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. [1]

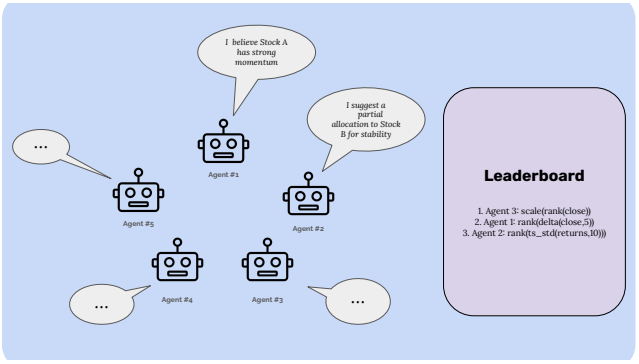

Figure 1: Overview of multi-agent trading framework.

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

[1] Code and data available at: https://github.com/Jerick-1380/multi-agent-alpha-generation

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Generative AI in Finance.

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

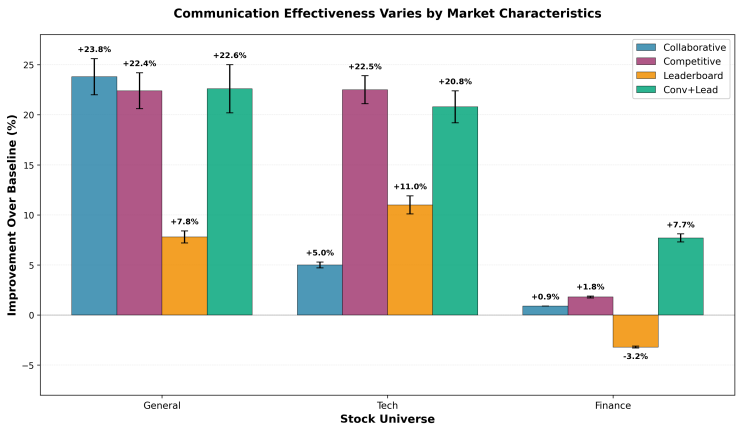

Figure 2: Communication effectiveness varies by market characteristics. Error bars show 95% confidence intervals across 30 iterations.

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.

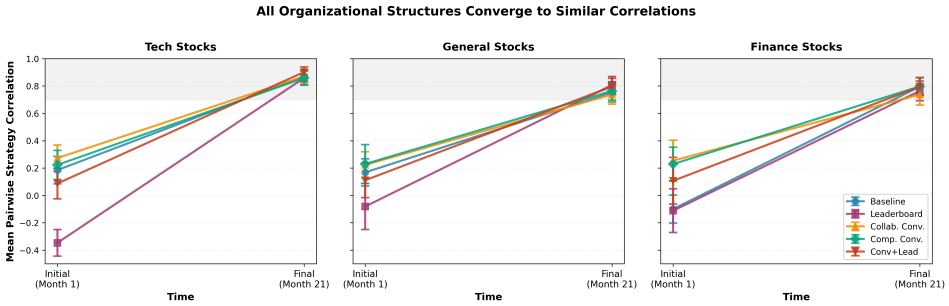

Figure 3: Mean pairwise strategy correlations from Month 1 to Month 21. All configurations converge to similar final correlations regardless of information sharing, including isolated baseline agents. Error bars show 95% confidence intervals across 30 iterations.

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.	$99.88\% \pm 5.57\%$	**47.22%** $\pm 2.27\%$	$72.69\% \pm 1.55\%$
Comp. Conv.	**116.50%** $\pm 6.31\%$	$46.68\% \pm 2.43\%$	$73.35\% \pm 1.98\%$
Conv. + Lead.	$114.92\% \pm 8.21\%$	$46.74\% \pm 4.07\%$	**77.54%** $\pm 3.95\%$
Best Improvement	**+22.5%**	**+23.9%**	**+7.7%**

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 ± 0.05	2.17 ± 0.17	2.00 ± 0.07
Leaderboard	2.32 ± 0.08	2.02 ± 0.14	2.01 ± 0.07
Collab. Conv.	2.26 ± 0.10	**2.38** ± 0.11	2.06 ± 0.03
Comp. Conv.	**2.47** ± 0.09	2.31 ± 0.12	2.06 ± 0.04
Conv. + Lead.	2.44 ± 0.14	2.30 ± 0.25	**2.13** ± 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 ± 0.100	0.870 ± 0.045	+0.684
	Leaderboard	-0.347 ± 0.097	0.859 ± 0.052	+1.206
	Collab. Conversation	0.273 ± 0.096	0.870 ± 0.048	+0.597
	Comp. Conversation	0.223 ± 0.106	0.859 ± 0.051	+0.636
	Conv. + Leaderboard	0.087 ± 0.112	0.902 ± 0.038	+0.815
General	Baseline	0.168 ± 0.100	0.752 ± 0.068	+0.584
	Leaderboard	-0.081 ± 0.168	0.807 ± 0.063	+0.888
	Collab. Conversation	0.223 ± 0.096	0.738 ± 0.071	+0.515
	Comp. Conversation	0.230 ± 0.142	0.764 ± 0.067	+0.534
	Conv. + Leaderboard	0.111 ± 0.128	0.796 ± 0.060	+0.685
Finance	Baseline	-0.100 ± 0.102	0.796 ± 0.065	+0.896
	Leaderboard	-0.112 ± 0.160	0.764 ± 0.072	+0.876
	Collab. Conversation	0.254 ± 0.150	0.738 ± 0.078	+0.484
	Comp. Conversation	0.229 ± 0.123	0.797 ± 0.064	+0.568
	Conv. + Leaderboard	0.107 ± 0.171	0.796 ± 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 "coopetition" 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.