A Unified Cloud-Edge Infrastructure for High-Frequency Financial Forecasting: Synergizing Stream Computing with LLM-Based Reasoning

Wesley Hollis
School of Computational Science and Engineering, Georgia Institute of Technology
wesley.hollis@gatech.edu

Caleb Prescott
Department of Finance and Risk Engineering, New York University Tandon School of Engineering
caleb.prescott@nyu.edu

Abstract

The integration of Large Language Models (LLMs) into the high-frequency financial forecasting domain represents a significant shift from purely numerical autoregressive models to multimodal, reasoning-based systems. However, the computational intensity of LLMs often conflicts with the millisecond-level latency requirements of modern capital markets. This paper introduces a unified cloud-edge infrastructure designed to harmonize high-speed stream computing with the cognitive reasoning capabilities of LLMs. We propose a hierarchical system architecture that offloads time-critical numerical processing to the edge while utilizing a cloud-based reasoning engine for contextual and geopolitical synthesis. By employing a dynamic orchestration layer, the system manages the inherent trade-offs between inference throughput and analytical depth. We provide a rigorous examination of system-level considerations, including deployment strategies, environmental sustainability, and the socio-technical implications of autonomous financial agents. The discussion extends to the governance frameworks necessary to maintain market integrity and the policy implications of deploying such complex infrastructures in a global economic context. Our findings suggest that the synergy of stream computing and agentic reasoning provides a more robust and adaptive forecasting mechanism than traditional quantitative methods, provided that the underlying infrastructure is designed with a focus on low-latency synchronization and hardware-level security. This research offers a comprehensive roadmap for the next generation of financial forecasting systems that prioritize both speed and contextual intelligence.

Keywords

Distributed Systems, Financial Forecasting, Stream Computing, Large Language Models, Edge Computing, Socio-Technical Infrastructure, Algorithmic Governance.

1. Introduction

The evolution of financial forecasting has been characterized by a relentless pursuit of speed and precision, moving from foundational linear models to the sophisticated deep learning architectures that dominate contemporary high-frequency trading. Despite these advancements, the inherent non-stationarity and extreme noise of financial time series data continue to pose significant challenges to purely statistical approaches. Traditional quantitative models often operate in a vacuum, focusing on numerical patterns while remaining blind to the qualitative drivers of market movement, such as geopolitical events, regulatory shifts, and social sentiment. The emergence of Large Language Models (LLMs) offers a transformative potential to bridge this gap, providing a "reasoning layer" that can synthesize unstructured data into actionable market context. However, the integration of LLMs into high-frequency environments requires a fundamental rethinking of system-level infrastructure.

High-frequency forecasting operates on the scale of milliseconds, whereas the inference cycle of a multi-billion parameter LLM typically spans several seconds. This latency disparity creates a structural bottleneck that prevents the direct application of LLMs in the execution path of a trade. This paper addresses this challenge by proposing a unified cloud-edge infrastructure that decouples the fast-path numerical stream from the slow-path contextual reasoning. In this paradigm, edge nodes located in proximity to exchange data centers handle the high-velocity ingestion and processing of tick data, while a distributed cloud backbone performs asynchronous agentic reasoning. The synergy between these two layers allows the system to generate forecasts that are not only statistically sound but also contextually informed, effectively transforming the forecasting task from a passive curve-fitting exercise into an active process of environmental scanning and logical deduction.

The motivations for developing such a unified infrastructure are rooted in both technical necessity and the socio-economic imperatives of modern capital markets. From a systems engineering perspective, the sheer volume of financial data generated globally necessitates a decentralized approach to minimize network congestion and tail latency. From a socio-technical standpoint, the increasing complexity of globalized finance demands systems that can explain their rationale, providing a layer of interpretability that is critical for risk management and regulatory compliance. This paper provides an exhaustive analysis of the architectural trade-offs involved in this synergy, emphasizing the need for robust, policy-aware deployment strategies that can withstand the inherent volatility of global financial infrastructures.

