Toward Reliable Evaluation of LLM-Based Financial Multi-Agent Systems: Taxonomy, Coordination Primacy, and Cost Awareness

Phat Nguyen[1][0009−0004−0671−2487] and Thang Pham[2][0000−0003−3984−641X]

[1] Georgia Institute of Technology, Atlanta GA 30332, USA
cherry.07.skr@gmail.com
[2] Adobe Inc., San Jose CA 95110, USA
thanpham@adobe.com

Abstract. Multi-agent systems based on large language models (LLMs) for financial trading have grown rapidly since 2023, yet the field lacks a shared framework for understanding what drives performance or for evaluating claims credibly. This survey makes three contributions. First, we introduce a four-dimensional taxonomy, covering architecture pattern, coordination mechanism, memory architecture, and tool integration; applied to 12 multi-agent systems and two single-agent baselines. Second, we formulate the *Coordination Primacy Hypothesis* (CPH): inter-agent coordination protocol design is a primary driver of trading decision quality, often exerting greater influence than model scaling. CPH is presented as a falsifiable research hypothesis supported by tiered structural evidence rather than as an empirically validated conclusion; its definitive validation requires evaluation infrastructure that does not yet exist in the field. Third, we document five pervasive evaluation failures (look-ahead bias, survivorship bias, backtesting overfitting, transaction cost neglect, and regime-shift blindness) and show that these can reverse the sign of reported returns. Building on the CPH and the evaluation critique, we introduce the *Coordination Breakeven Spread* (CBS), a metric for determining whether multi-agent coordination adds genuine value net of transaction costs, and propose minimum evaluation standards as prerequisites for validating the CPH.

Keywords: Multi-agent systems · LLM agents · financial decision-making · coordination mechanisms · portfolio optimization · trading evaluation

1 Introduction

Financial markets reward decision quality under uncertainty. When LLM-based agents entered this domain, the initial approach treated the entire investment workflow as a single prompt-to-trade pipeline. These monolithic systems face a fundamental constraint: a single context window cannot simultaneously perform macroeconomic regime identification, earnings quality assessment, technical pattern recognition, risk budgeting, and execution optimization. The dominant

paradigm now decomposes these tasks across cooperating agents, and systems such as FinCon [25], TradingAgents [22], and HedgeAgents [15] report substantial gains, with claimed cumulative returns ranging from 23% to over 400%. However, the field has not established whether these gains reflect genuine design advances or artifacts of inconsistent evaluation methodology.

This survey addresses that gap. Rather than cataloguing systems or ranking them by reported performance, we ask a more tractable prior question: *Given that cross-system comparisons are currently unreliable, what design choices are most worth investigating rigorously once evaluation standards improve?* Our approach proceeds in four steps: a taxonomy decomposing the design space (Section 3), documentation of five systematic evaluation failures (Section 4), the *Coordination Primacy Hypothesis* (CPH) derived from structural patterns that survive those failures (Section 5), and the *Coordination Breakeven Spread* (CBS) metric that operationalizes the hypothesis in deployment (Section 6). Our contributions are analytical; we do not report new empirical results.

2 Related Work

Ding et al. [6] survey LLM agents for financial trading but treat multi-agent coordination as one topic among many, without systematic comparison of coordination trade-offs. General multi-agent surveys [11,20] analyze communication protocols abstractly, without confronting domain-specific constraints such as coordination latency measurable in basis points of adverse price movement. Sun et al. [19] survey LLM-based multi-agent reinforcement learning; no published financial system has successfully integrated LLM agents with formal optimization guarantees. FinCon's verbal reward function is a step toward structured decision optimization but lacks formal convergence guarantees.

Our work differs in two respects. We treat financial multi-agent systems (MAS) as decision architectures rather than software architectures, so every design description is accompanied by its decision-quality implication. We also foreground evaluation methodology as a first-class concern: the gap between reported and robust performance is large enough to affect whether published claims should be acted upon.

