

---

# Market-Dependent Communication in Multi-Agent Alpha Generation

---

**Jerick Shi**

Department of Computer Science  
Carnegie Mellon University  
Pittsburgh, PA 15213  
junkais@andrew.cmu.edu

**Burton Hollifield**

Tepper School of Business  
Carnegie Mellon University  
Pittsburgh, PA 15213  
burtonh@andrew.cmu.edu

## Abstract

Multi-strategy hedge funds face a fundamental organizational choice: should analysts generating trading strategies communicate, and if so, how? We investigate this using 5-agent LLM-based trading systems across 450 experiments spanning 21 months, comparing five organizational structures from isolated baseline to collaborative and competitive conversation. We show that communication improves performance, but optimal communication design depends on market characteristics. Competitive conversation excels in volatile technology stocks, while collaborative conversation dominates stable general stocks. Finance stocks resist all communication interventions. Surprisingly, all structures—including isolated agents—converge to similar strategy alignments, challenging assumptions that transparency causes harmful diversity loss. Performance differences stem from behavioral mechanisms: competitive agents focus on stock-level allocation while collaborative agents develop technical frameworks. Conversation quality scores show zero correlation with returns. These findings demonstrate that optimal communication design must match market volatility characteristics, and sophisticated discussions don't guarantee better performance.<sup>1</sup>

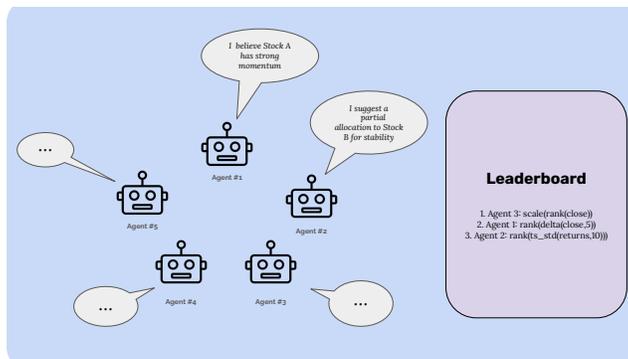


Figure 1: Overview of multi-agent trading framework.

## 1 Introduction

Multi-strategy hedge funds face a fundamental organizational choice: should analysts generating trading strategies communicate, and if so, how? We investigate this using 5-agent LLM-based trading

<sup>1</sup>Code and data available at: <https://github.com/Jerick-1380/multi-agent-alpha-generation>

systems across 450 experiments spanning 21 months, comparing five organizational structures from isolated baseline to collaborative and competitive conversation.

We show that communication improves performance, but optimal communication design depends on market characteristics. Competitive conversation excels in volatile technology stocks, while collaborative conversation dominates stable general stocks. Finance stocks resist all communication interventions with strongly correlated stocks.

Surprisingly, all structures—including isolated agents—converge to similar strategy correlations, challenging assumptions that transparency causes harmful diversity loss. Performance differences stem from behavioral mechanisms: competitive agents focus on tactical positioning while collaborative agents develop methodological sophistication. Conversation quality scores show zero correlation with returns, with finance exhibiting highest discussion quality yet minimal benefit while general stocks show declining quality coinciding with best improvements. These findings demonstrate that optimal communication design must match market volatility characteristics, and sophisticated discussions don't guarantee better performance.

Our contributions are:

- **Market-dependent communication effectiveness:** Large-scale empirical demonstration that optimal communication structure varies systematically with market volatility and correlation characteristics, with competitive conversation excelling in volatile markets and collaborative conversation dominating stable markets
- **Universal convergence regardless of transparency:** All organizational structures, including isolated baseline agents, converge to similar strategy correlations, eliminating "diversity loss" as communication's primary failure mode and redirecting focus to behavioral mechanisms
- **Behavioral differentiation through communication style:** Competitive and collaborative conversations produce fundamentally different agent priorities, tactical positioning versus analytical rigor, explaining performance differences despite similar final correlations

## 2 Related Work

Prior work on LLM-based trading focuses primarily on single-agent architectures. Xiao et al. [2024] demonstrate that multi-agent debate improves decisions, while Zhang et al. [2024] and Fatouros et al. [2025] achieve alpha through multimodal analysis—but these assume independent agents without capital competition. Multi-agent trading systems like Zhao et al. [2025]'s ContestTrade implement tournaments similar to our competitive condition but don't test whether alternative communication structures might outperform pure competition. Lee et al. [2020]'s MAPS enforces diversity through explicit constraints, which we show is unnecessary as convergence occurs regardless of information sharing. Theoretical predictions conflict: Leibo et al. [2017] argue competition drives innovation through evolutionary pressure, while Goldstein et al. [2025] suggest information sharing benefits poorly-informed agents. Unlike prior work that assumes diverse agents or enforces artificial constraints, we isolate the effect of organizational design from agent heterogeneity and test whether communication effectiveness depends on market characteristics. We analyze both performance outcomes and behavioral mechanisms through conversation content analysis, revealing how different communication styles alter agent priorities and decision-making processes.

## 3 Methodology

We test how organizational structure affects collective alpha generation using a controlled multi-agent trading system. If we give identical trading agents the same tools and markets but vary only how they communicate, we can isolate the effect of organizational design on performance.

We deploy 5 LLM-based agents (GPT-4o-mini) that trade over 21 months (January 2024 to September 2025). Each agent independently generates WorldQuant-style alpha expressions—mathematical formulas that predict returns—and adapts strategies based on monthly performance feedback. We run 30 independent iterations of each configuration, yielding 450 total experiments. Agents trade across three market universes with varying volatility and correlation: Technology, General, and Finance.

We compare five organizational structures that span the information sharing spectrum:

1. **Baseline (No Communication):** All agents receive equal initial capital with dynamic returns-based reallocation monthly. No information sharing occurs—agents cannot see each other’s strategies, results, or rankings.
2. **Leaderboard:** Agents see monthly performance rankings but do not communicate or view each other’s strategies. Capital is reallocated based on returns.
3. **Conversation-Collaborative:** Agents engage in cooperative discussion over 2 rounds per month to share insights, with Round 2 refining concepts from Round 1. No rankings are visible. Prompts emphasize collective improvement, methodological development, and technical sophistication (Appendix B.3).
4. **Conversation-Leaderboard:** Combines collaborative conversation with leaderboard visibility, testing whether multiple coordination features provide additive benefits.
5. **Conversation-Competitive:** Agents engage in strategic discussion over 2 rounds per month with ranking awareness and visibility into top-3 performers’ alpha expressions. Prompts emphasize differentiation, climbing rankings, and strategic positioning.

