

Ensemble LLM Consensus Architectures for Deterministic Trade Signal Generation: An Empirical Study of Multi-Persona AI Panels in Retail FinTech Applications

Michael Anton

Midas Invest, LLC, CEO

Edge Stack Trading, Developer

<https://edgestacktrading.com/>

michael.novamarket@gmail.com

Jean T. Wells, Esq., CPA

Howard University, Associate Professor

jean.wells@howard.edu

<https://doi.org/10.68414/HMXY3672>

Abstract

This paper presents an empirical study of a multi-persona large language model (LLM) consensus architecture deployed in a live retail FinTech platform for equity and options trade signal generation. Grounded in a prior quantitative trading framework (Anton, 2025) and building upon academic antecedents in multi-agent LLM systems (Xiao et al., 2024), the Edge Stack Trading platform deploys a six-persona AI analyst panel producing structured, constrained verdicts across a real-time multi-signal scoring architecture. Analysis of nine live trade reports generated between April and May 2026 reports a preliminary 87.5% directional accuracy rate across eight assessable signals in an exploratory proof-of-concept evaluation, with unanimous panel consensus strongly correlated with system confidence extremes. Inter-model disagreement functions as an auditable risk signal, and model API failures produce identifiable, recoverable degradation patterns consistent with graceful degradation requirements for financial AI systems. The architecture addresses the 2026 conference theme of Reliable and Deterministic AI for Financial Systems through structured verdict formats, role-constrained persona prompting, and explicit multi-signal deterministic scoring as mechanisms for AI governance in retail financial decision support. One incorrect signal (Palantir) reveals a boundary condition in which extreme valuation risk overrides technical alignment, informing a specific architectural recommendation.

Implications for HBCU financial technology education, democratized institutional-grade AI analysis, and regulatory auditability are discussed.

Keywords: multi-agent LLM, ensemble AI, trade signal generation, deterministic AI, FinTech, VolumeZoneOscillator, consensus architecture, AI auditability, retail trading, financial AI governance

Introduction

The emergence of large language models (LLMs) as analytical instruments in financial markets represents one of the most consequential technological shifts in retail FinTech since algorithmic trading. However, single-agent LLM deployments in financial contexts suffer from well-documented limitations: hallucination risk, absence of role-based specialization, lack of internal debate mechanisms, and insufficient auditability for regulatory contexts (Xiao et al., 2024). These limitations are especially acute in retail trading environments where investors act directly on AI-generated recommendations without institutional intermediation or validation infrastructure.

This paper examines the Edge Stack Trading platform, a live retail FinTech system deploying a six-persona LLM consensus panel for equity and options trade signal generation. The platform was developed from a theoretical foundation in the Multi-Factor Probabilistic Trading System (Anton, 2025), which formalized Bayesian-style probability stacking across multiple independent market signals: the VolumeZoneOscillator (VZO), NYSE TICK breadth, VIX volatility regime, sector ETF alignment, news sentiment, and earnings binary event risk. Edge Stack Trading extends this framework by replacing human signal synthesis with an ensemble of LLM agents, each assigned a distinct analytical persona, producing structured verdicts aggregated into a consensus recommendation.

Contributions are fourfold: (1) first empirical description of a live production multi-persona LLM consensus panel for retail equity trading; (2) analysis of inter-model agreement rates across nine live reports; (3) post-hoc accuracy assessment against observed market outcomes yielding a preliminary 87.5% directional accuracy across a small exploratory sample; and (4) examination of an API failure event as a graceful degradation case study. The platform also addresses equity of access, making institutional-grade AI analysis available at retail scale consistent with the democratization mission of HBCU financial education.

Literature Review

Theoretical Foundations in Probabilistic Trading

The Edge Stack Trading architecture derives from market microstructure theory and behavioral finance. Kyle (1985) and Easley and O'Hara (1987) established the informed trading hypothesis, demonstrating that sophisticated market participants reveal informational advantages through volume and order flow. The VolumeZoneOscillator operationalizes this insight by measuring the balance between volume on price advances versus declines, normalized to a bounded oscillator (Anton, 2025). Kahneman and Tversky (1979) and Tetlock (2007) provided behavioral finance grounding for sentiment-based signal components, demonstrating that media pessimism predicts short-term price pressure and that extreme sentiment conditions create exploitable contrarian opportunities.

Anton (2025) synthesized these streams into a multi-factor probabilistic framework, formalizing Bayesian probability stacking: given conditionally independent signals S_1 through S_n , the posterior probability $P(\text{Success}|S_1, \dots, S_n)$ substantially exceeds any individual signal's predictive power. The framework specified VZO zone-dependent entry rules, NYSE TICK thresholds, VIX regime protocols, earnings binary event management, and a three-tier position sizing decision tree, constituting the direct intellectual predecessor to the automated system examined here.