3 A Taxonomy of Design Patterns

3.1 System Selection

Candidate systems were drawn from the LLM-based financial trading literature published between 2023 and 2026. Inclusion required that a system employ distinct LLM roles for active trading or portfolio allocation and provide sufficient architectural detail for four-dimensional classification. Purely single-agent systems, rule-based pipelines, and systems lacking architectural documentation were excluded. These criteria yield 12 systems, selected to maximize coverage across the four taxonomy dimensions: FinCon [25], TradingAgents [22], HedgeAgents [15],

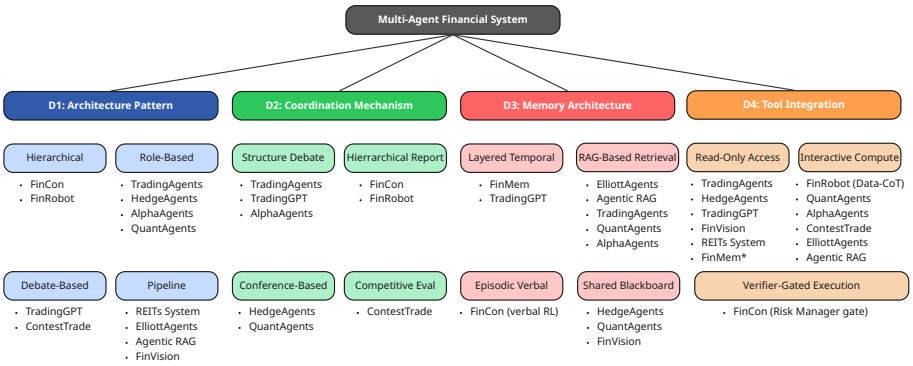

Fig. 1: Taxonomy of LLM-based multi-agent financial systems across four design dimensions. FinVision, ElliottAgents, REITs System, and Agentic RAG are not listed in Coordination Mechanism because they simply utilize a sequential pass-through (minimal coordination).

ContestTrade [28], FinVision [10], TradingGPT [13], QuantAgents [16], AlphaAgents [2], ElliottAgents [21], FinRobot [24], Agentic RAG [5], and Chinese Public REITs system [12]. We include FinMem [26] (memory-focused) and FinAgent [27] (tool-augmented) as single-agent baselines for comparison in Table 1.

3.2 Four Dimensions

We classify systems along four dimensions (Fig. 1), intended to decompose hybrid designs so that future controlled experiments can vary one dimension while holding others constant.

D1: Architecture Pattern. *Hierarchical* architectures use a manager agent to arbitrate specialist inputs; the manager's ability to weight conflicting inputs is the binding constraint. *Role-based* designs map agents to professional departments, supporting auditability. *Debate-based* architectures improve calibration under signal ambiguity but incur overhead costly when execution speed matters. *Pipeline* architectures offer low latency but no error correction.

D2: Coordination Mechanism. *Structured debate* improves accuracy over two to four rounds [7,4] but risks Degeneration-of-Thought [17], where agents converge to a shared wrong answer through social pressure. *Hierarchical reporting* uses selective knowledge propagation to reduce noise, ensuring only decision-relevant feedback reaches specialists. *Conference-based* coordination activates group discussions adaptively but requires precise triggers to avoid activating complex protocols during routine trading. *Competitive evaluation* rewards contrarian accuracy rather than consensus, avoiding consensus bias entirely. The

absence of competitive risk-adjusted returns among systems using a sequential pass-through provides indirect motivating evidence for the CPH (Section 5.2), though this observation is subject to the same evaluation confounds documented in Section 4 and should be treated as motivating rather than evidential.

D3: Memory Architecture. *Layered temporal* memory risks assuming fixed relevance decay during structural breaks. *RAG-based retrieval* allows for high-granularity data access, but introduces experience-following behaviour [23], amplifying anchoring bias. *Episodic verbal* memory supports compliance and auditability but risks update lag. *Shared blackboard* state enables real-time sharing but propagates errors system-wide.

