.4], reflects the same small-sample fragility in the raw, un-adjusted statistic). We therefore report the Z-statistic and its associated p-value in Table IV purely as a descriptive summary of the observed cell means, not as evidence against H_0 for Coordination, and draw no directional conclusion about Coordination-specific alpha from this dataset. Signaling's pooled statistic ($Z = 1.83$, nominal $p = 0.067$, cells of $n = 21, 21, 20$) carries a milder version of the same caveat — worth noting explicitly, its cross-sectional-control counterpart is nominally significant ($Z = 2.25$, $p = 0.024$), which we read as the same small-sample fragility rather than a confirmed effect, for the reasons given above. Only Prisoner's Dilemma and Bargaining have cell sizes ($n \geq 90$ at $N \leq 20$, $n \geq 35$ at $N = 60$) large enough for the pooled Stouffer statistic to be treated as an approximately valid test in the usual sense; Section 4.1 and Table II therefore remain the paper's only load-bearing significance claims.

4.4 Interpretation: an exposure-timing signal, not a direction signal

Taken together, Sections 4.1-4.3 support a specific, falsifiable interpretation of what this classification framework measures: for its largest directionally-actionable category (Prisoner's

Dilemma, $n = 145$ to 250 depending on horizon — Bargaining has more total labels overall, 872 vs. 738 , but fewer of them carry a directional Buy/Sell call, 115 vs. 284 , since GT-2 routes the great majority of Bargaining classifications to Hold), the framework's celebrated low accuracy is a restatement of market beta, not evidence that Prisoner's-Dilemma-classified disagreement points in the *wrong* direction more often than chance. This is an economically important distinction. A classification with genuine negative alpha would be a contrarian signal — profitable to fade. A classification with zero alpha, as we find here, is not a trading signal in either direction; it is a *volatility/uncertainty marker*, informative about when the market's disagreement structure is uninformative, without being informative about which way to bet. The half-Kelly abstention rule in [1] happens to be the mathematically correct response to *both* interpretations (negative or zero alpha both imply a non-positive optimal Kelly stake, so abstention is optimal either way), which is why the practical trading rule survives even though the causal story behind the reported accuracy gap does not.

4.5 Robustness: alternative beta estimators

The period-average drift $\bar{R}_N(\pi)$ in Section 4.1 has two properties a careful reader should treat as caveats rather than settled fine print: it is a single, coarse aggregate market factor rather than a sector- or style-adjusted exposure, and it is estimated in-sample — using every date in period π , including the very dates being tested — which biases the null in the framework's favor by construction. We report two alternative estimators that each address one of these concerns, and are transparent about where they agree with Section 4.1 and where they do not.

The first, a cross-sectional control, replaces the single period-wide constant $\bar{R}_N(\pi)$ with a date-specific factor: for each signal, we subtract the mean forward return of every ticker in the universe on that same calendar date, which nets out day-specific market-wide moves (the closest feasible proxy to a single-factor market model [6] available without external sector/style factor data) and is a materially stronger control than the coarse period average, since it holds the comparison group fixed at the exact date of each signal rather than averaging over the whole evaluation window. This per-date mean is computed over the variable active universe — every ticker carrying a scoreable forward return on that date — rather than over a fixed baseline restricted to the eleven continuously covered tickers, so the size of the comparison group rises from eleven to sixteen when the late-entering tickers join partway through the sample; Section 6

notes the resulting compositional caveat. Table V reports this cross-sectional alpha test for Prisoner's Dilemma alongside the main-text result.

TABLE V. Prisoner's Dilemma Alpha: Period-Average Drift (main text) vs. Cross-Sectional (per-date) Control. Columns 2-4 report the period-average drift estimator of Section 4.1; columns 5-7 the per-date cross-sectional control. As in Table IV, p-values are unadjusted for overlapping-window serial correlation.

N	n	α (period)	t	p	n	α (x-sect.)	t	p
5	250	-0.14%	-0.26	0.798	250	-0.05%	-0.11	0.911
20	221	+0.91%	+0.69	0.488	221	+0.78%	+0.71	0.477
60	145	-1.16%	-0.30	0.765	145	+1.85%	+0.55	0.580

The cross-sectional control leaves the paper's central claim intact: Prisoner's-Dilemma alpha remains statistically indistinguishable from zero at every horizon (all $p \geq 0.48$, pooled Stouffer's $Z = 0.67$, $p = 0.51$), and the point estimates shrink toward zero in magnitude relative to the period-average version at $N = 5$ and $N = 20$, if anything strengthening the null result once day-specific market moves are controlled for.