Multi-Agent LLM Frameworks in Finance

Xiao et al. (2024) proposed TradingAgents, a framework modeled on trading firm organization, featuring LLM agents in specialized roles: fundamental analysts, sentiment analysts, technical analysts, risk managers, and a fund manager. Operating through five sequential layers, the system demonstrated improvements in cumulative returns, Sharpe ratio, and maximum drawdown relative to Buy-and-Hold, MACD, KDJ+RSI, Zero Mean Reversion, and SMA baselines in backtesting on large-cap equities. Du et al. (2023) demonstrated that multi-agent debate mechanisms improve factuality and reasoning, providing theoretical support for consensus architectures. Yu et al. (2024) further showed that conceptual verbal reinforcement in LLM multi-agent systems enhances financial decision-making quality.

Critical gaps remain: all major frameworks operate in simulated backtests rather than live production systems; existing work does not examine inter-model disagreement as an auditable reliability signal; and no prior study has integrated VZO-based multi-timeframe analysis, options

PCR flow, and NYSE TICK extrapolation with LLM consensus panels. This paper addresses all three gaps through analysis of a live deployment with post-hoc accuracy validation.

AI Reliability and Auditability in Financial Systems

The 2026 conference theme of Reliable and Deterministic AI for Financial Systems highlights the critical need for explainable, auditable AI governance in financial contexts. The Edge Stack Trading architecture addresses these requirements through: (1) deterministic scoring, where confidence scores are computed from explicit weighted signal components with full input traceability; (2) structured constrained verdict formats preventing role drift; and (3) panel transparency, preserving all six analyst verdicts in each report output to enable complete reasoning chain review.

Methodology

Platform Architecture

Edge Stack Trading is a production retail FinTech platform delivering AI-assisted trade analysis across equities, ETFs, and options. It integrates a deterministic multi-signal scoring engine with a six-persona LLM analyst panel, operating on a proprietary on-premises AI inference cluster. A Python-based asynchronous RSS aggregation engine spanning 48 financial feeds provides real-time news data with concurrent fetching, MD5 headline deduplication, and normalized output.

The deterministic scoring engine computes system confidence (0–100%) from eleven weighted components: (1) market direction from SPY/QQQ/DIA composite; (2) VIX volatility regime; (3) cryptocurrency sentiment from a five-asset weighted basket; (4) NYSE TICK breadth extrapolated from a 60-stock basket to ~3,000 NYSE securities; (5) sector ETF alignment across all eleven SPDR sectors; (6) insider sentiment; (7) earnings quality from prior four-quarter history; (8) RSS news sentiment; (9) VZO alignment across one-minute, five-minute, and one-day timeframes; (10) options put-call ratio; and (11) leadership and fraud risk. Scores map to: $\geq 70\%$ = BUY/STRONG BUY; 50–69% = HOLD; $< 50\%$ = SELL/STRONG SELL, applied consistently across all asset types.

The Six-Persona LLM Panel

Six personas operate in parallel with no cross-agent communication, each constrained to ~178 words with a required imperative verdict label: (1) Fundamental Analyst — earnings quality and revenue trajectory; (2) Risk Manager — position sizing and risk-reward profile; (3)

Market Strategist — macro backdrop, VIX regime, and sector rotation; (4) News Sentiment Analyst — headline tone and social sentiment; (5) Senior Technical Analyst (panel leader) — synthesizes all data and issues the final consensus verdict with stated confidence; and (6) Financial Educator — plain-English translation for non-professional investors. Parallel rather than sequential execution prevents the telephone effect — progressive information distortion across sequential agent handoffs — documented as a limitation of sequential architectures (Xiao et al., 2024).

Data Collection, Sample, and Analytical Framework

Nine live reports generated between April 22 and May 23, 2026, are analyzed. Reports span individual equities (INTC, IEX, IBM, INTU, PLTR, MU, BXDC), a sector ETF (IGV), and an options strategy (NVDA), covering swing equity, day trade equity, cash-secured put, and long call structures. Post-hoc accuracy is assessed against market prices as of June 25, 2026, providing a 33–64 day outcome window. Accuracy is evaluated on a binary directional basis: BUY correct if price appreciated meaningfully; SELL/AVOID correct if price declined or the avoided trade was unprofitable; HOLD correct if the wait-for-confirmation posture was vindicated. Analysis examines: (1) confidence score distribution; (2) convergence between system score extremes and panel unanimity; (3) signal factors driving divergence; (4) post-hoc accuracy; and (5) failure mode resilience.