2. Conceptual Framework: Synergizing Numerical Streams and Agentic Reasoning

The conceptual foundation of our unified infrastructure rests on the distinction between "system one" and "system two" thinking as applied to computational finance. System one represents the fast, reactive, and numerical processing of market streams—the domain of traditional high-frequency trading. System two represents the slower, more deliberative, and logical reasoning—the domain of LLM-augmented analysis. In our proposed framework,

these two systems are not independent but are unified through a continuous feedback loop. Stream computing provides the immediate data substrate, while the LLM-based reasoning engine provides the interpretive priors that allow the system to adapt to regime shifts and unexpected market shocks.

The synergy is operationalized through a hierarchical agentic structure. At the edge, lightweight stream processors execute pre-computed models and filters on live data packets. These processors are informed by "contextual vectors" generated periodically by the cloud-based LLM. These vectors encapsulate a compressed representation of the current macroeconomic environment, sentiment trends, and historical analogies. When a new market event occurs, the edge nodes do not need to perform full LLM inference; instead, they apply the LLM's distilled reasoning to the incoming stream. This approach allows the system to maintain the high throughput of edge computing while benefiting from the deep cognitive insights of a cloud-scale LLM, effectively bridging the gap between raw data and informed action.

Furthermore, the integration of agentic reasoning into time series forecasting allows for a more robust handling of the "long tail" of financial events. Financial markets are frequently influenced by events that have no direct precedent in numerical datasets. An LLM, pre-trained on vast corpora of human knowledge, can recognize the qualitative markers of such events—such as a specific phrasing in a central bank's statement or the escalation of a regional conflict—and adjust the forecasting parameters accordingly. This logical deduction provides a layer of resilience that purely quantitative models lack, transforming the infrastructure into an "anti-fragile" entity that can potentially thrive in the face of uncertainty.

3. Unified Cloud-Edge Architecture and Orchestration

The physical and logical architecture of the system is designed to maximize data locality and minimize end-to-end latency. We propose a three-tier structure consisting of the Local Edge, the Regional Aggregator, and the Global Cloud. The Local Edge layer is composed of low-latency compute nodes deployed within colocation facilities of major financial exchanges. These nodes are responsible for high-frequency data ingestion, feature extraction, and the execution of the "fast-path" models. The architecture at this level prioritizes deterministic execution and minimal memory overhead, utilizing hardware acceleration such as FPGAs and specialized AI chips to ensure that the numerical stream is processed at wire speed.

The Regional Aggregator layer acts as a bridge between the edge and the cloud, managing the synchronization of data and the distribution of reasoning outputs. This layer performs intermediate-level synthesis, aggregating signals from multiple edge nodes to identify broader market correlations. It also serves as a caching layer for the LLM's reasoning traces. When the cloud-based reasoning engine generates a new insight, the regional aggregator distributes it to the relevant edge nodes, ensuring that the local forecasting models are always operating with the most current contextual priors. This hierarchical caching strategy is critical for preventing the cloud backbone from becoming a central point of congestion during periods of extreme market volatility.

The Global Cloud layer houses the large-scale LLM ensembles and the agentic reasoning framework. This is where the "heavy lifting" of linguistic processing and cross-domain synthesis occurs. The cloud infrastructure is designed for high-throughput batch processing of unstructured data, such as news feeds, social media, and regulatory filings. The orchestration layer within the cloud manages the allocation of computational resources across different agents, prioritizing those tasked with monitoring high-impact sectors or regions. This layer also implements a "speculative reasoning" protocol, where the LLM pre-computes potential market scenarios in the background, allowing the system to react instantly if a predicted scenario begins to materialize in the live edge stream.

4. Structural Trade-offs: Latency, Throughput, and Depth

The design of a unified cloud-edge infrastructure for finance involves a constant negotiation between three primary variables: latency, throughput, and reasoning depth. In the high-frequency domain, the cost of latency is often measured in millions of dollars per millisecond. However, the value of a forecast is also dependent on its accuracy and its ability to incorporate complex context. Our architecture manages these trade-offs through a policy-driven orchestration engine that dynamically adjusts the intensity of the reasoning process based on the perceived market risk. During periods of low volatility, the system may favor high throughput and low latency, relying primarily on edge-based numerical models. Conversely, during a market crisis, the system may sacrifice some throughput to allow for deeper, more comprehensive LLM-based analysis.