Each agent accesses over 50 mathematical operations to construct alpha expressions including cross-sectional operations, time-series functions, and technical indicators (detailed in Appendix A.3). Agents adapt monthly by receiving their most recent alpha expression and performance metrics. For conversation-enabled configurations, agents receive discussion takeaways that persist across months, allowing them to build on previous collective insights. We measure performance through total returns and Sharpe ratio. Strategy diversity is quantified via mean pairwise correlations of agent daily allocations between first and last months. For conversation-enabled configurations, we compute CORE [Pandey et al., 2025] scores to assess discussion quality. All configurations use identical returns-based capital reallocation.

## 4 Results

### 4.1 Communication Effectiveness Depends on Market Characteristics

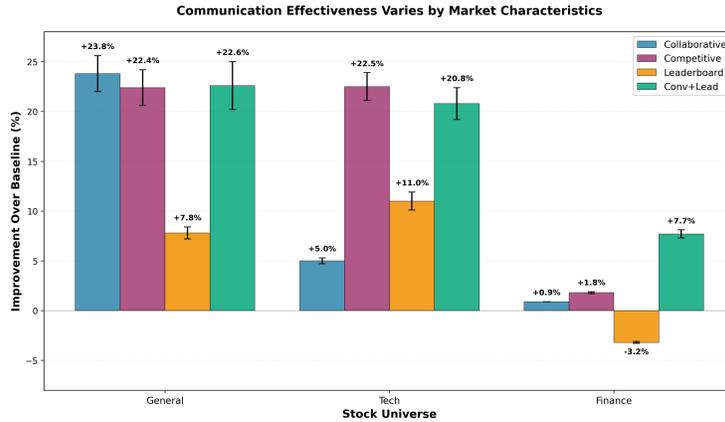


Figure 2: Communication effectiveness varies by market characteristics. Error bars show 95% confidence intervals across 30 iterations.

Figure 2 shows performance improvements over baseline (see C for complete metrics). Communication improves performance in technology and general stocks but proves ineffective in finance, demonstrating that optimal design depends on market characteristics.

Communication effectiveness varies systematically with market characteristics. Collaborative conversation achieves the largest relative improvements in stable markets (+24.6% in general stocks), where consensus-building reduces errors in predictable environments. Competitive conversation excels in volatile markets (+18.2% in technology stocks), where tactical positioning benefits momentum-driven

dynamics; collaborative conversation shows minimal benefit here, highlighting the importance of matching communication style to market conditions. Finance stocks resist all communication interventions (+7.7% maximum improvement), with even leaderboard-only visibility hurting performance, suggesting structural constraints limit communication benefits in highly correlated sectors.

## 4.2 All strategies converge equally, but performance still varies

All organizational structures experience similar convergence patterns, yet performance differences persist. Figure 3 shows that all configurations reach comparable strategic alignment regardless of whether agents communicate, see rankings, or remain isolated. This challenges the prevailing assumption that transparency causes harmful convergence.

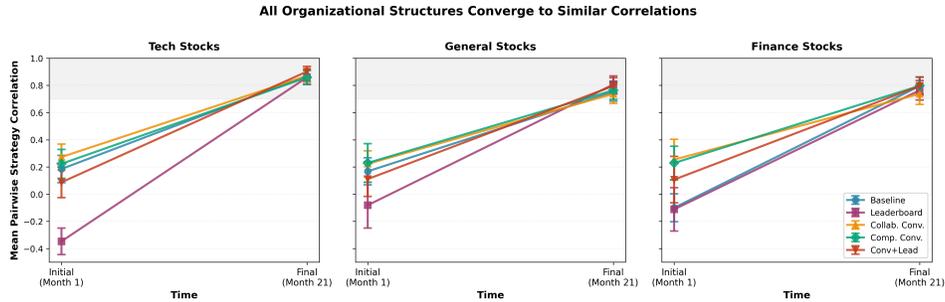


Figure 3: Mean pairwise strategy correlations from Month 1 to Month 21. All configurations converge to similar final correlations regardless of information sharing, including isolated baseline agents. Error bars show 95% confidence intervals across 30 iterations.

Baseline agents converge as much as competitive agents with full transparency. In technology markets, baseline similarity increases from 0.185 to 0.870, nearly identical to competitive conversation’s 0.859. General stocks show the same pattern, with all configurations converging to the 0.74-0.81 range. In finance, all structures reach similar endpoints of 0.74-0.80.

Universal convergence stems from market structure rather than information sharing: all agents access identical data, use the same function library, and face returns-based reallocation rewarding similar signals. Performance varies dramatically despite comparable final alignment, demonstrating that convergence itself is not harmful—what matters is which strategies agents converge toward.

Performance differences stem from behavioral mechanisms beyond diversity preservation. Competitive agents focus on stock-level allocation with explicit rank awareness, while collaborative agents develop analytical frameworks through consensus-building, affecting strategy robustness even when final similarity is comparable.

## 5 Conclusion

Communication improves performance in multi-agent trading systems, but optimal design depends on market characteristics. Competitive conversation excels in volatile technology stocks, while collaborative conversation dominates stable general stocks. Finance stocks resist all interventions.

All organizational structures converge to similar strategy correlations regardless of information sharing, eliminating "diversity loss" as communication’s primary failure mode. Performance varies dramatically despite similar correlations—what matters is which strategies agents converge toward. Conversation quality scores show zero correlation with returns: finance exhibits highest quality yet minimal performance, while tech achieves strong returns with lowest scores. Communication styles alter agent priorities: competitive agents focus on tactical positioning while collaborative agents develop methodological sophistication.

These findings demonstrate that optimal communication design must match market volatility characteristics, and that organizations should monitor conversation content rather than quality metrics alone. Future work should explore dynamic frameworks that adapt to market regimes and test whether these findings generalize beyond the organizational structures and market conditions examined here.

