Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Explainable Multi-Agent Financial Advisory: Design and Comparative Evaluation of Collaborative Reasoning Models for Investment Recommendation

Aman Goyal

atongoyal@gmail.com

Chandigarh University

Inderdeep Kaur

Chandigarh University

Varun Gupta

Chandigarh University

Sahil Rauniyar

Chandigarh University

Research Article

Keywords: financial AI agents, multi-agent systems, gen-erative AI, large language models, explainable AI, investment recommendation, machine learning, financial analytics, collabo-rative reasoning

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11780690/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Explainable Multi-Agent Financial Advisory: Design and Comparative Evaluation of Collaborative Reasoning Models for Investment Recommendation

Aman
Department of CSE
Chandigarh University
Mohali, Punjab, India (140413)
atongoyal@gmail.com
UID: 23BCS13452

Inderdeep Kaur
Department of CSE
Chandigarh University
Mohali, Punjab, India (140413)
inderdeep.sgtb@gmail.com

Varun Gupta
Department of CSE
Chandigarh University
Mohali, Punjab, India (140413)
23bcs11907@cuchd.in
UID: 23BCS11907

Sahil Rauniyar
Department of CSE
Chandigarh University
Mohali, Punjab, India (140413)
23bcs13633@cuchd.in
UID: 23BCS13633

***Abstract*—Generative-AI agents are increasingly proposed for investment analysis, yet two questions remain under-examined: *how should multiple specialist agents combine their opinions*, and *what does the combination cost in explainability*? This paper designs a financial multi-agent advisory framework in which four tool-grounded analyst agents—momentum, mean-reversion, risk, and macro-regime—emit bounded conviction scores, confidences, and structured evidence. On top of a single agent interface we formalize six collaborative reasoning models as *coalition-evaluable operators*: a best single agent, uniform voting, confidence-weighted fusion, iterative debate (DeGroot-style consensus formation), a hierarchical coordinator with risk-veto authority, and a learned ridge-regression fusion. Coalition evaluability makes exact Shapley attributions over agents computable in closed form (2^4 coalitions), yielding a unified explanation object—per-agent attributions, evidence rationale, disagreement-based uncertainty, and a natural-language brief—for every recommendation, regardless of the reasoning model. In a 4.6-year out-of-sample evaluation (2022–2026) on ten large-cap U.S. equities with transaction costs, confidence-weighted fusion and debate achieve Sharpe ratios of 1.03 and 1.02 versus 0.78 for a passive S&P 500 benchmark and 0.74 for the best single agent, while the hierarchical veto committee is over-conservative (0.52); block bootstrap shows this protocol ordering is robust (stable in 99.7% of resamples) even where absolute market outperformance is not statistically separable at this horizon. Explanation faithfulness degrades monotonically with reasoning interactivity (from 1.00 for additive committees to 0.78 for debate) even though exact Shapley attributions remain stable under input perturbation ($\geq$0.995 cosine similarity), quantifying an explainability cost of collaboration that simple deletion-based explanations would misreport. All data, code, figures, and tables are reproducible from a released Python artifact.**

***Index Terms*—financial AI agents, multi-agent systems, generative AI, large language models, explainable AI, investment recommendation, machine learning, financial analytics, collaborative reasoning**

I. Introduction

Large language models (LLMs) and the surrounding generative-AI tooling have moved rapidly into financial analysis: domain-specific foundation models [1], [2], [3], deep-reinforcement-learning trading stacks [4], memory-augmented trading agents [5], [6], and, most recently, *multi-agent* systems that mirror the role structure of an investment firm—analysts, researchers, risk managers, and portfolio managers debating a trade [7]. The multi-agent turn is well motivated: reasoning quality of LLMs improves with structured deliberation [8], [9], [10], and multi-agent debate has been shown to improve factuality on reasoning benchmarks [11], [12]. Orchestration frameworks such as AutoGen [13] and MetaGPT [14] make such systems easy to build.

Ease of construction, however, has outpaced evaluation discipline. Published financial-agent systems typically bundle one specific collaboration protocol with one specific prompt stack and one specific backtest, making it impossible to answer a basic design question: *holding the analysts fixed, which collaborative reasoning model should combine their opinions?* The question matters twice over. First, aggregation is a free design axis with material performance consequences. Second, financial advice is a high-stakes, increasingly regulated application: supervisory guidance on model risk [27] and the risk-tiered obligations of the EU AI Act [28] both demand that automated recommendations be explainable, auditable, and accompanied by calibrated uncertainty. A collaboration protocol that improves returns but destroys the ability to attribute a recommendation to evidence may be unusable in practice [19].

This paper studies the collaboration axis directly and in-

struments it for explainability from the start. We make four contributions:

1) **Architecture.** We design a financial multi-agent advisory framework (Fig. 1) in which specialist analyst agents consume engineered market features through tool calls and emit a typed opinion—a bounded conviction score, a confidence, and a structured evidence rationale. The committee layer is strictly interface-separated from the agents, so deterministic policy heads and LLM-instantiated agents are interchangeable.
2) **Unified formalization of collaborative reasoning.** We express six committee designs—single agent, uniform vote, confidence-weighted fusion, iterative debate, hierarchical coordination with veto, and learned fusion—as *coalition-evaluable operators* $F(S)$ defined on every subset S of agents (Sec. IV).
3) **Explainability framework.** Coalition evaluability makes *exact* Shapley attributions over agents computable (2^4 evaluations), giving every recommendation the same explanation object independent of the reasoning model. We evaluate explanations on faithfulness, stability, and completeness, and evaluate inter-agent disagreement as an uncertainty signal (Sec. V).
4) **Evidence.** On real daily data for ten large-cap U.S. equities with a strict calibration/test split (2017–2021 / 2022–2026) we compare all six reasoning models on portfolio performance, decision quality, and explanation quality (Sec. VII), and release the full artifact.