The second alternative, a trailing (expanding-window) drift estimate, directly addresses the in-sample-optimism concern: for a signal issued on date d , drift is computed using only observations with signal date strictly before d , so no information from d itself or any later date enters the beta estimate used to evaluate that signal. At $N = 5$ and $N = 20$ this produces results consistent with Table II (all $p \geq 0.30$). At $N = 60$, however, it produces a materially different and more decisive result: alpha becomes significantly negative ($n = 145$, $\bar{\alpha} = -13.15\%$, $t = -3.60$, $p = 0.0004$; pooled across horizons, $Z = -2.89$, $p = 0.0039$). We traced this divergence to its source rather than reporting it uncritically. Every $N = 60$ -scoreable Prisoner's-Dilemma signal in this dataset is dated between 2026-03-06 and 2026-06-05 ($N = 60$ forward returns cannot be computed for later dates within the sample's end date), a window falling almost entirely inside the original bull-market period. Within that window, the trailing drift is not merely large but strongly time-varying — it exceeds 42% for signals dated in March (reflecting the earliest, steepest part of the rally accumulated in the trailing window) and decays to roughly 25% by June, which happens to coincide closely with the period-average constant of 23.8% used in Table II. Because Prisoner's-Dilemma signals are not date-uniform across this window, evaluating each

one against its own (often much higher) trailing drift, rather than the single period-average constant, mechanically produces a more negative residual for the earlier-dated majority of signals than Table II's coarser estimator does. This is not a look-ahead artifact — the trailing estimator strictly uses only prior information at every date — but it is a genuine estimator-sensitivity finding: at the longest horizon and within a period of unusually strong, front-loaded market drift, the two beta estimators disagree on whether Prisoner's-Dilemma alpha is zero or modestly negative, and we cannot fully adjudicate between them with the sample currently available. Fig. 4 shows all three estimators side by side at each horizon. We view the disagreement itself, not either estimator's point estimate in isolation, as the correct summary of what this dataset supports at $N = 60$, and flag it as the paper's most important open robustness question.

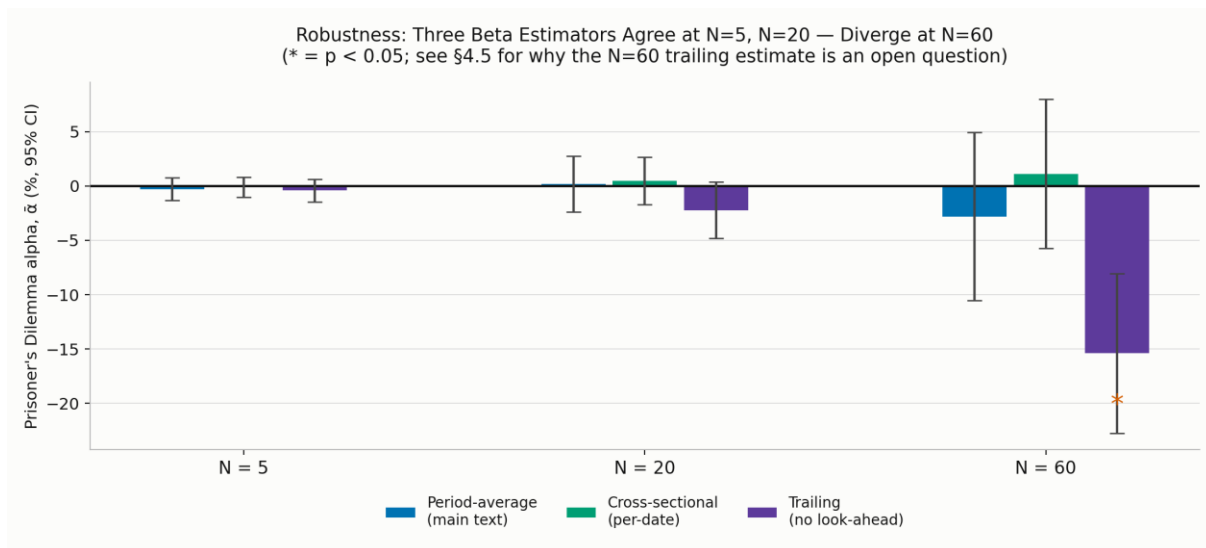


Fig. 4. Prisoner's-Dilemma alpha under three independent beta estimators, at each horizon. All three agree (95% CI spans zero) at $N = 5$ and $N = 20$. At $N = 60$, the trailing (no-look-ahead) estimator alone turns significantly negative — traced in the text to the narrow, front-loaded-drift bull-market window that is the only calendar range long enough to be $N = 60$ -scoreable, and reported as an open question rather than a settled result.