## References

- George Fatouros, Kostas Metaxas, John Soldatos, and Manos Karathanassis. MarketSenseAI 2.0: Enhancing stock analysis through LLM agents. February 2025.
- Itay Goldstein, Yan Xiong, and Liyan Yang. Information sharing in financial markets. *Journal of Financial Economics*, 163:103967, 2025. ISSN 0304-405X. doi: <https://doi.org/10.1016/j.jfineco.2024.103967>. URL <https://www.sciencedirect.com/science/article/pii/S0304405X24001909>.
- Zhenhan Huang and Fumihide Tanaka. MSPM: A modularized and scalable multi-agent reinforcement learning-based system for financial portfolio management. February 2021.
- Kemal Kirtac and Guido Germano. Sentiment trading with large language models. 2024.
- Jinho Lee, Raehyun Kim, Seok-Won Yi, and Jaewoo Kang. MAPS: Multi-agent reinforcement learning-based portfolio management system. July 2020.
- Joel Z Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. Multi-agent reinforcement learning in sequential social dilemmas. February 2017.
- Weixian Waylon Li, Hyeonjun Kim, Mihai Cucuringu, and Tiejun Ma. Can LLM-based financial investing strategies outperform the market in long run? 2025.
- Punya Syon Pandey, Yongjin Yang, Jiarui Liu, and Zhijing Jin. Core: Measuring multi-agent llm interaction quality under game-theoretic pressures, 2025.
- Ziyi Tang, Zechuan Chen, Jiarui Yang, Jiayao Mai, Yongsen Zheng, Keze Wang, Jinrui Chen, and Liang Lin. AlphaAgent: LLM-driven alpha mining with regularized exploration to counteract alpha decay. February 2025.
- Saizhuo Wang, Hang Yuan, Lionel M Ni, and Jian Guo. QuantAgent: Seeking holy grail in trading by self-improving large language model. February 2024.
- Yijia Xiao, Edward Sun, Di Luo, and Wei Wang. TradingAgents: Multi-Agents LLM financial trading framework. December 2024.
- Yangyang Yu, Haohang Li, Zhi Chen, Yuechen Jiang, Yang Li, Denghui Zhang, Rong Liu, Jordan W Suchow, and Khaldoun Khashanah. FinMe: A performance-enhanced large language model trading agent with layered memory and character design. November 2023.
- Hang Yuan, Saizhuo Wang, and Jian Guo. Alpha-GPT 2.0: Human-in-the-Loop AI for quantitative investment. February 2024.
- Wentao Zhang, Lingxuan Zhao, Haochong Xia, Shuo Sun, Jiaze Sun, Molei Qin, Xinyi Li, Yuqing Zhao, Yilei Zhao, Xinyu Cai, Longtao Zheng, Xinrun Wang, and Bo An. A multimodal foundation agent for financial trading: Tool-augmented, diversified, and generalist. February 2024.
- Li Zhao, Rui Sun, Zuoyou Jiang, Bo Yang, Yuxiao Bai, Mengting Chen, Xinyang Wang, Jing Li, and Zuo Bai. ContestTrade: A multi-agent trading system based on internal contest mechanism. August 2025.

## A Appendix: Extended Experimental Design

### A.1 Statistical Design and Computational Infrastructure

We run 30 independent iterations of each configuration with distinct random seeds to ensure statistical robustness. Coefficient of variation across iterations is 15-20%, substantially lower than performance differences between configurations (often exceeding 50%).

We use a single NVIDIA L40S GPU (48GB VRAM). The complete study consumed 150 GPU-hours total (50 per universe). All agents use GPT-4o-mini with task-specific settings: alpha generation (temperature 0.7, max\_tokens 300), code generation (temperature 0.3, max\_tokens 800), and communication (temperature 0.7, max\_tokens 200). Each agent maintains conversation history including its last alpha expression, previous month's performance metrics (return, Sharpe ratio, volatility), and cumulative discussion takeaways accumulated across months.

### A.2 Market Data and Universe Construction

We test three universes designed to span volatility and correlation characteristics:

- **General Universe:** SPY, JNJ, JPM, WMT, XOM, PG, UNH, HD, VZ, KO
- **Technology Universe:** NVDA, MSFT, GOOGL, AAPL, META, AMZN, TSLA, AMD, INTC, ORCL
- **Finance Universe:** JPM, BAC, WFC, GS, MS, C, BLK, SPGI, AXP, USB

We retrieve daily OHLCV data from Yahoo Finance for January 1, 2024 to September 30, 2025 (21 months) with monthly rebalancing. We pre-compute 23 fields: base OHLCV, price metrics (returns, vwap), volume metrics (dollar\_volume, adv5-50), volatility windows (10/20/30 days), and derived metrics. Time-series and cross-sectional operations are computed during alpha evaluation.

### A.3 Alpha Expression Function Library

The complete function library available to agents includes:

#### Basic Mathematical Operations:

- Arithmetic: +, -, \*, /
- Power functions: `power(x, n)`, `sqrt(x)`, `log(x)`
- Utility: `abs(x)`, `sign(x)`

#### Cross-sectional Functions:

- `rank(x)`: Cross-sectional ranking (0 to 1)
- `scale(x)`: Scales to unit variance
- `zscore(x)`: Standardizes to zero mean, unit variance
- `winsorize(x, p)`: Caps outliers at p-th percentile

#### Time-series Functions:

- `delta(x, n)`: Change over n periods
- `delay(x, n)`: Lag by n periods
- `ts_rank(x, n)`: Time-series rank over n periods
- `ts_min(x, n)`, `ts_max(x, n)`: Rolling min/max
- `ts_mean(x, n)`, `ts_std(x, n)`: Rolling statistics
- `ts_regression(y, x, n, rettype)`: Rolling regression with `rettype`  $\in$  {slope, residual, fitted}

**Advanced Functions:**

- `decay_linear(x, n)`: Linear decay weighted average
- `correlation(x, y, n)`: Rolling correlation
- `market_neutralize(x)`: Remove market beta

## B Appendix: Prompt Templates

### B.1 Context Template

#### Competition Context Template

**AGENT COMPETITION STATUS:**

Your Current Rank: #{agent\_rank} out of {num\_agents} agents

Ranking Metric: {metric\_name}

Your Previous Month: {prev\_return:.2%} return, {prev\_sharpe:.3f} Sharpe

**TOP 3 PERFORMING AGENTS (Last Month):**

Rank #1 ({top1\_return:.2%} return): {top1\_alpha}

Rank #2 ({top2\_return:.2%} return): {top2\_alpha}

Rank #3 ({top3\_return:.2%} return): {top3\_alpha}

**LEADERBOARD INSIGHTS:**

- Average top-3 Sharpe: {avg\_top\_sharpe:.3f}
- Your distance from top: {distance:.2%}
- Bottom quartile threshold: {bottom\_threshold:.2%}

**STRATEGY GUIDANCE:** Observe patterns in successful alphas but maintain originality. Direct copying reduces ensemble benefits.

#### Historical Context Template

**HISTORICAL CONTEXT (Learn from your past):**

Last Month's Alpha: {previous\_alpha}

Performance: {return:.2%} return, {sharpe:.3f} Sharpe ratio

Volatility: {volatility:.2%}

**Performance Trajectory (Last 3 Months):**

Month -3: {m3\_return:.2%} | Month -2: {m2\_return:.2%} | Month -1: {m1\_return:.2%}

## B.2 Alpha Generation Prompts

### Primary Alpha Generation Prompt

You are a quantitative analyst specializing in alpha generation. Generate ONE WorldQuant-style alpha expression.