Three findings stand out. (i) *Collaboration helps, but the protocol matters*: confidence-weighted fusion and debate deliver the best risk-adjusted performance (Sharpe 1.03/1.02 vs. 0.78 for the market and 0.74 for the best single agent), while a hierarchical veto committee forfeits nearly half of that Sharpe through over-conservatism—an ordering that holds in 99.7% of block-bootstrap resamples. (ii) *Interactivity taxes explainability*: leave-one-out explanations remain perfectly faithful for additive committees but degrade to 0.78 correlation with exact Shapley values under debate, a bias our closed-form attribution corrects at negligible cost. (iii) *The interpretability–performance trade-off is not monotone*: the most transparent high-performing design (confidence-weighted fusion) matches the most interactive one, supporting the view that high-stakes domains rarely need to buy performance with opacity [19].

II. Related Work

A. Generative AI and ML in finance

Domain-adapted language models such as BloombergGPT [1], FinGPT [2], and FinBERT [3] demonstrated that financial text understanding benefits from domain data. FinRL [4] established open deep-RL pipelines for trading, and Gu, Kelly, and Xiu [21] showed that machine learning materially improves empirical asset pricing, while cautioning that predictability is weak and regime dependent—consistent with the small information coefficients we observe. Agentic systems extend these models with memory and tools: FinMem [5] adds layered memory to a trading LLM, FinAgent [6] adds multimodal perception, and TradingAgents [7] organizes specialized LLM agents (fundamental, sentiment, technical, risk) into a firm-like workflow with debate between bull and bear researchers. Our work is complementary: rather than proposing one more integrated system, we hold the analyst layer fixed and isolate the *collaboration protocol* as the experimental variable, with explainability as a first-class metric.

B. Collaborative reasoning among agents

Structured reasoning improves LLM output quality [8], [9], [10]; multi-agent debate improves factuality and reasoning [11] though it can also amplify conformity [12]. General orchestration frameworks [13], [14] popularized conversational committees. The opinion dynamics we use to formalize debate descend from DeGroot's classical consensus model [15], which gives the protocol a transparent parametric form (openness, rounds) rather than an opaque prompt. Ensemble aggregation of forecasts is itself classical [23], [24], and momentum and reversal effects motivating our analysts are extensively documented [22].

C. Explainable AI for financial decisions

Post-hoc attribution methods such as LIME [18] and SHAP [17] approximate Shapley values [16] over input features because exact computation is exponential in the number of players. Our setting inverts the usual difficulty: at the *agent* level the committee has only four players, so exact Shapley attribution is cheap, and the approximation error that plagues feature-level XAI disappears. Surveys of XAI stress faithfulness and stability as necessary conditions for trustworthy explanations [20], and Rudin [19] argues for intrinsically interpretable models in high-stakes use; regulatory guidance [27], [28] makes these requirements concrete for finance. Finally, our evaluation design follows the warnings of Bailey *et al.* [25] on backtest overfitting: a single chronological split, a small hyperparameter grid frozen in-sample, and transaction costs throughout.

III. System Design

A. Agent abstraction

An analyst agent i is a function $a_i(\mathbf{x}_t) = (s_i, c_i, \mathcal{R}_i)$ mapping the information set $\mathbf{x}_t$ available at the close of decision date t to a conviction score $s_i \in [-1, 1]$ (bullish positive), a confidence $c_i \in (0, 1]$, and a structured rationale $\mathcal{R}_i$: the named evidence values the score was computed from. The typed interface is deliberately the same one an LLM analyst produces under function-calling with tool access (compute-indicator tools returning the same features), which is how generative-AI instantiations plug in without touching the committee or explanation layers. For the experiments reported here the agents are deterministic, tool-grounded policy heads, so that every number in this paper is exactly reproducible and no result depends on prompt phrasing or sampling temperature; the collaboration comparison is thereby isolated from LLM output variance.

