Multi-Agent Debate for Explainable Trading *

Juli Huang, Alanood Alrassan, Deveen Harischandra, Theodore Wu, Veljko Skarich, Matthew Hayes
Deparment of Engineering, Stanford University
{julih, alanoodr, deveen, wutheodo, vskarich, mhayes3}@stanford.edu

1 Problem Statement and Motivation

Quant trading firms rely on opaque statistical models, while current LLM trading products often generate fluent narratives without reliable alpha—essentially "talking" rather than reasoning. We target this gap by building a multi-agent debate system inside a stock-market simulator. Specialized agents (macro, value, risk) will debate trade proposals to produce auditable explanations linking news to action. Markets serve as a harsh testbed for this goal because traders must interpret complex narratives (e.g., Fed communications) rather than just forecast prices.

Our primary motivation is to test whether structured disagreement reduces common LLM failures like overconfidence, hallucinated causality, and groupthink. We will compare single-agent vs. multi-agent setups, evaluating whether improved reasoning quality (via T³/CRIT-style rubrics) correlates with better risk-adjusted returns. If debate improves performance, it supports the mechanism for decision-making under uncertainty; if not, we can distill the most effective constraints into stronger single-agent policies (Li et al. [2025], Xiao et al. [2024])

2 Background and Related Work

Most production trading systems prioritize robustness over interpretability, making it difficult to audit how conflicting signals are reconciled (FTI Consulting [2023]). To balance market realism with inspection, we utilize ABIDES and ABIDES-Gym, which model exchange latency and order-book dynamics while providing a standardized interface for repeatable experiments (Byrd and Hybinette 2021; Yao et al. 2024).

Our approach builds on recent agent-based finance research. While TradingAgents (Xiao et al. [2024]) and QuantAgents (Li et al. [2025]) demonstrate the utility of role-based LLM frameworks, and AlphaAgents (Zhao et al. [2025]) explores collaboration, we specifically leverage "Multi-Agent Debate" (Du et al. [2023]). By forcing agents to challenge reasoning over multiple rounds rather than just collaborating, we aim to expose weak causal claims and produce a transparent decision trail that standard ensemble methods often lack.

3 Proposed System Design

We propose an experimental framework with three components: (1) a market simulation environment in which agentic systems observe market information and execute trades; (2) a set of N multi-agent systems, each optimizing a stock portfolio; and (3) a standardized evaluation mechanism. Figure 1 illustrates the interaction among these components.

3.1 Market Simulation Environment

Existing market simulators such as PyMarketSIM (Mascioli et al. [2024]) and ABIDES (Byrd et al. [2019]) primarily target agents operating on high-precision numerical inputs. In contrast, our simulation must also emphasize qualitative information to study agentic reasoning behavior. Agents must observe historical price data together with textual inputs such as financial news (Dong et al. [2024]) and earnings call transcripts (Kurry [2025]).

*GitHub: https://github.com/TheClassicTechno/cs372research

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

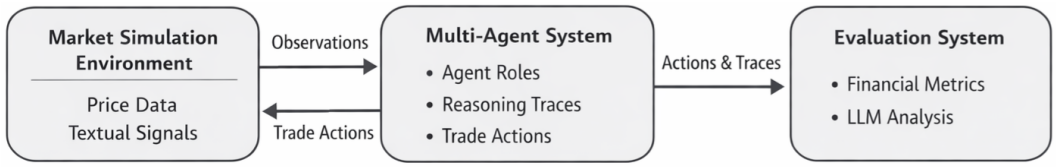

Figure 1: Multi-agent systems interact with the market simulation environment via observations and trade actions, and an evaluation system scores agent behavior and outcomes.

The observations within the environment will be sourced from a benchmark dataset covering a fixed time horizon and a predefined set of stock ticker symbols. At each timestep, agents receive identical observations and may execute trades via a broker API exposed as a standardized tool interface.

3.2 Multi-Agent Systems

We aim to evaluate multiple multi-agent architectures within this environment. The framework must support flexible agent configurations, inspection of reasoning traces, and iterative evaluation. We implement all agents using LangChain (Chase [2022]), which provides native support for these capabilities and facilitates reproducibility.

3.3 Evaluation System

For each system, we will use an automated evaluation pipeline that scores performance using the metrics in Table 1. Deterministic evaluators compute financial performance metrics, while LLM-based evaluators assess qualitative properties such as reasoning quality. All evaluators are implemented using the LangGraph framework.