5. A Decision-Theoretic Argument for Agentic AI as an Exposure Layer, Not an Alpha Layer

5.1 Formal setup

Consider a quantitative investment process with two candidate architectures for incorporating an agentic AI / multi-LLM system. In the *direct-alpha* architecture, the AI system's classified output g is used to select a directional bet whose expected value is assumed to depend on g ($E[R | g] \neq E[R]$). Fig. 5 contrasts the two architectures. In the *exposure-control* architecture, a separate, independently-sourced alpha signal ψ (a human portfolio manager's view, a traditional quant factor, or a different, independently-validated model) generates the direction and magnitude of a candidate trade, and the agentic AI system's classified output g is used only to scale the position size multiplicatively via a risk-budget function $\kappa(g) \in [0,1]$, i.e. the realized position is $\kappa(g) \cdot f(\psi)$ for some sizing function f .

Our empirical result in Section 4 is precisely a test of the direct-alpha architecture's core assumption for the Prisoner's-Dilemma category: we fail to reject $E[\alpha | g=PD] = 0$, i.e. $E[R | g=PD] = \bar{R}_N(\pi)$, so g carries no information about the *sign or magnitude* of R beyond the market's own average drift. Under the direct-alpha architecture, deploying capital on the strength of a PD classification alone therefore has an expected value equal to a passive market-drift bet — no worse, but also no better, than doing nothing directional and simply holding the index, net of any transaction costs the active bet incurs. Any transaction cost $c > 0$ makes the direct-alpha use of a PD classification *strictly dominated* by inaction:

$$E[R_{k,d,N} - c | g = PD] = \bar{R}_N(\pi) - c < \bar{R}_N(\pi) = E[\text{passive exposure}] \quad (6)$$

The platform's own observed behavior coincides with (6): the automated classifier acts on none of its 738 Prisoner's-Dilemma classifications. This coincidence is not itself evidence for (6), however. GT-2 reaches the same outcome structurally rather than through any cost calculation — the classifier returns no tradable direction for the 1-1-1 split and 2-1-Hold patterns that produce most Prisoner's-Dilemma labels, and the minority that do carry a direction are screened out by the confidence gate before any cost comparison arises. Equation (6) therefore explains why abstention is the correct policy, not why this particular implementation abstains.