Results and Findings

System Confidence Scores, Panel Consensus, and Post-Hoc Accuracy

Table 1 presents all nine reports with system confidence, verdict, panel dissent count, and post-hoc accuracy outcomes. Of eight assessable signals, seven were directionally correct (87.5%), though this result should be interpreted cautiously given the small sample size and exploratory nature of the evaluation.

Table 1
Live Edge Stack Trading Reports with Confidence Scores, Panel Dissent, and Post-Hoc Accuracy (June 25, 2026)

Ticker	Date	Score	Verdict	Dissent	Price at Signal	Price Jun 25	Result
INTC	Apr 23	84%	BUY	1/6	\$66.68	~\$135	✓ +102%

Ticker	Date	Score	Verdict	Dissent	Price at Signal	Price Jun 25	Result
IEX†	Apr 23	19%	SELL*	—†	\$205.90	N/A	Excl.†
IBM	Apr 22	59%	HOLD	4/6	\$251.80	~\$262	✓ Correct
NVDA	May 23	2%	AVOID	6/6	\$214.28	~\$199	✓ Avoided loss
INTU	May 22	2%	SELL	5/6	\$318.84	~\$259	✓ −19%
IGV	May 21	51%	HOLD	6/6	\$93.20	~\$94	✓ Flat/HOLD
PLTR	May 21	100%	BUY	0/6	\$137.85	~\$107–\$113	✗ −20%
MU	May 20	100%	BUY	0/6	\$726.00	~\$1,048+	✓ +44%+
BXDC	May 18	50%	HOLD	6/6	\$20.78	~\$21.84	✓ Correct

† IEX: Full API panel failure (404 errors). Deterministic engine operative; panel unavailable. Excluded from accuracy assessment (see Section 4.4).

* SELL verdict from deterministic scoring engine only.

System confidence scores clustered at functional extremes, with four of nine at or below 51% and three at 84% or above. This polarization reflects the Bayesian stacking mechanism: signal alignment drives composite scores toward extremes while signal conflict produces mid-range outputs (Anton, 2025). A clear convergence pattern emerged between score extremes and panel unanimity: both 100% system confidence reports (PLTR, MU) achieved 0/6 panel dissent, while both 2% reports (NVDA, INTU) generated 5–6 of 6 dissenting analysts. This convergent validity indicates that the deterministic scoring engine and the LLM panel, operating through different reasoning mechanisms from the same data, consistently agree at conviction extremes.

Notable Signal Outcomes

Two signals produced exceptional gains. INTC appreciated over 100% from \$66.68 to approximately \$135, driven by the semiconductor sector rally, the Apple-Intel chip partnership announcement, and BofA raising its price target to \$160. Only one of six panel analysts dissented (Market Strategist, citing elevated VIX and negative XLK sector alignment), a disagreement that post-hoc analysis confirms was overridden correctly by the consensus BUY. MU appreciated 44%+ from \$726, with Micron reporting record fiscal Q3 2026 revenue of \$41.5 billion on June 24, 2026, shattering consensus estimates by over \$6 billion. The unanimous

100% system BUY and zero panel dissent were fully validated. The INTU SELL at \$318.84 was validated by a 19% decline to ~\$259, with Goldman Sachs subsequently downgrading to Sell, securities fraud investigations launched, and the company's market capitalization declining \$131 billion from its prior year high. The NVDA AVOID correctly prevented a cash-secured put assignment that would have required purchasing stock at \$207.90 when the market price fell to ~\$199.

The Single Incorrect Signal: PLTR and Valuation Override

The single incorrect signal was PLTR, which generated a 100% system BUY and unanimous panel consensus recommending a cash-secured put at the \$135 strike. The stock declined 18–23% to \$107–\$113, resulting in a put assignment well below breakeven. Critically, the original report contained its own warning: the Risk Manager flagged extreme valuation (~200x forward earnings), the Senior Technical Analyst downgraded the system's 100% to 78% citing missing VIX data, and the concurrent IGV report had identified the SaaS sector as experiencing its worst quarter since 2008. These qualitative flags were present but overridden by technical signal alignment. This case contributes a specific architectural recommendation: implementing valuation-threshold hard constraints that prevent BUY signals when price-to-earnings multiples exceed defined limits, regardless of technical score. Table 2 presents the full panel verdict distribution for all nine reports.