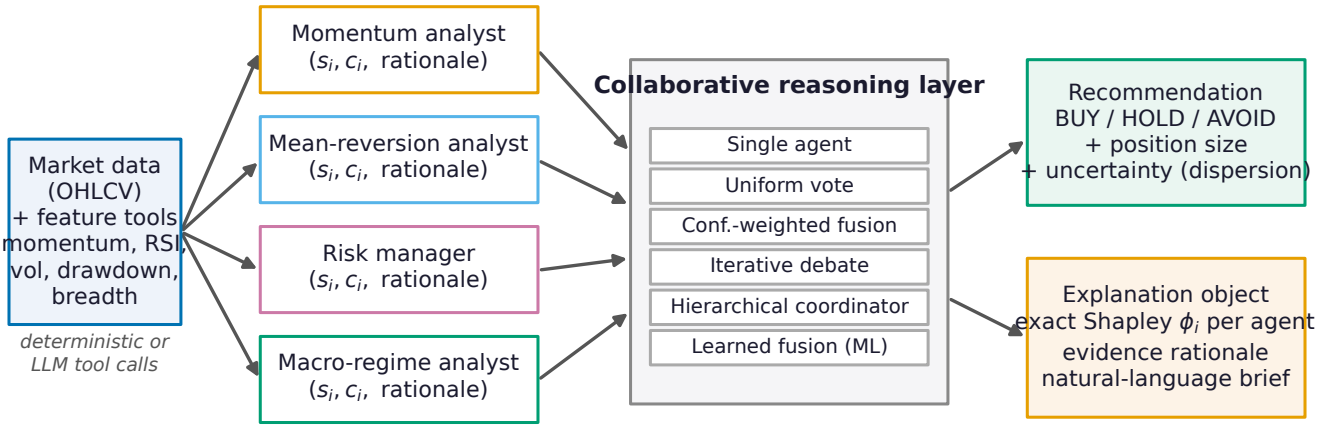


Fig. 1. The advisory framework. Analyst agents consume engineered features through tool calls and emit typed opinions $(s_i, c_i, \mathcal{R}_i)$: a conviction score $s_i \in [-1, 1]$, a confidence $c_i \in (0, 1]$, and an evidence rationale $\mathcal{R}_i$. A collaborative reasoning layer aggregates opinions into a recommendation; because every reasoning model is defined on arbitrary agent coalitions, the explanation object (exact Shapley attributions, rationale, uncertainty, natural-language brief) is computed identically for all six models. Agents are interface-interchangeable between deterministic policy heads (used in the reported experiments for exact reproducibility) and LLM-instantiated analysts.

B. Analyst agents

Four agents span complementary investment doctrines. All features use information up to the decision close only.

1) Momentum analyst: evidence: 12–1 month momentum $m = P_{t-21}/P_{t-252} - 1$, the 50/200-day moving-average trend gap, and the above/below 200-day state. Score: a weighted blend of $\tanh$-squashed sub-signals $(0.50/0.35/0.15)$; confidence rises with signal magnitude and sub-signal sign agreement [22].

2) Mean-reversion analyst: evidence: the 20-day Bollinger z-score and RSI(14). Score: contrarian $\tanh$ blend $(0.60/0.40)$; confidence rises with extremity of the deviation.

3) Risk analyst: evidence: the percentile of 21-day realized volatility within a rolling 1.5-year window, and the drawdown from the 1-year high. Score: negative when volatility is elevated or drawdowns deep (risk hostile), positive in benign conditions.

4) Macro-regime analyst: evidence: whether the S&P 500 proxy (SPY) trades above its 200-day average, its 100-day return, and market breadth (the fraction of universe names above their own 200-day average). This agent is market-level: it emits the same opinion for every ticker on a given date and contributes no cross-sectional ranking signal by design; its inclusion reflects the regime-dependence of strategy efficacy [29].

IV. Collaborative Reasoning Models

Let $N = \{1, \ldots, 4\}$ be the agent set with opinions $\{(s_i, c_i)\}$. A *coalition-evaluable operator* is a map $F(S)$ defined for every $S \subseteq N$ with $F(\emptyset) = 0$: the recommendation the committee would issue if only the agents in S were present (weights renormalized, dynamics run on S alone). All six models below satisfy this property, which Sec. V exploits.

1) Single (best single agent): $F(S) = s_{i^*}$ if $i^* \in S$ else 0, where i^* is the agent with the highest calibration-window information coefficient (momentum, in our data).

2) Vote (uniform ensemble): $F(S) = \frac{1}{|S|} \sum_{i \in S} s_i$.

3) Cwf (confidence-weighted fusion):

$$F(S) = \sum_{i \in S} w_i s_i, \qquad w_i = \frac{e^{\beta c_i}}{\sum_{j \in S} e^{\beta c_j}}, \tag{1}$$

with temperature β calibrated in-sample.

4) Debate (iterative consensus): Agents revise positions over R rounds of DeGroot-style dynamics [15]. With $\bar{s}^{(r)} = \sum_j w_j s_j^{(r)}$ the confidence-weighted group view and $\alpha_i = \alpha_0(1 - c_i)$ the openness of agent i,

$$s_i^{(r+1)} = (1 - \alpha_i)\, s_i^{(r)} + \alpha_i\, \bar{s}^{(r)}, \tag{2}$$

so confident agents concede little and diffident agents concede more—the quantitative skeleton of LLM debate protocols [11], [7]. The recommendation is $\bar{s}^{(R)}$; the residual dispersion $\sigma_R = \mathrm{std}_i\, s_i^{(R)}$ is retained as epistemic uncertainty.

5) Coord (hierarchical coordinator): Mirroring an investment-committee hierarchy, the proposal agents (momentum, mean-reversion) are fused by (1); the oversight agents then scale the position multiplicatively,

$$F = \Big(\textstyle\sum_{i \in \{\mathrm{M,V}\}} w_i s_i \Big) \cdot g_{\text{risk}} \cdot g_{\text{reg}}, \tag{3}$$

where $g_{\text{risk}} \in [0, 1]$ is an affine map of the risk score with a hard veto (g=0) under extreme volatility and drawdown, and $g_{\text{reg}} \in [0.3, 1]$ scales exposure to the market regime (gates apply to long proposals; absent oversight agents leave gates neutral).

6) LEARN *(learned fusion):* A machine-learning baseline: ridge regression fitted *on the calibration window only*, mapping per-agent terms to the 21-day forward return and squashed for comparability,

$$F(S) = \tanh\Big(\tfrac{1}{2\sigma}\big(b + \textstyle\sum_{i\in S}(u_i s_i + v_i c_i s_i)\big)\Big). \quad (4)$$

Because the linear index is additive per agent, coalition semantics are exact. Notably, the fitted weights turn the momentum agent's signal *contrarian* ($u_{\mathrm{M}} < 0$): in the calibration years, extreme high-confidence bearish momentum in beaten-down mega-caps preceded rebounds.

V. EXPLAINABILITY FRAMEWORK

Regulated financial advice needs three things an accuracy metric does not give: *who* drove a recommendation, *how sure* the system is, and *whether the stated reasons are the real ones* [20], [27]. Every recommendation in our framework therefore ships an *explanation object* with four parts.