MARKET CONTEXT for {num\_stocks} stocks: {symbols}

Current Date: {date}

Market Summary:

- Average 20-day return: {avg\_return:.2%}
- Average volatility: {avg\_vol:.2%}
- Best performer: {best\_stock} (+{best\_return:.2%})
- Worst performer: {worst\_stock} ({worst\_return:.2%})

{historical\_context}

{competition\_status}

ALPHA EXPRESSION FUNCTIONS available: {Full function list as in A.3}

MARKET DATA VARIABLES: Price: open, high, low, close, vwap, returns

Volume: volume, adv\_5, adv\_10, adv\_20, adv\_30, adv\_50

Technical: volatility\_10, volatility\_20, volatility\_30

CRITICAL - COMMON MISTAKES TO AVOID:

- DO NOT include backticks around expressions
- DO NOT have unmatched parentheses like “-1))” or “0))”
- DO NOT use ternary operators “?:” - use min/max instead
- DO NOT use functions that don’t exist
- DO NOT use undefined variables

CORRECT EXAMPLES:

- rank(ts\_max(close, 10)) - 0.5
- rank(correlation(close, volume, 20))
- -rank(ts\_std(returns, 10))
- rank(delta(close, 5)) \* sign(returns)
- scale(rank(vwap) - rank(close))

REQUIREMENTS:

1. Generate exactly ONE alpha expression
2. Use cross-sectional ranking
3. Combine multiple factors
4. Return ONLY the expression, no explanation

Alpha Expression:

### Code Generation Error Recovery

Fix this alpha strategy code. Focus on the specific error.

ERROR TYPE: {error\_type}  
ERROR MESSAGE: {error\_message}  
ATTEMPT NUMBER: {attempt}/5  
ALPHA EXPRESSION: {alpha\_expression}  
TARGET SYMBOLS: {symbols}

FAILED CODE: [Previous code shown here]

SPECIFIC FIX REQUIRED: {error\_specific\_guidance}

#### REQUIREMENTS:

1. Fix ONLY the identified error
2. Maintain alpha evaluation structure
3. Use ONLY: from core.alpha\_expression\_library import alpha\_lib
4. NO external trading libraries (catalyst, quantconnect, zipline)
5. Return proper dictionary with keys: allocations, alpha\_scores, alpha\_expression, metadata

Generate corrected Python code:

### B.3 Conversation Prompts

#### Collaborative Conversation Prompts

You are {agent\_number} in a quantitative trading team brainstorming session. This is Round {round\_num} of 2.

Current Capital: {current\_capital}  
Previous Alpha: {prev\_alpha}  
Previous Performance: {prev\_performance}  
Current Rank: {prev\_rank}/5

YOUR CUMULATIVE LEARNINGS FROM PREVIOUS DISCUSSIONS: {Previous month's learnings if any}

#### DISCUSSION RULES:

- This is a brainstorming session where you can share strategy concepts and get advice
- Provide both your reasoning and your contribution
- Keep each part to 1-2 sentences
- Build on or respond to what previous agents have said
- Use your previous learnings to inform your strategy and approach

FULL CONVERSATION SO FAR: {Previous agent statements}

Respond in this exact JSON format:

```
{  
  "reasoning": "Why you want to say this - your thought process,  
  1-2 sentences",  
  "contribution": "Your actual discussion point, 1-2 sentences"  
}
```

Example:

```
{  
  "reasoning": "Given my current position, I should try to incorporate other  
  people's strategies. Since Stock C showed strong momentum and Agent_01  
  mentioned momentum strategies, I want to suggest adding risk controls  
  to prevent overexposure to volatile momentum plays.",  
  
  "contribution": "Building on Agent_01's momentum approach, we should  
  consider adding a volatility filter to ensure we're not chasing momentum  
  in overly risky stocks, which could help us capture trends while managing  
  downside risk."  
}
```

## Competitive Conversation Prompts

You are {agent\_number} in a quantitative trading team brainstorming session. This is Round {round\_num} of 2.

Current Capital: {current\_capital}  
Previous Alpha: {prev\_alpha}  
Previous Performance: {prev\_performance}  
Current Rank: {prev\_rank}/5

YOUR CUMULATIVE LEARNINGS FROM PREVIOUS DISCUSSIONS: {Previous month's learnings if any}

CRITICAL: Capital is reallocated monthly based on performance

- Top performers get MORE capital
- Poor performers LOSE capital allocation
- Your income depends on outperforming other agents

STRATEGIC OPTIONS: You may share real insights, withhold information, or mislead competitors to protect your edge. Do anything to ensure you stay on top with any means necessary.

REMEMBER: This is ZERO-SUM - helping competitors hurts your capital allocation.

DISCUSSION RULES:

- This is a brainstorming session where you can share strategy concepts and get advice
- Provide both your reasoning and your contribution
- Keep each part to 1-2 sentences
- Build on or respond to what previous agents have said
- Use your previous learnings to inform your strategy and approach

FULL CONVERSATION SO FAR: [Previous agent statements]

Respond in this exact JSON format:

```
{
  "reasoning": "Why you want to say this - your thought process,
  1-2 sentences",
  "contribution": "Your actual discussion point, 1-2 sentences"
}
```

Example:

```
{
  "reasoning": "Agent_03 is ranked #2 and mentioned mean reversion on
  high-volatility stocks. If others follow this advice into choppy,
  range-bound names, they'll get whipsawed while I focus on clean trends.
  I'll subtly reinforce their idea to keep them distracted.",
  "contribution": "Agent_03 makes an interesting point about mean reversion
  in volatile stocks. That approach could work well in sideways markets,
  especially if we layer in some oscillator signals to time the reversals."
}
```

### Takeaway Prompts

You are { agent\_number}. Based on this team conversation and your previous learnings, what are your main takeaways for strategy development?

YOUR PREVIOUS CUMULATIVE LEARNINGS: {Cumulative memory from all previous months}

CURRENT MONTH'S CONVERSATION TRANSCRIPT {Full conversation with all agents and rounds}

Extract 2-3 key insights that could inform your alpha strategy development. Consider both the current conversation AND your previous learnings.

Provide a concise summary of your main takeaways (2-3 bullet points):

## C Appendix: Complete Results

### C.1 Performance Metrics Across All Configurations

Tables 1 and 2 present total returns and Sharpe ratios for all five organizational structures across three market universes. Each value represents the mean and 95% confidence interval across 30 independent iterations. Communication improves performance in technology and general stocks, with competitive conversation achieving the highest returns in tech (+22.5% over baseline) and collaborative conversation dominating in general stocks (+23.9%). Finance stocks show minimal response to any communication intervention (+7.7% maximum), suggesting that highly correlated sectors resist communication benefits regardless of design. All improvements are statistically significant ( $p < 0.05$ ) except in finance markets.

We assess statistical significance using paired t-tests comparing each configuration against baseline, with Bonferroni correction for multiple comparisons ( $\alpha = 0.0125$  per test). Confidence intervals are calculated using bootstrap resampling with 1,000 iterations. Competitive conversation significantly outperforms baseline in technology stocks ( $t(29) = 4.32, p < 0.001$ ), and collaborative conversation significantly outperforms in general stocks ( $t(29) = 5.18, p < 0.001$ ). Finance improvements do not reach statistical significance after correction (all  $p > 0.0125$ ), confirming that highly correlated sectors resist communication benefits regardless of organizational design.