Table 2

Six-Persona Panel Verdict Distribution Across All Nine Live Reports

Ticker	Fund. Anlst	Risk Mgr	Mkt Strat.	News Sntmt	Sr. Technical	Fin. Educ.
INTC	BUY	PROCEED	AVOID	BUY	BUY (consens.)	PROCEED
IEX†	ERROR	ERROR	ERROR	ERROR	ERROR	ERROR
IBM	HOLD	PROCEED	BUY	AVOID	HOLD (consens.)	MONITOR
NVDA	AVOID	AVOID	AVOID	AVOID	AVOID (consens.)	AVOID
INTU	AVOID	AVOID	AVOID	AVOID	SELL (consens.)	AVOID
IGV	HOLD	HOLD	HOLD	HOLD	HOLD (consens.)	HOLD
PLTR	BUY	PROCEED	PROCEED	PROCEED	BUY (consens.)	BUY

Ticker	Fund. Anlst	Risk Mgr	Mkt Strat.	News Sntmt	Sr. Technical	Fin. Educ.
MU	BUY	BUY	BUY	BUY	BUY (consens.)	BUY
BXDC	HOLD	HOLD	HOLD	HOLD	HOLD (consens.)	HOLD

† IEX: All personas returned API 404 errors.

API Failure Case Study: The IEX Degradation Event

The IEX report (April 23, 2026) illustrates graceful degradation: all six LLM personas returned 404 API errors due to an unavailable model endpoint. The failure produced consistent, identifiable error signatures — explicit error objects with type, message, and request ID — rather than silent omissions. Critically, the deterministic scoring engine was unaffected, computing a 19% SELL confidence from available market data. Simultaneously, the equity-level module independently produced an 87% STRONG BUY from VZO, chart trajectory, analyst ratings, and social sentiment. This system-to-equity divergence (19% SELL vs. 87% BUY) illustrates precisely the disambiguation function that the LLM panel provides: resolving conflicts between scoring layers requires qualitative reasoning that deterministic scoring alone cannot supply. The transparent error signature ensures that a human reviewer can immediately identify the degraded state before acting, satisfying auditability requirements central to the conference theme.

Discussion

Deterministic Scoring as an AI Governance Mechanism

The explicit, weighted signal architecture provides full output traceability: any confidence score can be decomposed into its constituent signal contributions, satisfying regulatory explainability requirements without reliance on post-hoc attribution methods applied to opaque neural networks. For example, NVDA's 2% score decomposes to: VZO alignment (-4 pts), macro direction (-10 pts), volatility caution (-10 pts), crypto sentiment (-3 pts), NYSE TICK (+1 pt), sector positive (+11 pts), earnings quality (+8 pts), news sentiment (+3 pts). Every component is auditable and disputable by a reviewing party — a governance property unavailable in end-to-end neural approaches. This aligns with the conference theme's emphasis on reliable and deterministic AI for financial systems.

Inter-Model Disagreement as a Risk Signal and PLTR Implications

Inter-model disagreement functions as an auditable risk signal: high dissent rates surface analytical tension requiring human review rather than automatic execution. IBM's 4/6 dissent rate correctly identified a VZO overbought timing problem the 59% system score alone would have obscured. However, PLTR demonstrates the inverse: unanimous consensus with 100% system confidence produced the study's only incorrect signal, showing that consensus is a necessary but insufficient condition for accuracy. The panel's own internal risk flags — extreme valuation, missing VIX data, sector distress — were present and overridden. Future iterations should implement valuation-based hard constraints and sector-distress override rules that cannot be overridden by technical alignment alone.

Equity of Access and HBCU Relevance

Edge Stack Trading makes institutional-grade multi-factor analysis — historically accessible only to professional traders with access to dedicated analytical infrastructure — available at retail scale. The Financial Educator persona specifically addresses accessibility for first-generation market participants. The platform has been applied in educational contexts at Howard University's School of Business in a Stock Market Investment Planning module, and represents a practical demonstration of the democratization mission central to HBCU financial technology education.

Implications, Limitations, and Future Research

Practical and Educational Implications

The Edge Stack Trading architecture carries several practical implications for the design of AI-assisted financial systems. First, the deterministic scoring layer provides a governance mechanism that is fully decomposable and auditable without reliance on post-hoc attribution methods. Practitioners and regulators can trace any system verdict to its constituent signal inputs, a property directly relevant to emerging explainability expectations in retail FinTech. Second, the use of structured, role-constrained LLM personas addresses the hallucination and role-drift risks documented in single-agent financial AI deployments, offering a replicable design pattern for developers building AI tools for retail investors. Third, the treatment of inter-model disagreement as an explicit risk signal—rather than noise to be suppressed—offers a practical audibility mechanism: high-dissent verdicts surface analytical tension for human review before execution, reducing the risk of automated compliance with erroneous signals.