A critical trade-off manifests in the "context window" management of the LLM. While a larger context window allows the model to consider more historical and qualitative data, it significantly increases the time required for inference. We employ a "semantic distillation" technique, where the cloud engine pre-processes large volumes of data into high-dimensional embeddings that can be quickly queried by the reasoning agents. This allows the system to maintain a vast "memory" of market conditions without the latency penalty of re-processing raw text during every reasoning cycle. The structural choice to use embeddings as the primary communication medium between the cloud and the edge is a fundamental design decision that enables the system to scale its reasoning depth without overwhelming the edge-bound network links.

Another dimension of this trade-off is the energy and computational cost of maintaining a high-frequency LLM infrastructure. The continuous operation of multi-billion parameter models in a cloud environment generates a significant environmental footprint. We address this through a "sustainability-aware" scheduling algorithm that modulates the model size and the frequency of reasoning cycles based on the immediate utility of the forecast. By transitioning to smaller, distilled versions of the LLM during stable market hours and reserving the full-scale ensembles for critical windows, the system optimizes its resource consumption while maintaining its analytical edge. This adaptive approach reflects a broader shift in systems design toward "frugal AI," where performance is balanced against the long-term ecological and economic costs of deployment.

5. Infrastructure Robustness and Adversarial Resilience

In the high-stakes environment of global finance, robustness is not merely a technical goal but a systemic requirement. Our unified infrastructure is designed to be resilient to both hardware failures and adversarial interventions. At the hardware level, we implement a "zero-trust" architecture where every edge node and cloud node must undergo continuous attestation. The use of Trusted Execution Environments (TEEs) ensures that the forecasting models and the LLM's reasoning traces are protected from unauthorized access or tampering, even if the underlying physical server is compromised. This hardware-level security is essential for maintaining the integrity of the forecasting process and preventing the leakage of proprietary trading strategies.

Adversarial resilience is a particularly complex challenge for LLM-based systems. Financial markets are inherently adversarial environments where participants may attempt to manipulate news or social sentiment to "hallucinate" specific market reactions in an LLM-based system. To mitigate this, our infrastructure incorporates a multi-agent validation layer. Every qualitative signal ingested by the system is cross-referenced by multiple independent agents, each using a different model architecture or data source. A reasoning output is only propagated to the edge if a consensus is reached, reducing the risk of the system being misled by isolated "fake news" or sentiment manipulation. This decentralized validation mimics the structure of institutional investment committees but operates with the speed and precision of a distributed computer system.

Furthermore, the system is designed to be resilient to "model drift" and the catastrophic forgetting often associated with online learning. Because the financial environment is constantly evolving, a static LLM will quickly lose its relevance. Our infrastructure implements a "shadow learning" pipeline where a new version of the reasoning engine is continuously trained on live data in the background. This shadow model is compared against the production model in real-time, and a transition is only made once the new model demonstrates superior performance on a held-out set of recent market data. This robust deployment lifecycle ensures that the system can adapt to structural changes in the global economy without the risk of sudden performance degradation.

6. Deployment Strategies and Global Data Sovereignty

Deploying a unified cloud-edge infrastructure across multiple international jurisdictions introduces significant challenges related to data sovereignty and regulatory compliance. Many countries now require that financial data be stored and processed within national borders, which complicates the design of a global reasoning engine. We address this through a "federated reasoning" approach, where localized cloud clusters perform reasoning on sensitive regional data, and only non-sensitive, high-level insights are shared with the global core. This allowed the system to maintain a global perspective while respecting the legal and ethical boundaries of each jurisdiction in which it operates.

The deployment strategy also accounts for the "fragmentation" of global liquidity. As capital

markets become more decentralized with the rise of regional exchanges and "dark pools," the infrastructure must be able to ingest and synthesize data from a highly diverse set of sources. Our system utilizes a modular data ingestion layer that can be easily extended to new venues or asset classes. This modularity is key to the system's long-term viability, allowing it to evolve alongside the market structures it is tasked with forecasting. The deployment is orchestrated via a containerized microservices architecture, which provides the flexibility to deploy specific agents or stream processors to the edge or the cloud as needed.