D4: Tool Integration. *Read-only access* depends on LLM numerical reasoning, a documented weakness. *Interactive computation* addresses this but introduces code correctness as a failure mode. *Verifier-gated execution* validates outputs before action and is preferred for institutional deployment.

3.3 Cross-System Observations

Table 1 reports metrics for the eight systems that publish them; the remaining six (TradingGPT, AlphaAgents, ElliottAgents, FinRobot, REITs System, Agentic RAG) are framework papers without standardized trading metrics. The evaluation quality assessment scores each system against the five criteria introduced in Section 4: no system satisfies more than two criteria. These eight systems differ along at least five methodological dimensions simultaneously: evaluation period, asset universe, market regime, cost model, and baseline choice. Normalizing across these dimensions would produce numbers that appear comparable but conceal the differences that matter most for practitioners. Instead, we assess how trustworthy each system's reported numbers are (Evaluation Quality column) and section 4 addresses the evaluation inconsistencies in depth.

Although the results in Table 1 are not directly comparable, a qualitative structural pattern is worth noting. Systems with explicit coordination mechanisms more often report extended evaluation horizons and live deployment, for example HedgeAgents posts a 405% three-year return and QuantAgents sustains 1.76–2.02 live Sharpe (both subject to the evaluation quality limitations in Section 4), whereas pipeline designs rarely disclose comparable long-horizon, risk-adjusted results. High Sharpe ratios observed over short, bullish windows (e.g., debate-based TradingAgents evaluations) highlight the importance of temporal robustness rather than establishing superiority. While not causal evidence, this pattern suggests coordination complexity may correlate with demonstrated robustness, motivating the CPH (Section 5).

Note: Metrics across rows are **not directly comparable** owing to differences in evaluation period, asset universe, market regime, and cost assumptions.

Table 1: Reported performance metrics. Evaluation quality scores each system against five criteria: (1) contamination control, (2) point-in-time universe, (3) rolling-window reporting, (4) net-of-cost returns, (5) regime coverage. ✓ = satisfied, × = not satisfied. † FinMem's reported 23% return on MSFT reversed to −22% under FINSABER controlled conditions.

System (Agents)	Sharpe Ratio	Cumulative Return	Annual Return	Max Drawdown	Eval Period	Evaluation Details	Evaluation Quality					
							Contam.	Point-in-Time	Rolling Win.	Net Costs	Regime Cov.	
FinAgent [27] (1+tools)	1.43–2.01	—	31.9–92.3%	5.57–13.2%	∼6 mo	Six datasets; stocks and crypto. Best ARR 92.27% on TSLA.	✓	×	×	×	✓	■ 2/5
FinCon [25] (7+1)	3.26	114%	—	16.2%	∼18 mo	Jan 2022–Jun 2023; 6 US stocks	✓	×	×	×	✓	■ 2/5
HedgeAgents [15] (4+1)	2.41	405%	71.60%	∼14%	3 yr	2021–2023; BTC, DJ30, Forex; multi-asset	✓	×	×	×	✓	■ 2/5
QuantAgents [16] (Multi)	3.11	∼300%	58.7%	16.86%	3 yr+live	Jan 2021–Dec 2023; NASDAQ-100.	✓	×	×	×	✓	
	1.76-2.02	98-112%	—	—	Q3 2024 - Q1 2025	Live trading A-stock and HK-stock markets						■ 2/5
ContestTrade [28] (Multi)	3.12	52.8%	—	12.41%	—	NASDAQ-100	✓	×	×	×	×	■ 1/5
FinVision [10] (4)	1.20–1.72	—	14.8–42.1%	12.09–14.38%	∼7 mo	AAPL, MSFT, AMZN; predominantly bullish window; pipeline architecture.	✓	×	×	×	×	■ 1/5
TradingAgents [22] (7)	5.60–8.21	23–27%	24.9–30.5%	0.91–2.1%	3 mo	Jan–Mar 2024; AAPL, GOOGL, AMZN. Sharpe ratio inflated by short bullish window.	✓	×	×	×	×	■ 1/5
FinMem [26] † (1)	0.23–2.67	23-61.7%	—	10.8–22.9%	∼1 yr	Oct 2022-Apr 2023	×	×	×	×	×	
	1.4 → -1.24	**23→ −22%**		14.9 → -29	6 mo → 8 mo	MSFT only, Under FINSABER controlled re-evaluation.						■ 0/5