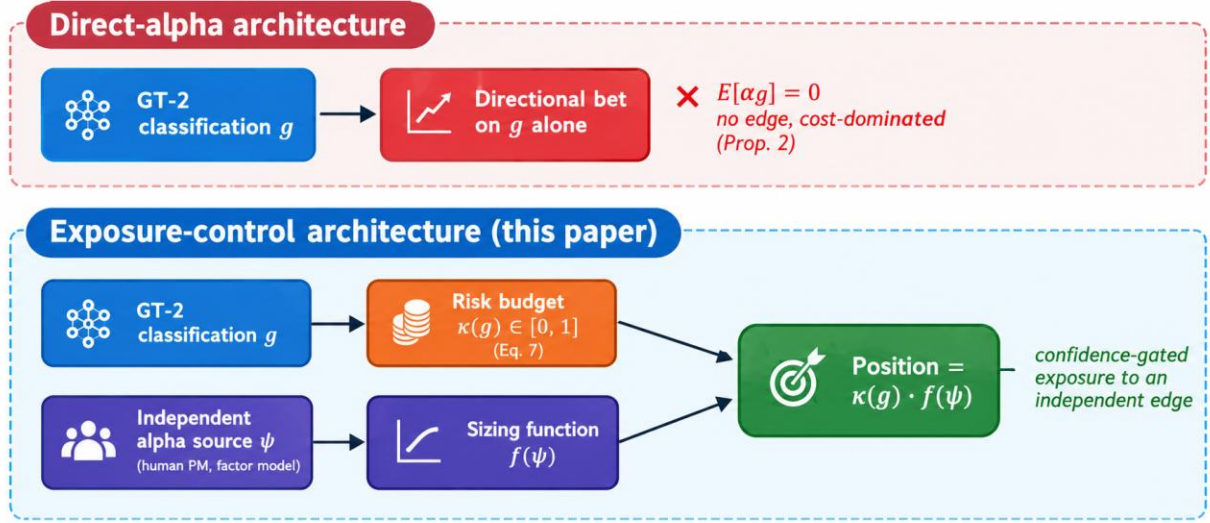


Fig. 5. The direct-alpha architecture (top) treats a classification g as a directional bet in its own right — dominated by inaction under Proposition 2 whenever $E[\alpha|g] \leq 0$. The exposure-control architecture (bottom, proposed here) instead uses g only to gate the size of a position built from an independently-sourced alpha signal ψ , via the risk-budget rule $\kappa(g)$ of Eq. (7).

5.2 Why exposure-control survives even when direct-alpha does not

The exposure-control architecture does not require $E[\alpha | g] \neq 0$ to add value; it requires only that the *variance* or *tail behavior* of realized outcomes differ systematically by g , or that g predicts a state of the world in which an independent alpha signal ψ 's own accuracy is degraded. Neither of these is ruled out by our zero-mean-alpha finding. A classification could correctly flag “this is a moment where LLM disagreement contains no information,” which is a valid and useful signal for *down-weighting* an LLM-derived ψ specifically (even while leaving a separately-sourced, human or classical-factor ψ unaffected), without that classification needing to carry any directional content of its own. Formally, if ψ is itself partly derived from the same disagreeing LLM ensemble, and g proxies for “this ensemble's current output is uninformative,” then $\kappa(g)$ should shrink toward zero exactly when $g = \text{Prisoner's Dilemma}$, which is a meta-level, ensemble-quality-control use of the classification that is fully consistent with our finding of zero *direction-level* alpha.

5.3 A risk-budgeting rule and its implication for agentic AI system design

This suggests a general risk-budgeting rule for agentic AI systems used as an input to a quantitative process, distinct from (and, we argue, more defensible than) using the same system's output as a direct trading signal:

$$\text{Position} = \kappa(g) \cdot f(\psi), \quad \kappa(g) = \max\left(0, \frac{\hat{p}(g) - \hat{p}_0}{1 - \hat{p}_0}\right) \quad (7)$$

where $\hat{p}(g)$ is the classifier's own estimated confidence for game type g , $\hat{p}_0$ is the beta-only base rate from (5) (the confidence a coin-flip classifier would achieve for free), and $f(\psi)$ is an independently-sourced directional and sizing signal. Equation (7) has two properties worth emphasizing for practitioners designing agentic-AI-augmented investment systems. First, it sets $\kappa(g) = 0$ whenever the classifier's own confidence is at or below the beta-only floor $\hat{p}_0$ — exactly the condition our empirical results show holds for Prisoner's Dilemma — so it automatically reduces to abstention in the case this paper studies, without requiring the system designer to know in advance which categories will turn out to be pure beta. Second, and more importantly for the broader field, it structurally separates the question “should the AI system's classification be trusted at all in this instance” (a calibration and confidence-floor question, answerable from backtested $\hat{p}(g)$ vs. $\hat{p}_0$ comparisons of exactly the kind in Table III) from the question “what trade should be made” (an alpha question, which this paper's results suggest agentic AI classification, at least as currently implemented, does not yet answer).

5.4 Forward-looking implication for the quantitative investment industry

The practical recommendation this analysis supports is architectural rather than a claim about any specific model's future capability: quantitative desks integrating LLM and multi-agent systems should, on current evidence, treat these systems' primary contribution as a conditioning variable on exposure — a real-time classifier of when the market's information environment is coherent enough for any signal (human or machine) to be acted upon with conviction — rather than assume that a sophisticated-sounding classification (game type, sentiment regime, disagreement structure) is itself an alpha source simply because it correlates with headline