From a policy perspective, the deployment of such systems raises questions about the "digital divide" in financial markets. Smaller firms may lack the resources to build or lease the sophisticated infrastructure required to run high-frequency LLM reasoning. This could lead to an even greater concentration of market power among a few elite institutions. We advocate for the development of "public-interest" financial infrastructures—shared, regulated platforms that provide standardized reasoning services to a broader range of participants. This would promote market fairness and prevent the emergence of technological monopolies that could destabilize the broader financial ecosystem.

7. Governance, Fairness, and Policy Implications

The governance of autonomous financial infrastructures is one of the most pressing socio-technical challenges of the current era. When systems like the one proposed here are capable of making real-time forecasts that influence global capital flows, the potential for unintended consequences is vast. We propose a governance framework based on "algorithmic accountability," where every decision made by the system is backed by a human-readable reasoning trace. The LLM’s ability to explain its rationale in natural language is a critical tool for this, allowing regulators and internal auditors to understand why a particular forecast was made, rather than treating the system as an opaque "black box."

Fairness in financial forecasting is often overlooked but is increasingly critical as systems become more autonomous. Bias in training data can lead to models that systematically disadvantage certain sectors, regions, or asset classes. Our infrastructure includes a "bias monitoring" agent that continuously audits the system’s forecasts for signs of systematic error or discrimination. If a bias is detected, the system triggers an automatic retraining of the affected agents using a more balanced dataset. This proactive approach to fairness is essential for maintaining the trust of market participants and ensuring that the infrastructure does not exacerbate existing economic inequalities.

The policy implications of this technology extend to the core of market stability. The widespread adoption of highly synchronized, LLM-augmented forecasting systems could lead to "crowded trades" and increased systemic volatility. If all major participants are using similar reasoning engines, they may all react to the same qualitative signals simultaneously, leading to flash crashes or sudden liquidity droughts. To address this, we suggest that regulatory bodies should oversee the "diversity" of financial AI models, encouraging a landscape of heterogeneous reasoning approaches. Policy should also mandate "circuit breakers" for AI agents, similar to those used for trading, that can automatically throttle the

system's influence during periods of extreme market stress.

8. Socio-Technical Implications and the Future of Financial Agents

The shift toward a unified cloud-edge infrastructure for financial forecasting is part of a broader transformation of the socio-technical landscape. As AI agents become the primary participants in capital markets, the role of human experts will move from execution to governance. This requires a new set of skills and a different kind of relationship between humans and machines. Our research emphasizes the need for "human-in-the-loop" systems where the machine handles the scale and speed of data, while the human provides the ethical oversight and long-term strategic direction. This collaborative intelligence is the most sustainable path forward for the industry.

Furthermore, the integration of LLMs into financial systems could lead to a "democratization of expertise," where high-level analytical tools that were previously only available to elite hedge funds are accessible to a wider audience. This could lead to more efficient and inclusive markets, provided that the underlying infrastructure is managed as a public good. However, this also requires a significant investment in digital literacy and a robust regulatory framework to prevent the misuse of these powerful tools. The future of financial agents is not just a technological question but a social and political one, requiring a dialogue between engineers, economists, and policymakers.

In the long term, we envision a "global reasoning mesh" where thousands of specialized agents, hosted on a unified cloud-edge infrastructure, collaborate to maintain the stability and efficiency of the global economy. This mesh would act as a collective immune system for capital markets, identifying risks and opportunities in real-time and providing the logical rationale needed to navigate them. The development of such a mesh is a multi-decade project that will require breakthroughs in distributed systems, cognitive science, and economic theory. This paper represents a first step in that direction, providing the architectural blueprint for a system that can bridge the gap between high-frequency data and deep contextual reasoning.

9. Conclusion

This paper has proposed a unified cloud-edge infrastructure for high-frequency financial forecasting that synergizes the speed of stream computing with the depth of LLM-based reasoning. We have detailed a hierarchical architecture that optimizes the distribution of computational workloads, allowing for contextually informed forecasts at the speed of modern capital markets. Through an analysis of structural trade-offs, we have shown how a policy-driven orchestration layer can manage the inherent tensions between latency and analytical complexity. Our discussion of robustness, sustainability, and governance has underscored the socio-technical nature of this endeavor, emphasizing that the success of such systems depends as much on their ethical and regulatory alignment as it does on their technical performance.