■ 4–5/5 Relatively credible ■ 2–3/5 Partial credibility ■ 0–1/5 Low credibility

4 Evaluation Failures in the Published Literature

The design diversity in Section 3 means that any observed performance difference between two systems could reflect a genuine design advantage, a difference in evaluation conditions, or both. Five systematic failures prevent these explanations from being distinguished, making cross-system comparison unreliable as a basis for design conclusions.

4.1 Look-Ahead Bias

LLMs trained through 2024 may have encountered financial outcomes for periods used in backtesting, effectively retrieving rather than predicting. StockBench [3] addresses this with DJIA data from March to July 2025; most LLM-based agents fail to outperform buy-and-hold under these conditions. A second manifestation is feature leakage through retrieval: imprecisely timestamped RAG databases can inject future information into historical queries. FinAgent's multi-step retrieval and Agentic RAG's cross-encoder re-ranking are both vulnerable in the absence of documented timestamp controls.

4.2 Survivorship Bias

Most systems evaluate on stock universes selected at evaluation time, excluding delisted companies that are disproportionately poor performers. Elton et al. [8] estimated 0.9% annual survivorship bias in mutual fund returns; for individual stock selection the effect is larger. FINSABER [14] addresses this with historical index constituent lists.

4.3 Backtesting Overfitting

LLM-based multi-agent systems have extensive hyperparameters (agent count, debate rounds, memory depth, temperature, prompt templates, evaluation windows), creating combinatorial space prone to overfitting. FinMem's reported 23.26% cumulative return on MSFT became −22.04% under a slightly different but equally defensible window with transaction costs included [14]. A sign reversal of this magnitude is consistent with an overfitted system.

4.4 Transaction Cost Neglect

Round-trip costs of 10 to 20 basis points can compound to 25 to 50 percentage points of annual drag for daily-trading systems. Of systems surveyed, only FINSABER and StockBench explicitly model transaction costs. HedgeAgents' 405%, FinCon's 114%, and ContestTrade's 52.8% are all gross of costs. This failure is particularly consequential for MAS: coordination-driven signal improvements may increase trading frequency without proportionally improving per-trade alpha, converting a nominal performance advantage into net underperformance.

4.5 Regime-Shift Blindness

Most evaluations cover six to twelve months within a single market regime, providing no cross-regime evidence. Only HedgeAgents explicitly addresses regime adaptation; its three-year evaluation spanning 2021–2023 is the strongest available cross-regime evidence among surveyed systems. TradingAgents reports an extraordinary Sharpe of 5.60 to 8.21 based solely on a three-month bullish window (January–March 2024) during rallies in AAPL, GOOGL, and AMZN. A Sharpe ratio at this level, annualized from a single favourable regime, is statistically consistent with trend following in a favorable regime rather than genuine risk-adjusted alpha, and is consistent with regime shift blindness producing unreliable metrics.

4.6 Consolidated Minimum Standards

1. **Contamination control.** Evaluation period should post-date model training, or a post-training ablation should be provided.
2. **Point-in-time universe.** Asset universe should reflect historical index composition at each evaluation date.

3. **Rolling-window reporting.** Performance across multiple non-overlapping windows with variance estimates.
4. **Net-of-cost returns.** Explicit transaction cost model covering commissions, half-spread, and market impact.
5. **Regime coverage.** Evaluation spanning multiple regimes or explicit adversarial stress testing.