1) Exact Shapley attribution: Treating F as a characteristic function over agents, the attribution of agent i is the Shapley value [16]

$$\phi_i = \sum_{S\subseteq N\setminus\{i\}} \frac{|S|!\,(|N|-|S|-1)!}{|N|!}\big[F(S\cup\{i\}) - F(S)\big]. \quad (5)$$

Feature-level XAI must approximate (5) [17], [18]; at agent granularity $|N| = 4$, the 2^4=16 coalition evaluations cost microseconds, and the efficiency axiom guarantees $\sum_i \phi_i = F(N)$ exactly—the explanation provably accounts for the whole recommendation, for *every* reasoning model including the non-additive debate and gated coordinator.

2) Evidence rationale: The per-agent $\mathcal{R}_i$ values (momentum, trend gap, RSI, volatility percentile, breadth, ...) ground each attribution in named, auditable market facts.

3) Disagreement as uncertainty: The pre-aggregation dispersion $\sigma_0 = \mathrm{std}_i(s_i)$ (and, for DEBATE, the residual σ_R) is reported as an uncertainty estimate, which Sec. VII-D evaluates against realized accuracy.

4) Natural-language brief: A deterministic template renders the score, attributions, rationale, and dissent into a short analyst brief (Fig. 6); with LLM-instantiated agents the same fields prompt a generative summary.

Explanation quality metrics. We evaluate explanations on: **faithfulness**—the mean per-decision Pearson correlation between ϕ and leave-one-out marginals $\Delta_i = F(N) - F(N\setminus\{i\})$, measuring how well cheap deletion tests recover the exact attribution; **stability**—the mean cosine similarity between ϕ computed on original and perturbed inputs (i.i.d. lognormal price noise, 25 bps daily, 8 seeds), an explanation-robustness criterion [20]; and **completeness**—the reconstruction gap $|\sum_i \phi_i - F(N)|$, which (5) drives to zero by construction.

VI. EXPERIMENTAL SETUP

1) Data: Daily adjusted OHLCV (dividends reinvested) from Yahoo Finance for ten liquid U.S. large caps spanning six GICS sectors—AAPL, MSFT, NVDA, AMZN, GOOGL, JPM, XOM, JNJ, PG, CAT—plus SPY as the market proxy, 2015-01-02 through 2026-08-13 (2,920 trading days per name).

2) Protocol: Decisions every 5 trading days. Each model scores all names; the portfolio holds the top k=3 names with score above a threshold θ in equal weight (unfilled slots stay in cash at zero interest), with a 1-day execution lag and 10 bps one-way transaction costs on turnover; weights drift between rebalances. Direction hit rate counts recommendations with $|F| > \theta$ whose sign matches the 21-day forward return; the information coefficient (IC) is the per-date Spearman rank correlation between scores and forward returns.

3) Calibration/test split: All hyperparameters are selected on 2017–2021 only and frozen: best single agent (momentum, calibration IC +0.086), β=8, α_0=0.6, R=5, ridge λ=1, and per-model $\theta \in \{0.02, 0.05, 0.10\}$ by calibration Sharpe. All reported results use the untouched test window 2022-01-01 to 2026-08-13 (232 decision dates, ∼2,300 scored decisions per model).

4) Statistical inference: Sampling uncertainty is quantified with a stationary block bootstrap [26] over the test-window daily returns (B=2,000, mean block 21 days, fixed seed), resampling all strategies and the benchmark *jointly* so cross-strategy correlation is preserved; we report 95% percentile intervals for Sharpe differences and bootstrap probabilities for protocol orderings.

5) Reproducibility: The pipeline is ∼1,400 lines of Python (NumPy, pandas, Matplotlib); agents are deterministic, so the artifact regenerates every table, figure, and inline number in this paper from the cached raw data.[1]

TABLE I
OUT-OF-SAMPLE PORTFOLIO PERFORMANCE, 2022-01–2026-08, NET OF 10 BPS COSTS. CAGR = COMPOUND ANNUAL GROWTH RATE; MAXDD = MAXIMUM DRAWDOWN.

Strategy	CAGR	Vol.	Sharpe	Sortino	MaxDD	Calmar
Single agent (momentum)	16.2%	24.2%	0.74	1.11	-27.5%	0.59
Uniform vote	14.4%	18.6%	0.81	1.09	-24.9%	0.58
Confidence-weighted fusion	21.4%	21.1%	**1.03**	1.46	-31.7%	0.67
Iterative debate	21.5%	21.3%	1.02	1.44	-31.0%	0.69
Hierarchical coordinator	9.0%	20.5%	0.52	0.74	-37.0%	0.24
Learned fusion (ridge)	18.3%	22.9%	0.85	1.19	-33.6%	0.55
SPY buy–and–hold	12.8%	17.5%	0.78	1.08	-24.5%	0.52
Equal-weight universe	27.8%	30.3%	0.96	1.47	-40.1%	0.69

VII. RESULTS

A. Investment performance

Table I and Fig. 2 give the headline comparison. CWF and DEBATE are statistically indistinguishable from each other (Sharpe 1.03 and 1.02; CAGR 21.4% and 21.5%) and clearly ahead of passive SPY (0.78, 12.8%), the best single agent (0.74), and uniform voting (0.81). The calendar-year decomposition (Table II) shows where the edge lives: in the

[1]This paper is a research study of collaborative reasoning and explainability; nothing in it constitutes investment advice.

TABLE II
CALENDAR-YEAR RETURNS (TEST WINDOW). †2026 THROUGH AUG. 13.