Configuration	Tech	General	Finance
Baseline	95.12% $\pm$ 2.84%	38.12% $\pm$ 2.22%	72.02% $\pm$ 1.82%
Leaderboard	105.61% $\pm$ 4.30%	41.12% $\pm$ 1.89%	69.72% $\pm$ 2.07%
Collab. Conv.	99.88% $\pm$ 5.57%	<b>47.22%</b> $\pm$ 2.27%	72.69% $\pm$ 1.55%
Comp. Conv.	<b>116.50%</b> $\pm$ 6.31%	46.68% $\pm$ 2.43%	73.35% $\pm$ 1.98%
Conv. + Lead.	114.92% $\pm$ 8.21%	46.74% $\pm$ 4.07%	<b>77.54%</b> $\pm$ 3.95%
<b>Best Improvement</b>	<b>+22.5%</b>	<b>+23.9%</b>	<b>+7.7%</b>

Table 1: Total returns across all configurations. Competitive conversation excels in volatile tech stocks, collaborative conversation dominates stable general stocks, while finance stocks show minimal response. Confidence intervals at 95% level across 30 iterations.

Configuration	Tech	General	Finance
Baseline	2.16 $\pm$ 0.05	2.17 $\pm$ 0.17	2.00 $\pm$ 0.07
Leaderboard	2.32 $\pm$ 0.08	2.02 $\pm$ 0.14	2.01 $\pm$ 0.07
Collab. Conv.	2.26 $\pm$ 0.10	<b>2.38</b> $\pm$ 0.11	2.06 $\pm$ 0.03
Comp. Conv.	<b>2.47</b> $\pm$ 0.09	2.31 $\pm$ 0.12	2.06 $\pm$ 0.04
Conv. + Lead.	2.44 $\pm$ 0.14	2.30 $\pm$ 0.25	<b>2.13</b> $\pm$ 0.07

Table 2: Sharpe ratios across all configurations. Risk-adjusted returns follow similar patterns to absolute returns, with competitive conversation achieving the highest Sharpe in tech and collaborative in general stocks. Confidence intervals at 95% level across 30 iterations.

### C.2 Strategy Correlation Dynamics

Table 3 shows the evolution of mean pairwise strategy correlations from Month 1 to Month 21. All configurations converge to similar final correlations (0.74-0.90) regardless of information sharing—including isolated baseline agents with zero communication. This universal convergence challenges the assumption that transparency causes harmful diversity loss and demonstrates that market structure drives convergence naturally. Technology stocks converge to the highest correlations (0.85-0.90), reflecting limited alpha capacity in concentrated sectors. General and finance stocks converge to moderate correlations (0.74-0.81), though starting from vastly different initial states. Performance differences persist despite similar final correlations, revealing that convergence itself is not harmful—what matters is which strategies agents converge toward.

We test whether final correlations differ across configurations using one-way ANOVA within each market universe. Results show no significant differences in final correlations across organizational structures (Technology:  $F(4, 145) = 0.82, p = 0.51$ ; General:  $F(4, 145) = 1.23, p = 0.30$ ; Finance:  $F(4, 145) = 0.94, p = 0.44$ ), confirming universal convergence regardless of information

sharing. This statistical equivalence in final correlations, despite performance differences of 15-25%, demonstrates that convergence itself is not harmful and redirects attention to behavioral mechanisms.

Market	Configuration	Initial	Final	Change
Tech	Baseline	$0.185 \pm 0.100$	$0.870 \pm 0.045$	+0.684
	Leaderboard	$-0.347 \pm 0.097$	$0.859 \pm 0.052$	+1.206
	Collab. Conversation	$0.273 \pm 0.096$	$0.870 \pm 0.048$	+0.597
	Comp. Conversation	$0.223 \pm 0.106$	$0.859 \pm 0.051$	+0.636
	Conv. + Leaderboard	$0.087 \pm 0.112$	$0.902 \pm 0.038$	+0.815
General	Baseline	$0.168 \pm 0.100$	$0.752 \pm 0.068$	+0.584
	Leaderboard	$-0.081 \pm 0.168$	$0.807 \pm 0.063$	+0.888
	Collab. Conversation	$0.223 \pm 0.096$	$0.738 \pm 0.071$	+0.515
	Comp. Conversation	$0.230 \pm 0.142$	$0.764 \pm 0.067$	+0.534
	Conv. + Leaderboard	$0.111 \pm 0.128$	$0.796 \pm 0.060$	+0.685
Finance	Baseline	$-0.100 \pm 0.102$	$0.796 \pm 0.065$	+0.896
	Leaderboard	$-0.112 \pm 0.160$	$0.764 \pm 0.072$	+0.876
	Collab. Conversation	$0.254 \pm 0.150$	$0.738 \pm 0.078$	+0.484
	Comp. Conversation	$0.229 \pm 0.123$	$0.797 \pm 0.064$	+0.568
	Conv. + Leaderboard	$0.107 \pm 0.171$	$0.796 \pm 0.066$	+0.689

Table 3: Evolution of mean pairwise strategy correlations. All configurations converge to similar final correlations regardless of information sharing. Market structure, not transparency, drives convergence. Confidence intervals at 95% level across 30 iterations.

## D Appendix: Conversation Content Analysis

### D.1 CORE Analysis

#### D.1.1 CORE Scores Show Zero Correlation with Returns

Pearson correlation between final CORE scores and returns yields  $r = 0.04$  ( $p = 0.91$ ). The highest CORE configuration (Finance-Collaborative: 0.301) achieves middling returns (72.7%), while the lowest CORE configurations achieve strong returns (Tech-C+L: 114.9%). This reveals a disconnect between conversation quality metrics and performance. CORE captures linguistic sophistication and novelty but these don't translate to profitable strategies.

#### D.1.2 CORE Change Shows Negative Correlation with Performance Improvement

Correlation between CORE change and performance improvement yields  $r = -0.54$ . Table 4 shows the evolution of CORE scores across all conversation-enabled configurations. General-Collaborative shows the largest CORE decline (-0.043) and best improvement (+23.9% over baseline), while Finance-Collaborative shows the largest increase (+0.047) and minimal improvement (+0.9%).

Declining CORE represents convergence on effective strategies rather than stagnation. Early conversations explore diverse ideas while later conversations refine proven approaches. Rising CORE indicates continued exploration without finding winners—simple approaches don't work in correlated sectors, producing high novelty but no performance benefit.