No system in our survey satisfies all five (see evaluation quality scores in Table 1). FinCon, HedgeAgents, FinAgent, and QuantAgents each satisfy only 2/5, while TradingAgents, ContestTrade, and FinVision reach just 1/5. Notably, FinMem scores 0/5, with its reported 23% return on MSFT reversing to -22% under controlled re-evaluation. Building a benchmark satisfying all five simultaneously is among the most pressing infrastructure needs in this field.

5 The Coordination Primacy Hypothesis

5.1 Motivation

The evaluation failures above make precise quantitative comparison unreliable, but structural observations remain informative even when specific return figures do not: which systems survive the transition to live trading, which coordination patterns appear consistently across independent research groups, and which designs collapse under controlled re-evaluation. The CPH is derived from these patterns rather than from cross-system performance rankings.

5.2 Hypothesis Statement

Coordination Primacy Hypothesis holds that the inter-agent coordination protocol is the most consequential structural factor in trading decision quality among the four taxonomy dimensions, exerting greater influence than model selection.

This is a falsifiable claim: upgrading the LLM backbone within a fixed coordination protocol should yield smaller performance improvements than replacing the coordination protocol, holding all other design choices constant. The hypothesis does not assert that coordination is sufficient; only that it is the most consequential dimension to optimize.

5.3 Supporting Evidence (Tiered)

Tier 1 – Live-Market Benchmarking (Strongest). Available evidence suggests that framework architecture is a more consequential predictor of profitability than model selection: weaker models within sophisticated coordination structures tend to outperform frontier models in linear pipelines across the benchmarks examined. This is the most credible available evidence for the CPH, though regime diversity covered by AMA [18] remains limited and the finding should be treated as strongly suggestive rather than definitive.

Tier 2 – Ablation Studies (Moderate). In FinCon and TradingAgents, removing the coordination reduced the Sharpe ratio by 15–30%; while substituting a smaller model produced only 5–8% variance. These ablations are author-reported and should be treated as suggestive rather than confirmatory.

Tier 3 – Theoretical Scaling Arguments (Tentative). Formal results [9] suggest that increasing agent count without an optimized coordination topology yields diminishing returns and increased inter-agent interference. This is consistent with the CPH but does not directly test it in financial settings.

5.4 Why the CPH Cannot Yet Be Validated

Definitive validation requires a controlled experiment varying D2 while holding D1, D3, D4, and LLM backbone constant, evaluated on contamination-free data with rolling windows and net-of-cost returns. This experiment has not been conducted because the five evaluation failures make its prerequisites unavailable. The failures are not merely a general critique; they are the specific obstacle blocking the field's most important untested hypothesis. Addressing them is a prerequisite for testing the CPH, not a parallel concern.

To transition toward empirical validation, we propose a Cross-Architecture Factorial Design. By isolating *Coordination Logic* as the primary independent variable and *LLM Parameter Scale* as a control variable, researchers can quantify the Marginal Alpha contributed by the protocol. We suggest a benchmark of 500 simulated trading days across three distinct market regimes (Bull, Bear, and Sideways) to ensure the coordination advantage is robust against regime-specific model biases.

6 Coordination Trade-offs and the CBS Metric

If coordination protocol is the most consequential design dimension, understanding its costs and risks is essential for any practitioner acting on that hypothesis. This section synthesises four trade-off axes and introduces the CBS as the metric that operationalizes the CPH in deployment.

6.1 Key Trade-off Axes

Cost and performance. Inference costs scale linearly with agent count, while coordination costs scale quadratically in fully connected topologies. At current API pricing, a seven-agent system incurs roughly \$0.50–\$2.00 per daily decision, negligible for medium-frequency strategies but material at higher frequencies. Practical budgets of three to seven agents with two to three interaction rounds are consistent across the literature. Hybrid designs escalating selectively to frontier models for complex reasoning steps offer order-of-magnitude cost reductions.