Strategy	2022	2023	2024	2025	2026†
Single agent	-8.6%	14.9%	41.5%	9.2%	22.7%
Uniform vote	4.7%	19.5%	17.0%	3.8%	21.8%
Confidence-weighted fusion	12.4%	28.3%	31.5%	0.7%	27.5%
Iterative debate	10.3%	23.5%	26.8%	9.2%	29.5%
Hierarchical coordinator	1.1%	6.1%	11.6%	10.6%	12.1%
Learned fusion	9.1%	36.5%	32.8%	-8.1%	19.2%
SPY buy–and–hold	-18.2%	26.2%	24.9%	17.7%	14.7%

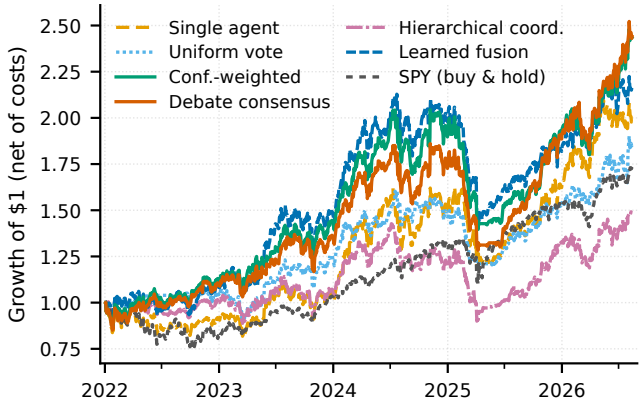


Fig. 2. Out-of-sample growth of $1, net of transaction costs. Line styles provide a redundant encoding for grayscale printing.

2022 bear market CWF returned $+12.4\%$ against -18.2% for SPY, because the regime and risk agents pull committee scores below the entry threshold and the portfolio de-risks into cash; the committees then participate in the 2023–2024 recovery. The equal-weight universe benchmark illustrates the concentration alternative: higher CAGR (27.8%, driven by NVDA's run) but at -40.1% maximum drawdown and a lower Sharpe (0.96) than CWF.

Two committee designs underperform, informatively. VOTE treats a market-level regime agent as one-fourth of every stock-level decision regardless of its confidence, diluting the cross-sectional signal. COORD is the striking one: giving the risk agent multiplicative veto authority (Eq. 3) halves the Sharpe ratio (0.52) and *worsens* the maximum drawdown (-37.0%): the gates re-admit exposure late after volatility spikes, so the committee sells weakness and misses V-shaped recoveries—over-conservatism is not safety. LEARN sits in between (0.85 Sharpe) with the best hit rate (58.0%) but the weakest 2025, when its calibration-era contrarian tilt fought a trending market.

B. Decision-level quality

Table III reports recommendation-level metrics. Hit rates sit in the 51–58% band and mean ICs are small ($|\overline{\text{IC}}| \leq 0.044$), consistent with the weak-predictability regime documented for equity ML [21]; portfolio results are driven by the extreme tail of the score distribution that top-k selection actually trades,

TABLE III
DECISION-LEVEL RECOMMENDATION QUALITY (TEST WINDOW). COVERAGE = FRACTION OF SCORED DECISIONS WITH $|F| > \theta$; DISP. = MEAN PRE-AGGREGATION DISPERSION OF AGENT SCORES. THE SINGLE AGENT HAS NO COMMITTEE DISPERSION.

Model	Hit rate	Coverage	$\overline{\text{IC}}$	t(IC)	Disp.
Single (mom.)	52.1%	96%	+0.010	+0.41	–
Uniform vote	54.4%	93%	-0.044	-1.83	0.50
Conf.-weighted	55.4%	97%	+0.012	+0.49	0.50
Debate	55.5%	95%	+0.011	+0.42	0.18
Hier. coord.	50.8%	86%	-0.044	-1.77	0.42
Learned (ridge)	58.0%	97%	-0.006	-0.24	0.50

TABLE IV
STATISTICAL AND COST ROBUSTNESS (TEST WINDOW). ΔSHARPE VS. SPY WITH STATIONARY-BOOTSTRAP 95% CI AND ONE-SIDED p-VALUE; SHARPE WHEN ONE-WAY COSTS RISE TO 25 BPS; ONE-WAY ANNUAL TURNOVER (×NAV).

Model	ΔSharpe vs. SPY [95% CI]	$p_{\leq 0}$	Sh.@25 bp	Turn.
Single (mom.)	-0.04 $[-0.72, +0.64]$	0.57	0.67	12×
Uniform vote	$+0.04$ $[-0.79, +0.83]$	0.50	0.57	30×
Conf.-weighted	$+0.25$ $[-0.57, +1.02]$	0.30	0.81	30×
Debate	$+0.24$ $[-0.57, +1.06]$	0.31	0.81	29×
Hier. coord.	-0.26 $[-1.02, +0.50]$	0.77	0.31	29×
Learned (ridge)	$+0.07$ $[-0.59, +0.70]$	0.45	0.69	24×

which is why VOTE's negative full-rank IC coexists with a positive portfolio Sharpe. LEARN attains the best hit rate (58.0%) yet not the best Sharpe: directional accuracy and risk-adjusted portfolio return are related but distinct objectives, and reporting only one of them—as recommendation-system evaluations often do—overstates certainty about system quality.

C. Statistical and cost robustness

Table IV separates what this evaluation can and cannot establish. Absolute outperformance of the market is *not* statistically separable at a 4.6-year horizon: the best protocol's Sharpe advantage over SPY (CWF: $+0.25$, 95% CI $[-0.57, +1.02]$, $p_{\leq 0} = 0.30$) has a confidence interval spanning zero, exactly as expected for equity strategies over windows this short. The paper's comparative claims, however, are robust: because all protocols are evaluated on identical agents, dates, and resampled paths, protocol *orderings* difference out common market noise—CWF beats the veto hierarchy in 99.7% of resamples, the best single agent in 85.9%, and DEBATE beats uniform voting in 74.9%. Raising one-way costs from 10 to 25 bps at the committees' $\sim 30\times$ annual one-way turnover (vs. $12\times$ for the single agent) lowers every Sharpe (CWF/DEBATE to 0.81/0.81) but leaves the protocol ordering unchanged; at weekly cadence, cost efficiency, not signal quality, becomes the binding constraint on absolute performance.