For HBCU financial technology education, the platform demonstrates how institutional-grade multi-factor AI analysis can be made available at retail scale and integrated into coursework. The Financial Educator persona, designed to translate technical verdicts into plain-English explanations, directly addresses the needs of first-generation market participants and novice investors who may lack the background to interpret raw quantitative signals. The architecture's application in Howard University's School of Business Stock Market Investment Planning module illustrates the pedagogical potential of live AI financial systems as teaching instruments, offering students exposure to production-grade analysis pipelines as part of their training.

Limitations

This study has several important limitations that constrain the strength of conclusions that can be drawn. The empirical evaluation is based on only nine live trade reports, with eight assessable signals, which is insufficient to establish statistically robust claims regarding system accuracy or generalizability. The reported 87.5% directional accuracy should be interpreted as a preliminary proof-of-concept result rather than a validated performance benchmark. Without a larger sample, it is not possible to determine whether this accuracy rate significantly exceeds what would be expected by chance, and no confidence intervals or formal hypothesis tests are presented to support the directional accuracy claim.

Additional limitations include: (1) binary rather than magnitude-adjusted accuracy assessment, which does not account for the economic significance of correct or incorrect signals; (2) the absence of risk-adjusted performance metrics, transaction cost analysis, or benchmark comparisons against simpler trading rules or human analyst judgments; (3) a sample period coinciding with specific semiconductor and SaaS sector dynamics that may not be representative of other market regimes; (4) limited reproducibility due to the proprietary nature of the platform, which constrains other researchers' ability to evaluate or replicate the proposed architecture; and (5) the accuracy criteria for BUY, SELL, AVOID, and HOLD recommendations, while described, could be defined with greater precision to enable systematic replication. Claims regarding system reliability and AI governance should be understood as exploratory findings from a live proof-of-concept deployment rather than as validated production-benchmark assertions.

Future Research

Future research should prioritize expanding the empirical dataset substantially: longitudinal accuracy assessment across hundreds of live reports spanning multiple market regimes and asset classes is necessary before stronger conclusions can be drawn. Such datasets would enable persona-specific accuracy analysis by market regime, as well as formal testing of valuation-threshold hard constraints and sector-distress override rules identified by the PLTR case. Benchmarking against established baselines—including Buy-and-Hold, MACD, and human analyst consensus—and regulatory analysis relative to SEC and FINRA explainability guidance for AI-assisted financial advice represent important additional avenues for future inquiry.

Conclusion

This paper presents the first empirical study of a live production multi-persona LLM consensus panel for retail equity and options trade signal generation, demonstrating a preliminary 87.5% directional accuracy rate across eight assessable signals on the Edge Stack Trading platform (April–May 2026). Key contributions include: fully decomposable deterministic scoring supporting regulatory auditability; structured constrained persona verdicts preventing role drift; inter-model disagreement as an auditable risk signal; and transparent API failure signatures maintaining human oversight capacity. These design principles offer a replicable governance framework that directly serves the HBCU community’s mission of financial empowerment and equitable access to sophisticated analytical tools.

References

- Anton, M. (2025, December 22). *Multi-factor probabilistic trading system: A quantitative framework for intraday equity trading* [Professional trading manual]. Unpublished manuscript.
- Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2023). *Improving factuality and reasoning in language models through multiagent debate*. arXiv. <https://doi.org/10.48550/arXiv.2305.14325>
- Easley, D., & O'Hara, M. (1987). Price, trade size, and information in securities markets. *Journal of Financial Economics*, 19(1), 69–90. [https://doi.org/10.1016/0304-405X\(87\)90029-8](https://doi.org/10.1016/0304-405X(87)90029-8)
- Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2), 263–292. <https://doi.org/10.2307/1914185>
- Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica*, 53(6), 1315–1335. <https://doi.org/10.2307/1913210>
- Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance*, 62(3), 1139–1168. <https://doi.org/10.1111/j.1540-6261.2007.01232.x>
- Xiao, Y., Sun, E., Luo, D., & Wang, W. (2024). *TradingAgents: Multi-agents LLM financial trading framework*. arXiv. <https://doi.org/10.48550/arXiv.2412.20138>
- Yu, Y., Yao, Z., Li, H., Deng, Z., Jiang, Y., Cao, Y., Chen, Z., Suchow, J. W., Cui, Z., Liu, R., Xu, Z., Zhang, D., Subbalakshmi, K., Xiong, G., He, Y., Huang, J., Li, D., & Xie, Q. (2024). FinCon: A synthesized LLM multi-agent system with conceptual verbal reinforcement for enhanced financial decision making. *Advances in Neural Information Processing Systems*, 37, 137010–137045.