Debate and latency. Each debate round introduces one to three seconds of latency; a two-round debate can incur five to twenty basis points of adverse

price movement, potentially exceeding the signal improvement it provides. This latency cost is frequency-dependent, with coordination benefits most pronounced at medium-frequency horizons where holding periods are long enough to absorb the delay. Debate depth should be calibrated to both market conditions and asset type: direct execution during high-conviction signals minimises latency cost, while liquid equities can tolerate deeper coordination more readily than illiquid or volatile assets.

Memory depth and regime drift. Historical precedents from a prior regime introduce anchoring bias when structural conditions change. Existing designs such as FinMem's decay-based memory and FinCon's episodic updating partially address this but rely on fixed temporal structures rather than event-driven adaptation. Explicit regime-change detection is needed to trigger belief revision; absent such mechanisms, memory-equipped agents should incorporate circuit breakers suspending retrieval under elevated drawdown or volatility.

Planner-executor depth. Ablation results from FinCon and TradingAgents indicate that removing independent risk assessment degrades risk-adjusted performance. A minimal three-stage pipeline (signal generation, risk gate, execution) appears sufficient for most production settings, with additional stages yielding diminishing returns unless they contribute distinct information.

6.2 The Coordination Breakeven Spread

The trade-off axes above share a common structure: coordination improves signal quality but incurs a cost, and no published metric determines whether the improvement justifies that cost for a given instrument. We formalize this via the **Coordination Breakeven Spread (CBS)**. Let $\Delta p(d)$ denote the expected improvement in entry/exit price from coordination depth d, and let s denote the bid–ask spread (round-trip cost = $2s$). Defining

$$\mathrm{CBS}(d) = \frac{\Delta p(d)}{2}$$

coordination is optimal only if $s < \mathrm{CBS}(d)$; when the instrument's spread exceeds the CBS, coordination overhead is not recovered and the system should revert to single-agent operation.

Under reasonable assumptions, the implied CBS lies in the low single-digit basis-point range for typical Sharpe levels, indicating that coordination is economically viable primarily in highly liquid instruments with narrow spreads. In practice, $\Delta p(d)$ can be estimated as the difference in volume-weighted average execution price between a coordinated and single-agent baseline over matched trading windows, net of latency-induced slippage; where unavailable, it can be bounded using coordination-driven Sharpe improvements converted to basis points via average holding period. Precise calibration beyond these approximations remains an open empirical task.

Relationship to the Evaluation Failures and the CPH. CBS directly addresses transaction cost neglect (Section 4.4) by converting coordination gains

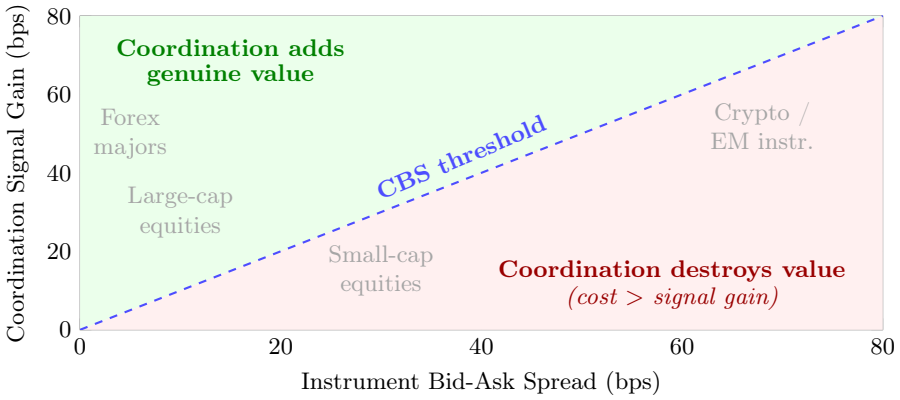

Fig. 2: Conceptual illustration of the CBS threshold. Instruments in the upper-left region (low spread, high coordination gain) are candidates for multi-agent coordination; those in the lower-right are not. Asset-class regions are directional and illustrative only; empirical CBS values cannot be computed from currently published results owing to transaction cost neglect (Section 4.4).

into a spread threshold. It is regime-dependent, as spreads widen during volatility spikes. In deployment, practitioners can test the CPH by running coordinated and single-agent systems in parallel and observing whether coordination consistently clears the CBS threshold.