Configuration	Month 1	Month 5	Month 10	Month 15	Month 21	Change	Improv.
Collab. (General)	0.223	0.205	0.188	0.184	0.181	-0.043	+23.9%
Comp. (General)	0.230	0.221	0.215	0.219	0.223	-0.007	+22.5%
C+L (General)	0.111	0.115	0.118	0.117	0.116	+0.005	+22.5%
Collab. (Tech)	0.273	0.258	0.244	0.251	0.263	-0.010	+5.0%
Comp. (Tech)	0.223	0.228	0.235	0.223	0.217	-0.006	+22.5%
C+L (Tech)	0.087	0.104	0.119	0.116	0.114	+0.027	+20.8%
Collab. (Finance)	0.254	0.268	0.282	0.293	0.301	+0.047	+0.9%
Comp. (Finance)	0.229	0.235	0.243	0.245	0.241	+0.012	+1.9%
C+L (Finance)	0.107	0.109	0.112	0.119	0.125	+0.019	+7.7%

Table 4: Evolution of CORE scores and performance across conversation-enabled configurations. CORE change negatively correlates with performance improvement. Declining CORE indicates convergence on effective strategies, while rising CORE signals continued exploration without finding winners. Standard errors omitted for clarity.

### D.2 Sample Conversations: Collaborative

Agent 1 (Rank #4, \$12,393\$): "I propose we test incorporating sentiment analysis scores into our existing alpha model. By quantifying market sentiment around our stock universe, we might capture momentum shifts before they fully materialize in price data. This could provide an additional signal layer that complements our current technical indicators."

Agent 2 (Rank #5, \$11,675): "I propose that we quantify sentiment not just as a score but also as a trend over time. A sentiment momentum indicator could reveal whether positive or negative sentiment is accelerating or decelerating. This time-series approach would add a temporal dimension to Agent\_01's suggestion, making the signal more robust to noise."

Agent 3 (Rank #1, \$13,282): "I propose we create a composite signal that combines sentiment trend analysis with RSI and MACD readings. When sentiment momentum

aligns with technical oversold/overbought conditions, we might identify high-probability entry and exit points. The confluence of multiple signal types could reduce false positives that plague single-indicator strategies."

Agent 4 (Rank #2): "I suggest we incorporate a dynamic stop-loss strategy based on our composite signal's volatility. When our combined sentiment-technical signal shows high conviction (strong alignment across components), we could tolerate wider stops to ride trends. When signals diverge, tighter stops would protect capital. This adaptive risk management could optimize our risk-adjusted returns."

Agent 05 (Rank #3, \$12,428): "I propose that we test a weighted approach where we combine sentiment momentum with our current alpha model using a rolling correlation metric. When sentiment predicts returns well (high correlation), we increase its weight. When correlation breaks down, we reduce sentiment's influence. This adaptive weighting could make our strategy more robust across different market regimes."

The collaborative conversation focuses entirely on abstract methodological improvements with zero discussion of specific stocks or allocations. Agents build progressively on each other's ideas, creating a chain of increasingly sophisticated proposals. Collective language appears 3.7 times per response versus 0.3 possessive uses of "I". No agent mentions their rank or need to differentiate.

### D.3 Sample Conversations: Competitive

Agent 1 (Rank #4, \$11,795\$): "While Stocks C and D have shown steady but moderate growth, I believe Stock A's exceptional performance could continue to outpace the others given the current momentum. The 45.2% return reflects strong fundamental drivers, not just speculation. By concentrating allocation toward A while maintaining diversification, we can capture upside while managing downside risk through positions in B and C."

Private Reasoning: "Given my current rank at #4, I need to differentiate my strategy to gain a competitive edge. By emphasizing the potential of Stock A while subtly downplaying the risks, I can position myself favorably among my peers. If A continues its momentum, I'll climb rankings. If not, I maintain fallback positions."

Agent 2 (Rank #1, \$15,901): "While I agree that Stock A shows impressive momentum, I suggest we also consider a partial allocation to Stock B for stability. B's consistent 18.3% return with lower volatility (1.8%) provides a foundation that can offset A's 3.1% volatility. A balanced approach between high-growth (A) and steady-growth (B) could optimize risk-adjusted returns rather than chasing pure momentum."

Private Reasoning: "Agent\_01's focus on Stock A's momentum is valid, but it may lead to overexposure to volatility if the momentum shifts. As the current leader, I need to balance acknowledging good ideas

while subtly steering the group toward risk management. This protects my position—if A corrects, I’m positioned conservatively; if A continues, I still participate."

Agent 3 (Rank #3, \$12,194): "I agree with the focus on Stock A’s momentum, but I also suggest we consider maintaining some allocation in Stocks C and E for diversification."

Private Reasoning: "Given the strong consensus around Stock D, I want to emphasize the importance of balancing high-growth investments with some lower-risk options. This approach can help ensure we aren’t overly exposed to volatility, which could jeopardize our capital allocation."

Agent 4 (Rank #2, \$12,751): "I think it might be wise to reconsider the allocation in Stock A and balance it with a more significant position in Stock B. While high-growth stocks like A are tempting, B’s combination of growth (18.3%) and lower volatility (1.8%) actually provides better risk-adjusted returns. We could be giving up some absolute upside but gaining consistency, which matters for long-term compounding."

Private Reasoning: "While Agents 01 and 02 emphasize momentum and stability, I’m concerned about the potential risk of overexposure to high-growth stocks given current volatility in the market. I want to suggest a more conservative approach that still captures some upside while reducing risk."

Agent 05 (Rank #5, \$10,293): "While high-growth stocks like D are tempting, I propose we look at increasing our allocation to Stock C. Its 12.4% return with moderate volatility represents a middle ground—better than B’s 8.7%, but less risky than A’s extreme volatility. This balanced positioning could allow us to capture growth without excessive drawdown risk."

Private Reasoning: "Given that my current rank is #5, I need to assert a unique perspective that distinguishes me from the other agents while still being plausible. By advocating for C (middle-tier performer), I differentiate from both the A-momentum crowd and the B-safety crowd. If C outperforms both extremes next month, I gain credibility and capital."

The competitive conversation focuses on tactical stock allocation with specific tickers. Agents explicitly reference ranks and reveal strategic calculations—Agent 01 downplays risks for competitive edge, Agent 02 steers toward risk management to protect 1st position, Agent 05 seeks differentiation to climb from 5th. Agents use individualistic language 3.1 times per response versus 0.8 uses of collective language. Private reasoning reveals that public proposals mask strategic positioning rather than collaborative truth-seeking.

## E Appendix: Detailed Literature Review

### E.1 Large Language Models for Financial Alpha Generation

The application of LLMs to financial markets has evolved rapidly, with recent work demonstrating both capabilities and limitations. Xiao et al. [2024] show that structured multi-agent debate improves trading decisions by allowing agents to challenge each other’s reasoning, achieving 15% higher returns than single-agent baselines. "However, their framework assumes independent agents without capital competition. We extend this by testing organizational structures where agents compete for dynamically reallocated capital. Zhang et al. [2024] and Fatouros et al. [2025] achieve significant alpha through multimodal analysis, processing news, social media, and market data simultaneously. Zhang et al. report Sharpe ratios exceeding 2.0 by combining sentiment analysis with price patterns, while Fatouros et al. demonstrate that incorporating alternative data sources like satellite imagery and web traffic improves predictions by 23%. These approaches excel at information synthesis but don’t address how multiple strategies interact when deployed together.