D. Debate dynamics and disagreement as uncertainty

Fig. 3(a) shows consensus formation under Eq. (2): dispersion falls monotonically from 0.50 to 0.18 over five rounds—agents converge, they are not forced to agree—while

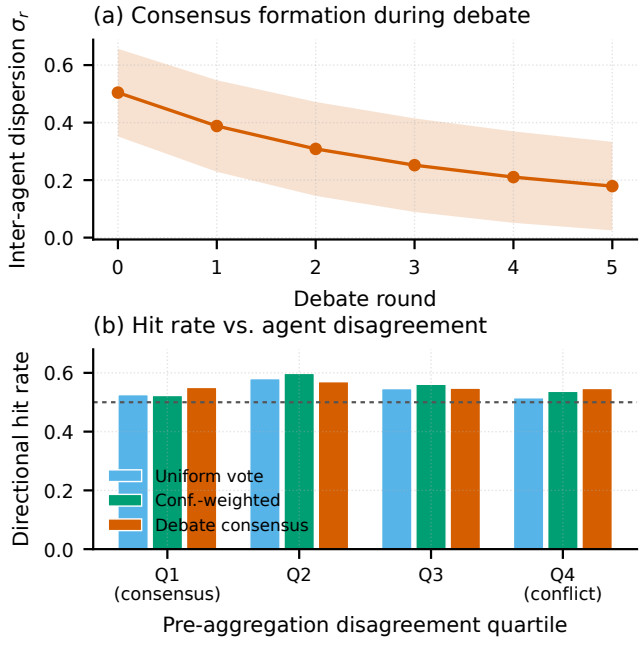


Fig. 3. (a) Debate dynamics: inter-agent dispersion per round (mean ± 1 s.d. across all test decisions), from σ_0=0.50 to σ_5=0.18. (b) Directional hit rate by quartile of pre-aggregation disagreement; the dashed line marks the 50% coin-flip level.

TABLE V
EXPLANATION QUALITY (TEST WINDOW). FAITH. = MEAN CORRELATION OF LEAVE-ONE-OUT MARGINALS WITH EXACT SHAPLEY ATTRIBUTIONS; STAB. = MEAN COSINE SIMILARITY OF ϕ UNDER 25 BPS INPUT NOISE; $|\Sigma\phi - F|$ = RECONSTRUCTION GAP OF THE SHAPLEY EXPLANATION.

Model	Faith.	Stab.	$\lvert\Sigma\phi - F\rvert$	Top agent (share)
Single (mom.)	1.000	0.995	0.0000	Momentum (100%)
Uniform vote	1.000	0.995	0.0000	Regime (32%)
Conf.-weighted	0.839	0.996	0.0000	Regime (38%)
Debate	0.780	0.996	0.0000	Regime (38%)
Hier. coord.	0.938	0.997	0.0000	Momentum (57%)
Learned (ridge)	0.988	0.999	0.0000	Regime (29%)

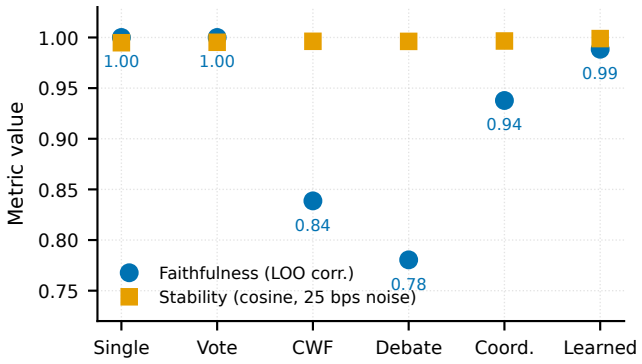


Fig. 4. Explanation faithfulness degrades as committee reasoning becomes more interactive, while exact-Shapley stability stays uniformly high.

final recommendations track CWF closely (Sharpe 1.02 vs. 1.03). Debate thus buys a *calibrated consensus trace* (a per-recommendation record of who conceded and how far) at no performance cost, but also at no performance gain: with numeric opinions and confidence-weighted influence, iterated deliberation is largely a re-parameterization of confidence weighting. Claims that debate *improves* decisions [11] should therefore be tested against this equivalence class, not against single agents only.

Fig. 3(b) evaluates disagreement as an uncertainty signal. The pattern is an inverted U: hit rates peak in the second disagreement quartile (up to 59.8% for CWF) and fall toward both extremes; unanimous committees are *not* the most accurate ones. Full consensus concentrates in strong-trend states where the contrarian analyst's dissent has already capitulated—exactly where reversals surprise. Dispersion is therefore usable for triage (route mid-disagreement decisions to autonomy, extremes to human review) but is not a monotone confidence score.

E. Explanation quality

Table V and Fig. 4 contain the paper's central explainability result. The reconstruction gap is $\sim 10^{-4}$ or below for every model—the efficiency axiom holds numerically—and stability under 25 bps input noise is uniformly high (≥ 0.995 cosine similarity), so exact agent-level Shapley explanations are both complete and robust for all six reasoning models. What varies is *faithfulness of the cheap alternative*: leave-one-out deletion explanations correlate perfectly with Shapley attributions for additive committees (SINGLE, VOTE: 1.00), nearly for the quasi-linear learned fusion (0.99), but degrade through the coordinator (0.94) and confidence-weighted fusion (0.84, where deleting an agent renormalizes every other weight) down to 0.78 for debate, whose iterated influence makes single-deletion tests systematically misstate who moved the committee. The practical reading for regulated deployment: as collaboration becomes more interactive, deletion-style audits—the default in industry tooling—become quietly unfaithful, while exact coalition attribution remains cheap at agent granularity (2^4 evaluations) and should be preferred.