Asset-Class and Regime Dependence. The CBS threshold varies substantially (Fig. 2). For large-capitalisation US equities (1–2 bps spreads), coordination-driven signal improvements may plausibly exceed costs at daily frequency. For small-capitalisation equities (10–50 bps), mid-capitalisation cryptocurrency (20–100 bps), or emerging market instruments, the threshold is considerably higher. A system applying coordination uniformly will over-coordinate during crisis periods when spreads are wide and under-coordinate during calm periods when coordination overhead is relatively more costly. We propose CBS as a standard reporting requirement alongside Sharpe ratio and maximum drawdown.

7 Conclusion and Future Directions

This survey has argued that five systematic evaluation failures make cross-system comparisons of LLM-based multi-agent financial systems unreliable, and that addressing them is a prerequisite for any credible claim about what drives performance. From the structural patterns that remain observable despite these failures, we formulated the CPH and introduced the CBS as its deployment metric. The logical chain is: the taxonomy provides vocabulary for controlled comparison; the evaluation critique explains why comparison is currently unreliable; the CPH identifies what is most worth testing; and the CBS defines

what testing it requires in practice. We identify three high-priority directions for future work.

Controlled validation of the CPH. The definitive experiment holds architecture topology, memory design, tool integration, and LLM backbone constant while varying only the coordination mechanism across identical contamination-free data with rolling-window, net-of-cost evaluation. A community benchmark providing this infrastructure would either confirm the CPH and redirect research effort, or reveal that coordination and model quality require joint optimization.

Small language model specialist architectures. No published system implements a production-ready hybrid in which small models handle routine subtasks (sentiment classification, compliance checking) and escalate to frontier models only for complex reasoning. Related work [1] suggests that fine-tuned small models can match frontier models on specialized tasks, with inference cost reductions of one to two orders of magnitude, though this finding derives from general agentic settings and its applicability to financial multi-agent systems remains an open question.

Systemic risk from correlated AI trading. If multiple institutions deploy similar LLM-based multi-agent architectures, correlated signals could amplify rather than dampen market volatility. Existing regulatory frameworks (IMF, CFTC, EU AI Act), calibrated for single-agent AI systems, do not account for emergent coordination effects across independent deployments.

The most consequential contributions will come not from adding agents or scaling models, but from designing coordination mechanisms that demonstrably improve risk-adjusted, net-of-cost, regime-robust decision quality, and from building the evaluation infrastructure needed to verify such claims.

Disclosure of Interests. The authors declare no competing interests. No external funding was received in support of this work.

References

1. Belcak, P., Molchanov, P., Dong, H., Muralidharan, S., et al.: Small language models are the future of agentic AI. arXiv preprint arXiv:2506.02153 (2025), nVIDIA
2. BlackRock: AlphaAgents: Multi-agent LLM for equity portfolios. arXiv preprint arXiv:2508.11152 (2025)
3. Chen, Y., et al.: StockBench: Can LLM agents trade stocks profitably in real-world markets? arXiv preprint arXiv:2510.02209 (2025)
4. Choi, J., Zhu, S., Li, T.: Debate or vote: Which yields better decisions in multi-agent large language models? In: Advances in Neural Information Processing Systems (NeurIPS) (2025), spotlight
5. Cook, J., et al.: Agentic RAG for fintech (2025), preprint
6. Ding, Y., Li, J., Wang, X., Chen, Y.: Large language model agent in financial trading: A survey. arXiv preprint arXiv:2408.06361 (2024)
7. Du, Y., et al.: Improving factuality and reasoning in language models with multiagent debate. In: Proceedings of the 41st International Conference on Machine Learning (ICML) (2024)