accuracy statistics that have not been beta-adjusted. This has a direct, falsifiable implication for how such systems should be validated before deployment: any reported accuracy or win-rate statistic for an LLM- or agent-classified trading signal should be accompanied by the decomposition in Equations (1)-(2), against that signal's own realized universe drift over the same evaluation window, not against a fixed 50% or fixed-threshold benchmark that ignores the regime the backtest happened to run in. A signal that clears this bar — nonzero α , not merely nonzero raw accuracy — has a stronger claim to being deployed as a direct alpha source; one that does not, as we show is the case for Prisoner's-Dilemma classification in this dataset, should be deployed as an exposure-control input under (7) instead, where it can still add measurable value (via $c > 0$ cost avoidance, as in (6), and via ensemble-confidence shrinkage, as in Section 5.2) without over-claiming a directional edge the data does not support. As agentic AI systems proliferate across quantitative desks, we expect this beta/alpha accounting discipline — applied to the classification layer itself, not only to the final portfolio — to become a standard due-diligence step, in the same way factor-based performance attribution became standard practice for equity managers after early single-number performance claims [9] were shown to be substantially explained by market and style exposures rather than manager skill.

6. Limitations

Section 4.5 already addresses, rather than merely flags, the two most obvious objections to the main-text decomposition: a cross-sectional per-date control confirms the zero-alpha result under a stronger, more local market-neutrality assumption, and a trailing (no look-ahead) drift estimator confirms it at $N = 5$ and $N = 20$ but surfaces a genuine, unresolved estimator sensitivity at $N = 60$ that we report rather than paper over. We restate the residual limitations here. The Coordination and Signaling categories remain too small in this automated-classifier sample ($n = 7$ to 21 per full-period cell, and as few as 2 in the in/out-of-sample split cells) to draw a confident conclusion in either direction under any of the three estimators used in this paper; our null result is strongest and best-supported for Prisoner's Dilemma, which has adequate sample size at every horizon under every estimator. Bargaining, whose $N = 5$ and $N = 20$ cells are also adequately powered and also null, departs from the null at $N = 60$ (Section 4.1a); that departure rests on 35 observations largely confined to the bull-market window, does not survive the trailing-drift estimator, and we do not treat it as established. Even the cross-sectional control in Section 4.5 is

a single-factor (equal-weighted universe mean, computed daily rather than over the full period) model, and the universe it averages over is not constant: five of the sixteen tickers enter the watch list partway through the sample (Section 3), so the per-date comparison group is narrower in the earlier part of the window than in the later part; a genuine multi-factor decomposition (sector, momentum, size, following e.g. Fama-French-style factor construction [7]) could still reveal alpha that an aggregate market factor obscures, though this would require external factor data and a substantially larger cross-sectional universe than the sixteen primarily-technology tickers available here to estimate with adequate power. A further limitation is common to all three estimators: 242 of the 1,836 classified observations carry a game-type label but no usable close price in the platform's log, and are dropped from every return-based test; we have no reason to believe this missingness is correlated with game type, but we cannot demonstrate that it is not. The $N = 60$ divergence documented in Section 4.5 means we cannot rule out a modest negative Prisoner's-Dilemma alpha specifically at the longest horizon and specifically within the bull-market sub-window where the great majority of current $N = 60$ observations happen to fall; resolving this will require additional trading days beyond the current sample's end date, which the trailing estimator's out-of-sample discipline currently makes unavailable. Finally, this analysis, like [1] and [4], is conducted entirely within a primarily technology-sector (fifteen of the sixteen tickers; MRK is a pharmaceutical), primarily long-only universe over a period containing no sustained bear market, so the beta term itself — under any of the three estimators used here — may not generalize to regimes where market-wide drift is negative or near zero for extended periods.

7. Conclusion

We show that the reported low accuracy of Prisoner's-Dilemma-classified multi-LLM disagreement signals in a game-theoretic classification framework is, in an automated replication over 142 platform report-dates, statistically indistinguishable from what a beta-only, uninformed directional bet would achieve once the market's own average forward return is subtracted (all $|t| < 0.74$, joint Stouffer's $Z = 0.08$, $p = 0.94$), and this null result survives a stronger, per-date cross-sectional control at every horizon and an out-of-sample trailing-drift estimator at $N = 5$ and $N = 20$, but not the trailing estimator at $N = 60$, where it alone turns significantly negative while the other two estimators remain null and we report that disagreement as an open question rather than