Fig. 5 shows that the reasoning model also reshapes the *content* of explanations: confidence weighting concentrates influence in the high-conviction regime agent (38%), the hierarchy concentrates it in momentum (57%) while shrinking its own oversight gates to single digits, and learned fusion spreads influence most evenly. Two committees with identical inputs and similar returns can thus offer materially different accounts of *why*—an auditability consideration invisible to performance metrics.

F. Worked example

Table VI traces one real, deliberately conflicted decision: GOOGL on 2023-01-09, after a -41% drawdown, below its 200-day average, in a bear regime. The momentum analyst is bearish at high confidence (-0.92, c=0.96); the mean-reversion analyst dissents ($+0.32$). Every heuristic commit-

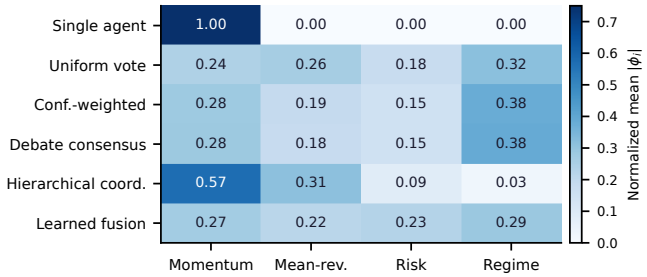


Fig. 5. Global agent importance (mean $|\phi_i|$, row-normalized): the reasoning model reshapes *whose voice* the recommendation reflects.

TABLE VI
WORKED EXAMPLE: GOOGL, 2023-01-09. TOP: AGENT OPINIONS AND EVIDENCE. BOTTOM: COMMITTEE OUTPUTS WITH SHAPLEY ATTRIBUTIONS (M/V/R/G = MOMENTUM/MEAN-REV./RISK/REGIME). REALIZED 21-DAY FORWARD RETURN: $+7.4\%$.

Agent	Key evidence	s_i	c_i
Momentum	12–1 mom. -31.1%; 50/200 trend -14.5%; below 200-d SMA	-0.92	0.96
Mean-rev.	Bollinger $z = -0.56$; RSI(14) $= 43$	$+0.32$	0.37
Risk	vol. percentile 0.39; drawdown -41%	-0.26	0.48
Regime	SPY below 200-d SMA; breadth 0.5	-0.74	0.79

Committee	F	ϕ_{M}	ϕ_{V}	ϕ_{R}	ϕ_{G}
Single (mom.)	-0.92	-0.92	$+0.00$	$+0.00$	$+0.00$
Uniform vote	-0.40	-0.42	$+0.34$	-0.01	-0.31
Conf.-weighted	-0.87	-0.54	$+0.10$	-0.09	-0.34
Debate	-0.89	-0.54	$+0.10$	-0.10	-0.35
Hier. coord.	-0.91	-0.98	$+0.12$	-0.02	-0.03
Learned (ridge)	$+0.92$	$+0.15$	$+0.23$	$+0.22$	$+0.32$

tee sides with momentum (scores -0.40 to -0.92)—and the attributions show precisely how: under CWF, momentum and regime carry $\phi = -0.54$ and -0.34 while the dissent contributes $+0.10$. LEARN, whose calibration made high-confidence momentum contrarian, recommends *buy* ($+0.92$). The realized 21-day return was 7.4%: the transparent committees were transparently wrong, and the explanation object (Fig. 6) records the overruled dissent—exactly the audit trail one wants when reviewing failures, and direct evidence that explainability and correctness are separate properties requiring separate evaluation.

VIII. DISCUSSION AND LIMITATIONS

1) Design guidance: Within this framework, confidence-weighted fusion is the efficient point: top-tier performance, one interpretable equation, near-perfect explanation stability, and only a moderate faithfulness penalty; debate adds a consensus trace and residual-dispersion uncertainty at equal performance and the largest faithfulness penalty; hard veto hierarchies need explicit re-entry logic before they add safety. The non-monotone interpretability–performance relationship

AVOID GOOGL — committee conviction -0.89 (post-debate dispersion 0.06: high consensus). Momentum analyst (conf. 0.96): 12–1 momentum -31.1%, price below its 200-day average with a -14.5% 50/200 trend gap. Regime analyst (conf. 0.79): index below its 200-day average; market breadth 0.5. Mean-reversion analyst dissents ($+0.32$, conf. 0.37): RSI 43 after a -41% drawdown. Attribution ϕ: momentum -0.54, regime -0.35, risk -0.10, mean-reversion $+0.10$. Realized 21-day return: $+7.4\%$ (recommendation incorrect).

Fig. 6. Natural-language brief auto-generated from the DEBATE explanation object for the decision of Table VI. The brief records the dissent and is explicit when hindsight proves the committee wrong.

is consistent with Rudin's argument [19] that high-stakes accuracy rarely requires opacity.

2) Limitations: The universe is ten mega-caps chosen from today's index leaders, so all long-only results (including the equal-weight benchmark) inherit survivor/selection bias; cross-sectional breadth is far below institutional scale. The evaluation is one chronological split; the block bootstrap quantifies sampling variability within it—and shows absolute market outperformance is not established (Sec. VII-C)—but cannot substitute for longer horizons or additional regimes, so backtest-overfitting risk remains [25]; our claims are comparative across protocols, not absolute, and cash earns zero interest. Agents are deterministic policy heads: this isolates the collaboration variable and guarantees reproducibility, but measures no prompt-level phenomena of LLM committees (sycophancy, conformity cascades [12]), and our debate is a numeric consensus process, not argumentative text exchange. Sentiment and fundamental statement data—natural LLM-agent inputs—are absent. Finally, explanation quality is evaluated with functional metrics, not user studies with investors or supervisors.