8. Elton, E.J., Gruber, M.J., Blake, C.R.: Survivorship bias and mutual fund performance. Review of Financial Studies **9**(4), 1097–1120 (1996)
9. Estornell, A., Liu, Y.: Multi-LLM debate: Framework, principals, and interventions. In: Advances in Neural Information Processing Systems (NeurIPS) (2024)
10. Fatemi, S., Hu, Y.: FinVision: A multi-agent framework for stock market prediction. In: Proceedings of the 5th ACM International Conference on AI in Finance (ICAIF) (2024), arXiv:2411.08899
11. Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N.V., Wiest, O., Zhang, X.: Large language model based multi-agents: A survey of progress and challenges. In: Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI) (2024), arXiv:2402.01680
12. Li, X.: Design and empirical study of a large language model-based multi-agent investment system for chinese public REITs. arXiv preprint arXiv:2602.00082 (2026)
13. Li, Y., et al.: TradingGPT: Multi-agent system with layered memory and distinct characters for enhanced financial trading performance (2023), preprint
14. Li, Y., Kim, B., Cucuringu, M., Ma, Y.: Can LLM-based financial investing strategies outperform the market in long run? arXiv preprint arXiv:2505.07078 (2025), kDD 2026 Datasets & Benchmarks Track
15. Li, Y., et al.: HedgeAgents: A balanced-aware multi-agent financial trading system. In: Companion Proceedings of the ACM Web Conference (WWW Companion) (2025)
16. Li, Y., et al.: QuantAgents: Towards multi-agent financial system via simulated trading. arXiv preprint arXiv:2510.04643 (2025)
17. Liang, Y., et al.: Degeneration-of-thought: Multi-agent debate can harm reasoning in large language models. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2024)
18. Qian, Y., et al.: When agents trade: Live multi-market trading benchmark for LLM agents. arXiv preprint arXiv:2510.11695 (2025)
19. Sun, Y., Huang, H., Pompili, D.: LLM-based multi-agent reinforcement learning: Current and future directions. arXiv preprint arXiv:2405.11106 (2024)
20. Talebirad, Y., Nadiri, A.: Multi-agent collaboration: Harnessing the power of intelligent LLM agents. arXiv preprint arXiv:2306.03314 (2023)
21. Wawer, M., Chudziak, B.: Integrating traditional technical analysis with AI: A multi-agent LLM-based approach to stock market forecasting. In: Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART). pp. 100–111 (2025)
22. Xiao, Y., et al.: TradingAgents: Multi-agents LLM financial trading framework. In: Proceedings of the 39th AAAI Conference on Artificial Intelligence (AAAI) (2025)
23. Xiong, Z., et al.: How memory management impacts LLM agents: An empirical study of experience-following behavior. arXiv preprint arXiv:2505.16067 (2025)
24. Yang, H., et al.: FinRobot: An open-source AI agent platform for financial applications using large language models. arXiv preprint arXiv:2405.14767 (2024)
25. Yu, W., et al.: FinCon: A synthesized LLM multi-agent system with conceptual verbal reinforcement for enhanced financial decision making. In: Advances in Neural Information Processing Systems (NeurIPS) (2024)
26. Yu, Y., et al.: FinMem: A performance-enhanced LLM trading agent with layered memory and character design. In: AAAI Spring Symposium (AAAI-SS) (2024)
27. Zhang, Y., et al.: A multimodal foundation agent for financial trading: Tool-augmented, diversified, and generalist. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) (2024)
28. Zhao, Y., et al.: ContestTrade: Competitive multi-agent trading (2025), preprint