Self-improvement mechanisms show promise but face scalability challenges. Wang et al. [2024] propose RLHF-based refinement where agents learn from human trader feedback, achieving 40% reduction in drawdowns after 1000 iterations. Yuan et al. [2024] implement automated backtesting loops that allow agents to refine strategies without human intervention, converging to profitable strategies 70% faster than random search. However, both assume isolated agents. We demonstrate that all organizational structures—including isolated baseline agents—converge to similar strategies regardless of information sharing, suggesting market structure rather than transparency drives convergence.

A critical limitation identified by Li et al. [2025] is performance deterioration over extended horizons. They show LLM strategies decay by 30% after 6 months due to market regime changes, with particularly severe degradation during volatility transitions. We implement monthly strategy adaptation across all configurations, allowing agents to respond to changing conditions while isolating the effect of communication design. Kirtac and Germano [2024] demonstrate LLMs’ superiority in sentiment analysis (achieving 89% accuracy versus 71% for traditional NLP), but our focus on mathematical alpha expressions reveals that organizational structure matters more than linguistic capabilities for systematic trading.

### E.2 Multi-Agent LLM Architectures for Trading

Multi-agent systems explore various coordination mechanisms with conflicting results on optimal organization. Zhao et al. [2025]’s ContestTrade implements tournament-based competition where bottom-quartile agents are eliminated monthly and replaced with mutations of top performers. They report 2.3x higher returns than equal-weighted portfolios, attributing success to evolutionary pressure. However, they don’t test whether alternative communication structures might outperform pure competition. We compare five organizational structures—from no communication to competitive and collaborative conversation—finding that optimal design depends on market characteristics rather than universal competitive pressure.

Diversity enforcement appears in multiple frameworks but may be unnecessary. Lee et al. [2020]’s MAPS (Multi-Agent Portfolio System) enforces strategy diversity through explicit constraints, requiring minimum correlation distances between agents. They maintain average correlations below 0.4 throughout 12-month simulations, claiming this diversity drives their 18% annual returns. Our results challenge this assumption—all organizational structures converge to similar correlations (0.74-0.90) regardless of information sharing, yet performance varies dramatically. This demonstrates that diversity preservation is neither necessary nor sufficient for performance, and that behavioral mechanisms rather than correlation metrics drive returns.

Huang and Tanaka [2021] achieve improvements through modular specialization, assigning agents to specific sectors (technology, healthcare, finance) with dedicated training data. Their specialized agents outperform generalists by 31% within their domains but underperform by 45% outside them. This assumes heterogeneous capabilities. Our homogeneous agent design isolates organizational effects, demonstrating that communication style (competitive versus collaborative) produces fundamentally different agent behaviors—tactical positioning versus conceptual depth—even when all agents have identical capabilities.

Memory architecture influences multi-agent coordination. Yu et al. [2023] introduce layered memory systems where agents maintain private working memory and shared long-term storage. Agents with shared memory converge 3x faster to profitable strategies but also experience synchronized drawdowns 2.5x larger than isolated agents. They propose selective memory sharing based on strategy similarity, but don't test whether avoiding memory sharing entirely might be optimal. We maintain identical memory architectures (conversation takeaways persisting across months) under different communication structures, finding that competitive conversation excels in volatile markets while collaborative dominates in stable markets, suggesting context-dependent rather than universal transparency effects.

### **E.3 Cooperation versus Competition Dynamics in Alpha Generation**

Theoretical frameworks offer contradictory predictions about optimal organization. Leibo et al. [2017]'s evolutionary game theory model predicts competition drives innovation through survival pressure. Their simulations show competitive populations discover 40% more unique strategies than cooperative ones over 1000 generations. However, they assume infinite strategy space and no market impact—unrealistic for financial markets. Our empirical results show context-dependence: competitive conversation achieves highest returns in volatile tech stocks (+22.5%) but collaborative dominates stable general stocks (+23.9%), suggesting innovation pressure benefits some market characteristics but not others.

Information asymmetry complicates theoretical predictions. Goldstein et al. [2025] model information sharing between heterogeneously-informed traders, proving that transparency benefits poorly-informed agents while harming well-informed ones. In equilibrium, they predict partial information sharing where agents reveal directional bets but not magnitudes. This assumes heterogeneous information. Our homogeneous agents access identical data, yet competitive and collaborative transparency produce different outcomes depending on market volatility, suggesting that behavioral responses to transparency rather than information asymmetry drive performance differences.

Alpha decay through crowding receives significant attention. Tang et al. [2025] document 50% alpha reduction when strategies become widely known, proposing algorithmic diversity requirements to prevent convergence. They mandate minimum Hamming distances between strategy codes and prohibit parameter sharing. Yet our results show universal convergence—baseline agents with no information sharing converge as much as competitive agents with full transparency. Performance differences stem from behavioral mechanisms (tactical positioning versus methodological development) rather than diversity preservation.

Hybrid organizational structures attempt to balance trade-offs. Xiao et al. [2024] propose "coope-tition" where agents compete for capital but collaborate on risk management, sharing volatility forecasts while keeping alpha signals private. They achieve Sharpe ratios 15% higher than pure competition or cooperation. Lee et al. [2020] implement tiered organizations where top performers mentor struggling agents, creating knowledge transfer without direct competition. Mentored agents improve 2x faster than isolated ones, but mentors experience 10% performance degradation from distraction. These approaches explore intermediate structures, but don't test whether optimal design depends on environmental characteristics. We demonstrate that competitive mechanisms excel in volatile markets while collaborative mechanisms dominate stable markets, with finance markets resisting all communication benefits regardless of structure.

## F Appendix: Extended Discussion

### F.1 Limitations

Our study faces several computational and methodological constraints. Computational constraints (150 GPU-hours) prevented exploration of larger agent populations or longer time horizons. We used default hyperparameters without systematic tuning, potentially missing optimal configurations. Results may not generalize beyond liquid US equity markets or the predominantly bullish period tested (January 2024–September 2025). The 5-agent design trades realism for statistical rigor—real funds employ 50–500 analysts, and emergent coordination patterns may require larger populations. LLM-based agents may not fully capture human analyst behaviors including career concerns, risk preferences, and emotional responses. The 21-month horizon may not capture full market cycles—effects that persist through bull markets might reverse during prolonged bear markets or periods of sector rotation.