3) Ethics and regulatory context: The system is a research prototype for studying collaboration and explainability, not an advisory product; deploying such agents for real investors triggers model-risk and high-risk-AI obligations [27], [28], requires human oversight, and must communicate uncertainty honestly—which, per Sec. VII-D, means resisting the temptation to sell committee consensus as confidence.

IX. CONCLUSION

We designed a financial multi-agent advisory framework whose collaborative reasoning layer is a swappable, coalition-evaluable operator, and used that property to give six committee designs—from a single analyst to debate, hierarchy, and learned fusion—an identical, exact, Shapley-based explanation object. A 4.6-year out-of-sample study on real equity data shows that the collaboration protocol is a first-order design choice (Sharpe 0.52–1.03 across protocols over the same agents and data), that confidence-weighted fusion and debate lead while veto hierarchies over-conserve, that committee disagreement is an inverted-U uncertainty signal suited to triage rather than a confidence score, and that deletion-based audits grow unfaithful precisely as collaboration grows interactive—while exact agent-level attribution stays cheap, stable, and complete. Future work: instantiating the same interfaces with

LLM analysts over news and fundamentals; learning the collaboration operator itself; reinforcement-learned position sizing; and user studies of explanation utility with practitioners and regulators.

ACKNOWLEDGMENT

The authors thank the maintainers of the open-source Python scientific stack used throughout this study.

REFERENCES

[1] S. Wu *et al.*, "BloombergGPT: A large language model for finance," *arXiv:2303.17564*, 2023.
[2] H. Yang, X.-Y. Liu, and C. D. Wang, "FinGPT: Open-source financial large language models," *arXiv:2306.06031*, 2023.
[3] D. Araci, "FinBERT: Financial sentiment analysis with pre-trained language models," *arXiv:1908.10063*, 2019.
[4] X.-Y. Liu *et al.*, "FinRL: A deep reinforcement learning library for automated stock trading in quantitative finance," *arXiv:2011.09607*, 2020.
[5] Y. Yu *et al.*, "FinMem: A performance-enhanced LLM trading agent with layered memory and character design," *arXiv:2311.13743*, 2023.
[6] W. Zhang *et al.*, "A multimodal foundation agent for financial trading: Tool-augmented, diversified, and generalist," in *Proc. ACM SIGKDD*, 2024, arXiv:2402.18485.
[7] Y. Xiao, E. Sun, D. Luo, and W. Wang, "TradingAgents: Multi-agents LLM financial trading framework," *arXiv:2412.20138*, 2024.
[8] J. Wei *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," in *Adv. NeurIPS*, 2022.
[9] S. Yao *et al.*, "ReAct: Synergizing reasoning and acting in language models," in *Proc. ICLR*, 2023.
[10] N. Shinn *et al.*, "Reflexion: Language agents with verbal reinforcement learning," in *Adv. NeurIPS*, 2023.
[11] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch, "Improving factuality and reasoning in language models through multiagent debate," in *Proc. ICML*, 2024.
[12] T. Liang *et al.*, "Encouraging divergent thinking in large language models through multi-agent debate," in *Proc. EMNLP*, 2024.
[13] Q. Wu *et al.*, "AutoGen: Enabling next-gen LLM applications via multi-agent conversation," *arXiv:2308.08155*, 2023.
[14] S. Hong *et al.*, "MetaGPT: Meta programming for a multi-agent collaborative framework," in *Proc. ICLR*, 2024.
[15] M. H. DeGroot, "Reaching a consensus," *J. Amer. Statist. Assoc.*, vol. 69, no. 345, pp. 118–121, 1974.
[16] L. S. Shapley, "A value for n-person games," in *Contributions to the Theory of Games II*, Princeton Univ. Press, 1953, pp. 307–317.
[17] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Adv. NeurIPS*, 2017.
[18] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," in *Proc. ACM SIGKDD*, 2016, pp. 1135–1144.
[19] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Mach. Intell.*, vol. 1, pp. 206–215, 2019.
[20] A. B. Arrieta *et al.*, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, 2020.
[21] S. Gu, B. Kelly, and D. Xiu, "Empirical asset pricing via machine learning," *Rev. Financial Studies*, vol. 33, no. 5, pp. 2223–2273, 2020.
[22] N. Jegadeesh and S. Titman, "Returns to buying winners and selling losers: Implications for stock market efficiency," *J. Finance*, vol. 48, no. 1, pp. 65–91, 1993.
[23] H. Markowitz, "Portfolio selection," *J. Finance*, vol. 7, no. 1, pp. 77–91, 1952.
[24] W. F. Sharpe, "Mutual fund performance," *J. Business*, vol. 39, no. 1, pp. 119–138, 1966.
[25] D. H. Bailey, J. M. Borwein, M. López de Prado, and Q. J. Zhu, "Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance," *Notices Amer. Math. Soc.*, vol. 61, no. 5, pp. 458–471, 2014.
[26] D. N. Politis and J. P. Romano, "The stationary bootstrap," *J. Amer. Statist. Assoc.*, vol. 89, no. 428, pp. 1303–1313, 1994.
[27] Board of Governors of the Federal Reserve System, "Supervisory guidance on model risk management," SR Letter 11-7, 2011.
[28] European Parliament and Council, "Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (AI Act)," 2024.
[29] A. W. Lo, "The adaptive markets hypothesis," *J. Portfolio Manage.*, vol. 30, no. 5, pp. 15–29, 2004.