4 Evaluation Plan and Benchmarks

Evaluation Axis	**Metrics / Dataset**
Financial Performance	PnL, Sharpe ratio, Maximum drawdown
Group Reasoning Quality	Pearl-style cases from agent debates ($\sim$ 100-200)
Individual Agent Reasoning Quality	100 general financial cases (control)

Table 1: Evaluation metrics for downstream trading performance and causal reasoning quality.

We plan to evaluate financial performance and reasoning quality on both a group and individual basis on the T^3 benchmark based on Pearl's causality framework Pearl [2009]. Learnings from group debate will be encoded in prompts that will be used to persist causal reasoning advances. We also plan to use RCA Chang [2026], CRIT Chang [2023], Chain of Thought, Wei et al. [2022], and human-in-the-loop for post-mortem debate feedback and analysis.

4.1 Research Questions

We investigate the impact of multi-agent debate on causal reasoning quality and downstream financial decision making through the following research questions:

RQ1: Does multi-agent debate improve causal reasoning quality in financial decision making? For each trading episode, agents first observe market information and engage in multi-agent debate, producing causal and counterfactual claims such as "The CPI surprise caused the selloff," or "If rates rise tomorrow, bank stocks will fall." We extract claims and scenarios from these statements, construct T^3-style causal reasoning cases, and evaluate causal reasoning quality using the T^3 benchmark.

RQ2: Does improved causal reasoning correlate with downstream trading performance? We plan to perform a joint analysis between T^3 reasoning scores and standard trading metrics, including profit and loss (PnL), Sharpe ratio, and maximum drawdown. Results are categorized as exhibiting positive correlation, negative correlation, or no significant correlation.

RQ3: Does debate reduce known LLM failure modes? We test whether multi-agent debate mitigates common large language model failures, including: overconfidence, sycophancy, hallucinated causality, unjustified Level-2 intervention claims, counterfactual reasoning without invariant structure.

References

David Byrd, Maria Hybinette, and Tucker Hybinette Balch. Abides: Towards high-fidelity market simulation for ai research, 2019. URL https://arxiv.org/abs/1904.12066.

Edward Y. Chang. Prompting large language models with the socratic method. *arXiv preprint arXiv:2303.08769*, 2023. URL https://arxiv.org/abs/2303.08769.

Edward Y. Chang. Internal reasoning vs. external control: A thermodynamic analysis of sycophancy in large language models. *arXiv preprint arXiv:2601.03263*, 2026. URL https://arxiv.org/abs/2201.11903.

Harrison Chase. LangChain, October 2022. URL https://github.com/langchain-ai/langchain.

Zihan Dong, Xinyu Fan, and Zhiyuan Peng. Fnspid: A comprehensive financial news dataset in time series, 2024.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multi-agent debate. *arXiv preprint arXiv:2305.14325*, 2023. URL https://arxiv.org/abs/2305.14325.

FTI Consulting. Artificial intelligence in trading and portfolio management. *FTI Consulting Insights*, 2023. URL https://www.fticonsulting.com/insights/articles/artificial-intelligence-trading-portfolio-management.

Kurry. Sp 500 earnings transcripts dataset, 2025. URL https://huggingface.co/datasets/kurry/sp500_earnings_transcripts.

Xiangyu Li, Yawen Zeng, Xiaofen Xing, Jin Xu, and Xiangmin Xu. Towards multi-agent financial system via simulated trading. *Findings of EMNLP*, 2025. URL https://aclanthology.org/2025.findings-emnlp.945.pdf. arXiv preprint.

Chris Mascioli, Anri Gu, Yongzhao Wang, Mithun Chakraborty, and Michael Wellman. A financial market simulation environment for trading agents using deep reinforcement learning. In *Proceedings of the 5th ACM International Conference on AI in Finance*, ICAIF '24, page 117–125, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400710810. doi: 10.1145/3677052.3698639. URL https://doi.org/10.1145/3677052.3698639.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, UK, 2 edition, 2009.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022. URL https://arxiv.org/abs/2201.11903.

Yijia Xiao, Edward Sun, Di Luo, and Wei Wang. Tradingagents: Multi-agents llm financial trading framework. *arXiv preprint arXiv:2412.20138*, 2024. URL https://arxiv.org/pdf/2412.20138. Accessed: February 1, 2026.

Yifan Zhao, Ming Li, and Hao Chen. Alphaagents: Large language model based multi-agents for equity portfolio construction. *arXiv preprint arXiv:2508.11152*, 2025. URL https://arxiv.org/abs/2508.11152.