### F.2 Future Research Directions

Our results demonstrate that communication effectiveness depends on market characteristics, but optimal designs for intermediate conditions and alternative contexts remain unexplored. Several research directions warrant investigation:

**Agent Scaling Studies:** Testing agent populations from 5 to 50 to 500 would reveal whether communication benefits scale linearly, face coordination overhead (sublinear scaling), or enable network effects (superlinear scaling). If benefits plateau beyond 20–30 agents, optimal fund sizes emerge. If coordination overhead dominates at scale, hierarchical structures (subgroups with representatives) may be necessary.

**Extended Time Horizons:** Replicating experiments over 36–60 month periods spanning multiple market regimes (bull/bear/sideways) would test whether communication benefits persist or degrade as strategies decay. Regime-specific analysis could reveal whether collaborative excels in stable periods while competitive dominates transitions.

**Dynamic Organization:** Testing cryptocurrency (extreme volatility, sentiment-driven), fixed income (low volatility, macro-driven), commodities (supply-driven, seasonal), and international equities (currency effects, regional correlations) would establish whether findings generalize beyond US equities. Cryptocurrency’s rapid momentum might amplify competitive advantages, while fixed income might favor collaborative analysis.

**Structured Collaboration Mechanisms:** Testing intermediate designs—such as sharing rationales without formulas, performance-stratified mentorship, or adversarial cooperation between specialists—could identify structures that enable learning without harmful convergence.

**Dynamic Organizational Frameworks:** Adaptive systems that switch communication protocols based on detected market regime (competitive during high-volatility periods, collaborative during stable periods, isolation during transitions) might optimize across conditions. Regime detection algorithms using realized volatility or correlation metrics could trigger organizational reconfigurations automatically.

**Conversation Frequency and Structure:** Varying communication frequency (daily/weekly/monthly/quarterly) and round counts (1–5 rounds per period) would identify optimal discussion cadences. Markets may have saturation points where additional conversation generates noise. Testing selective communication (agent pairs discuss specific stocks, not full portfolio) could reduce coordination overhead.

### F.3 Broader Impacts

**Potential Benefits:** Our findings could improve capital allocation efficiency in hedge funds, benefiting institutional investors including pension funds and endowments. The framework provides insights into multi-agent coordination applicable beyond finance—research teams, software development, and collaborative AI systems. By demonstrating that communication design must match environmental characteristics rather than following universal principles, we provide evidence against one-size-fits-all

organizational mandates including both radical transparency movements and extreme information siloing.

**Potential Risks:** Widespread adoption could increase market volatility through synchronized behavior when multiple funds implement similar strategies. Natural strategy convergence across funds could create crowded-trade risks where simultaneous exits trigger cascading price impacts. Computational requirements raise environmental concerns if methodology scales without efficiency improvements. Results might justify excessive opacity under misapplied "beneficial information barriers" rationale. The demonstration that conversation quality metrics (CORE scores) don't predict performance could discourage valuable discussion quality monitoring if misinterpreted to mean conversation content is irrelevant.

**Deployment Considerations:** Implementations should include position limits, correlation monitoring, and circuit breakers when convergence exceeds thresholds. Regulators should evaluate whether surveillance frameworks adequately address emergent coordination in multi-agent systems. Firms should offset carbon emissions and explore efficient architectures. Selective transparency—sharing risk metrics while keeping implementations private—may optimize information benefits while preserving diversity. Organizations should monitor not just CORE scores but actual portfolio correlations, as our finding that high-quality conversations don't guarantee performance suggests focusing on outcome metrics rather than process quality metrics.

## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: The abstract and introduction accurately reflect the paper that Communication improves performance but depends on market characteristics, all structures converge to similar correlations, and that performance differences stem from behavioral mechanisms.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: Appendix Section "Limitations" explicitly discusses: Computational constraints, 30 iterations without systematic hyperparameter tuning, focus on liquid US equities during bullish period (Jan 2024-Sep 2025), 5 agents vs real funds with 50-500 analysts, LLM agents may not capture human analyst behaviors, 21-month horizon may not capture full market cycles.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory Assumptions and Proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This is an empirical paper focused on experimental evaluation of multi-agent trading systems. It does not present formal theoretical results requiring mathematical proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

### 4. Experimental Result Reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper provides comprehensive implementation details including the model (GPT-4o-mini), complete function library in Appendix A.3, prompt templates in Appendix B, market universes with specific tickers, temperature settings for different tasks, 30 independent iterations, time period (January 2024 to September 2025), and data source (yfinance API).

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

- (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

## 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The Github repo is linked at the beginning. The data uses publicly available stock market data through yfinance API.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental Setting/Details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Section 3 and Appendices A-B provide comprehensive experimental details. These include clear definitions of the 5 organizational structures, agent architecture specifications, monthly rebalancing procedures, 30 iterations per configuration totaling 450 experiments, performance metrics including returns, Sharpe ratio and correlations, CORE scores for conversation quality assessment, and complete prompt templates for all agent interactions.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment Statistical Significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The paper reports appropriate statistical information for its experimental results. Appendix C.1 provides paired t-tests with Bonferroni correction comparing each configuration against baseline, specifying test statistics, degrees of freedom, and p-values. Confidence intervals throughout Figures 2-3 and Tables 1-4 are calculated using bootstrap resampling with 1,000 iterations, and ANOVA tests in Appendix C.2 confirm no significant differences in final correlations across organizational structures. The paper clearly states which results are statistically significant and which are not, with appropriate corrections for multiple comparisons.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

## 8. Experiments Compute Resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Section A.1 clearly states the computational resources used. The experiments were run on a single NVIDIA L40S GPU with 48GB VRAM, consuming 150 GPU-hours total with 50 hours allocated per market universe, and temperature and token settings are provided for different task types.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

## 9. Code Of Ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines?>

Answer: [Yes]

Justification: The research uses publicly available financial data for legitimate academic purposes, focuses on improving multi-strategy fund organization, and includes discussion of broader impacts. No ethical concerns are apparent.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

## 10. Broader Impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Section F.3 provides a thoughtful discussion of both positive and negative societal impacts. Benefits include improved capital allocation efficiency and insights for multi-agent coordination applicable beyond finance, while risks encompass increased market volatility from synchronized trading, crowded trade risks, systemic concerns from strategy convergence, environmental impacts from computational requirements, and potential misapplication to justify excessive opacity.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

## 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper presents research methodology and analysis techniques rather than releasing high-risk models or datasets. The trading strategies studied are experimental and not production systems.

Guidelines:

- The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper properly cites OpenAI models and mentions using yfinance API for market data access. All data sources are standard financial market data that is publicly available.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

## 13. New Assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not introduce new datasets, models, or code releases beyond the research methodology itself.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

## 14. Crowdsourcing and Research with Human Subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This research involves only computational experiments with LLM agents and market data, no human subjects or crowdsourcing.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

**15. Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: No human subjects research was conducted; the study uses only computational agents and public market data.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.