resolving it in either direction. Neither reading undermines the practical value of the abstention rule the original framework derives — abstaining is the correct response whether the true alpha is negative or merely zero, by Proposition 2 — but the balance of evidence across horizons and estimators shows that the accuracy gap used to motivate that rule should be attributed, at minimum for the shorter and more heavily evidenced horizons, to a common risk-management logic (avoid acting on uninformative signals net of cost) rather than to genuine directional skill encoded in the classification. We formalize this distinction as a decomposition with two proven propositions, propose a risk-budgeting rule (Eq. 7) that captures the value a classification framework can still deliver once its alpha claim is set aside, and argue this points toward a specific, testable architecture for the next generation of agentic-AI-augmented quantitative investing: LLM and multi-agent systems as a real-time exposure-control and signal-confidence layer, sitting alongside — rather than replacing — independently-validated sources of directional alpha.

References

- [1] Q. Zhang, “Game-Theoretic Scaffolding in Financial Analysis: Why Human Direction of LLMs Outperforms Autonomous Models,” SSRN working paper no. 6705178, 2026. doi:10.2139/ssrn.6705178. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6705178
- [2] A. Lopez-Lira and Y. Tang, “Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models,” SSRN working paper no. 4412788, 2023.
- [3] Y. Chen et al., “StockBench: Can LLM Agents Trade Stocks Profitably In Real-world Markets?” arXiv:2510.02209, 2025.
- [4] Q. Zhang, “Benchmarking Multi-LLM Investment Signals: Accuracy, Consensus, and the Augmented Analyst Edge in Daily Stock Recommendations,” SSRN working paper no. 6515058, 2026. doi:10.2139/ssrn.6515058. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6515058
- [5] E. O. Thorp, “The Kelly criterion in blackjack, sports betting, and the stock market,” in *Handbook of Asset and Liability Management*, Vol. 1, S. A. Zenios and W. T. Ziemba, Eds. Amsterdam: Elsevier, 2006, ch. 9, pp. 385-428.

- [6] W. F. Sharpe, "Asset allocation: Management style and performance measurement," J. Portfolio Management, vol. 18, no. 2, pp. 7-19, 1992.
- [7] E. F. Fama and K. R. French, "Common risk factors in the returns on stocks and bonds," J. Financial Economics, vol. 33, no. 1, pp. 3-56, 1993.
- [8] S. A. Stouffer, E. A. Suchman, L. C. DeVinney, S. A. Star, and R. M. Williams Jr., The American Soldier: Adjustment During Army Life (Studies in Social Psychology in World War II, Vol. 1), Princeton University Press, 1949.
- [9] M. C. Jensen, "The Performance of Mutual Funds in the Period 1945-1964," J. Finance, vol. 23, no. 2, pp. 389-416, 1968. doi:10.1111/j.1540-6261.1968.tb00815.x
- [10] W. K. Newey and K. D. West, "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," Econometrica, vol. 55, no. 3, pp. 703-708, 1987.
- [11] K. Liu and L. Chen, "OOM-RL: Out-of-Money Reinforcement Learning Market-Driven Alignment for LLM-Based Multi-Agent Systems," arXiv:2604.11477, 2026.
- [12] T. Zhu, W. Zhao, R. Sun, B. Luan, J. Lu, S. Wang, J. Li, D. Jiang, Y. He, and Z. Bai, "From Knowing to Doing: A Memory-Controlled Benchmark for LLM Trading Agents on Stock Markets," arXiv:2605.28359, 2026.
- [13] H. Zhang, Z. Wang, X. Sun, and J. Wang, "Many Men, Many Minds: A Multi-Agent LLM Approach to Disagreement and Global Asset Pricing," SSRN working paper no. 5962814, 2025. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5962814

Declarations

Availability of data and materials: The datasets analyzed during this study are available from the corresponding author on reasonable request; aggregated summary statistics sufficient to replicate all reported tables are included in the manuscript. Every value in Tables I-V is regenerated from the platform's own on-disk signal and price record by a single script (scripts/beta_decomposition.py in the platform repository), which emits the complete result set as machine-readable JSON; no reported value is entered by hand. Competing interests: The author declares no competing interests. Funding: This research received no specific grant from

any funding agency. Author contributions: QZ is the sole author, responsible for conceptualization, data collection, analysis, and manuscript preparation.