The transition toward agentic, reasoning-based forecasting represents a fundamental shift in the landscape of quantitative finance. By integrating qualitative context with numerical

precision, we can build infrastructures that are more adaptive, interpretable, and resilient. However, this evolution also brings new risks that must be carefully managed through proactive governance and robust system design. As we move closer to a world of autonomous financial agents, the unified cloud-edge model provides the necessary foundation for a more intelligent and stable global economy. The future of finance lies in the synergy of speed and reason, and the infrastructure described here is a vital component of that future.

References

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 308-318.

2. Acharya, V. V., & Richardson, M. (2009). Causes of the financial crisis. Critical Review, 21(2-3), 195-210.

3. Arumugam, R., & Bhargavi, R. (2019). A survey on modern trainable systems for time series forecasting. IEEE Access, 7, 70113-70135.

4. Bommasani, R., et al. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

5. Brown, T., et al. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901.

6. Cartea, A., Jaimungal, S., & Penalva, J. (2015). Algorithmic and High-Frequency Trading. Cambridge University Press.

7. Chen, L., & Zheng, Z. (2023). LLM-augmented financial analysis: Challenges and opportunities. Journal of Financial Data Science, 5(4), 12-28.

8. Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified data processing on large clusters. Communications of the ACM, 51(1), 107-113.

9. Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica, 987-1007.

10. Ghoshal, B., & Tucker, A. (2022). Scalable inference for deep learning in finance. Quantitative Finance, 22(10), 1845-1860.

11. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

12. Goyal, N., et al. (2023). High-throughput inference for large language models: A systems perspective. ACM SIGOPS Operating Systems Review, 57(1), 45-56.

13. Hendershott, T., Jones, C. M., & Menkveld, A. J. (2011). Does algorithmic trading improve liquidity? The Journal of Finance, 66(1), 1-33.

14. Kaplan, J., et al. (2020). Scaling laws for neural language models. arXiv preprint arXiv:2001.08361.

15. Kirilenko, A. S., et al. (2017). The Flash Crash: High-frequency trading in an electronic market. The Journal of Finance, 72(3), 967-998.

16. Lo, A. W. (2017). Adaptive Markets: Financial Evolution at the Speed of Thought. Princeton University Press.

17. Lopez de Prado, M. (2018). Advances in Financial Machine Learning. Wiley.

18. Liu, T. (2026). PCA-APT Stress Index for Market Drawdowns.

19. Narayanan, D., et al. (2019). PipeDream: Generalized pipeline parallelism for DNN training. Proceedings of the 27th ACM Symposium on Operating Systems Principles.

20. O'Neil, C. (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown.

21. Pasquale, F. (2015). The Black Box Society: The Secret Algorithms That Control Money and Information. Harvard University Press.

22. Rajbhandari, S., et al. (2020). ZeRO: Memory optimizations toward training trillion parameter models. SC20: International Conference for High Performance Computing, Networking, Storage and Analysis.

23. Shalf, J. (2020). The future of computing beyond Moore's Law. Philosophical Transactions of the Royal Society A, 378(2166).

24. Stoica, I., et al. (2017). Ray: A distributed framework for emerging AI applications. 13th USENIX Symposium on Operating Systems Design and Implementation.

25. Vaswani, A., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

26. Shih, K., Deng, Z., Chen, X., Zhang, Y., & Zhang, L. (2025, May). DST-GFN: A Dual-Stage Transformer Network with Gated Fusion for Pairwise User Preference Prediction in Dialogue Systems. In 2025 8th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE) (pp. 715-719). IEEE.

27. Wu, S., et al. (2023). BloombergGPT: A large language model for finance. arXiv preprint arXiv:2303.17564.

28. Zaharia, M., et al. (2012). Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. 9th USENIX Symposium on Networked Systems Design and Implementation.

29. Zhang, L., et al. (2021). Deep reinforcement learning for automated stock trading: An ensemble strategy. SSRN Electronic Journal.

30. Zhou, Y., et al. (2022). Mixture-of-experts with exponential selection. arXiv preprint arXiv:2202.08906.

31. Kaplan, J. D., & McCandlish, S. (2020). Scaling laws for neural language models. arXiv preprint arXiv:2001.08361.