Optimizing Multi-Modal Financial Intelligence via Resource-Aware Distributed Scheduling for Large Language Model Enhanced Time Series Inference Pipelines

Alistair Vance
Department of Computer Science and Engineering, University of North Texas
avant@unt.edu

Matthew Pennington
Department of Information Systems, Virginia Commonwealth University
m.pennington@vcu.edu

Abstract

The modern financial landscape is increasingly defined by the synthesis of high-frequency numerical data and unstructured linguistic context. As Large Language Models (LLMs) evolve from general-purpose assistants to specialized reasoning engines, their integration into time series forecasting pipelines has become a primary objective for institutional financial intelligence. However, the computational intensity of transformer-based architectures introduces significant latency and resource contention within distributed systems, often rendering real-time inference unfeasible. This paper explores the architectural requirements and systemic optimizations necessary for multi-modal financial intelligence systems. We propose a resource-aware distributed scheduling framework designed specifically to manage the heterogeneous demands of concurrent time series analysis and LLM-driven semantic synthesis. By examining the structural trade-offs between inferential precision and execution throughput, the research highlights the necessity of hardware-aware orchestration in large-scale financial deployments. The discussion extends to the socio-technical implications of such systems, focusing on algorithmic governance, environmental sustainability, and the critical need for robustness in volatile market environments. Through a comprehensive system-level analysis, we demonstrate how optimized scheduling protocols can mitigate the bottleneck of cross-modal data fusion, ensuring that financial intelligence remains both semantically deep and temporally relevant. The paper concludes with an examination of the policy and ethical frameworks required to govern autonomous financial agents in a globalized, multi-modal economy.

Keywords

Financial Intelligence, Distributed Scheduling, Large Language Models, Time Series Inference, Multi-Modal Systems, Resource-Aware Computing, Socio-Technical

Infrastructure.

1. Introduction

The digital transformation of global capital markets has led to an unprecedented convergence of disparate data modalities. Historically, financial forecasting was partitioned into two distinct silos: quantitative analysis, which focused on the statistical properties of numerical time series, and fundamental analysis, which relied on the qualitative interpretation of news, reports, and geopolitical events. The emergence of Large Language Models has fundamentally disrupted this binary by providing a bridge between linguistic reasoning and numerical pattern recognition. Today, "financial intelligence" is increasingly defined as the ability to perform real-time synthesis of these modalities to identify alpha that remains invisible to purely frequentist models. However, the move toward LLM-enhanced time series inference introduces a massive expansion in computational complexity.

Large-scale systems designed for financial inference must operate under extreme temporal constraints while managing the non-trivial memory and compute footprints of multi-billion parameter models. In a distributed environment, the challenge is not merely one of capacity but of orchestration. Financial workflows are inherently bursty and sensitive to tail latency, as a delay of a few milliseconds in a forecasting pipeline can result in significant financial loss or the degradation of systemic stability. Consequently, the development of resource-aware distributed scheduling becomes the central engineering challenge. This research investigates how scheduling protocols can be optimized to balance the heavy, long-running nature of LLM token generation with the lightweight, high-velocity requirements of numerical time series processing.

Beyond the immediate technical hurdles, the deployment of multi-modal financial intelligence carries significant socio-technical weight. As these systems become integrated into the core infrastructure of global finance, they influence market liquidity, price discovery, and the distribution of economic risk. The robustness of the underlying distributed system is no longer just a matter of corporate efficiency but a prerequisite for financial stability. This paper adopts an interdisciplinary perspective, examining the engineering specificities of distributed scheduling alongside the broader implications for governance, sustainability, and fairness in algorithmic decision-making. By aligning high-throughput system design with the nuances of financial reasoning, we propose a blueprint for a more resilient and context-aware financial infrastructure.

2. Conceptual Foundations of Multi-Modal Financial Synthesis

Multi-modal financial synthesis represents a shift from correlation-based forecasting to causal-augmented reasoning. Traditional time series models are frequently hindered by the "non-stationarity" of financial data, where historical patterns fail to persist during regime shifts. LLMs provide the semantic context necessary to identify the "why" behind a numerical movement, such as a sudden spike in volatility following a central bank announcement. The integration of these models into inference pipelines allows for a richer representation of the market state, where numerical features are contextualized by linguistic narratives. This

conceptual fusion, however, requires a system architecture that can maintain causal consistency across disparate data streams.

The primary architectural challenge in this synthesis is the "temporal alignment" of modalities. Numerical data flows at microsecond intervals, whereas linguistic context—such as news articles or social media sentiment—arrives sporadically and requires significant processing time to be converted into a machine-readable embedding. A naive system would experience "stale context," where the LLM's interpretation lags behind the current numerical price action, leading to erroneous forecasts. Optimized pipelines must therefore utilize speculative execution and predictive buffering to ensure that the semantic layer is ready at the exact moment the numerical model requires a context-aware update. This requires a scheduler that understands the temporal interdependencies of the modalities.

Furthermore, the reasoning capabilities of LLMs in finance extend beyond simple sentiment analysis to complex scenario simulation. An LLM-enhanced pipeline can simulate the impact of hypothetical geopolitical events on a specific portfolio, providing a layer of "narrative stress testing" that was previously manual and labor-intensive. From a systems perspective, this necessitates a high degree of parallelization and the ability to manage "multi-tenant" reasoning tasks without exhausting the memory bandwidth of the distributed cluster. The conceptual goal is a system where numerical signals and linguistic narratives exist in a continuous feedback loop, facilitated by an underlying infrastructure that prioritizes the most information-dense tasks.

3. Distributed Infrastructure for LLM-Enhanced Inference

The physical and logical architecture of a financial inference system must be designed for maximum hardware utilization and minimum inter-node communication overhead. We propose a tiered distributed infrastructure that partitions the compute cluster into specialized reasoning zones. The first zone consists of "High-Velocity Edge Nodes" responsible for real-time numerical feature extraction and low-latency statistical modeling. These nodes utilize specialized hardware such as Field-Programmable Gate Arrays (FPGAs) to minimize jitter. The second zone involves "Deep Reasoning Clusters" where the large language models reside. These clusters are composed of high-bandwidth memory (HBM) enabled GPUs and AI accelerators, optimized for the massive matrix multiplications required by the transformer architecture.

Orchestrating these zones requires a unified resource-aware scheduler that treats the entire cluster as a single compute fabric. Traditional schedulers, which operate on a "first-come, first-served" or simple priority basis, are inadequate for financial multi-modality because they fail to account for the "asymmetric compute profiles" of the tasks. Token generation in an LLM is a sequential, memory-bound process, while time series inference is typically a parallel, compute-bound task. Our proposed scheduler utilizes a "dynamic affinity" protocol, where tasks are mapped to nodes based on the real-time availability of memory bandwidth and interconnect speed. This minimizes the "data movement bottleneck" that often plagues large-scale distributed AI.

The robustness of this infrastructure is maintained through a decentralized health-check protocol. In a financial environment, a single node failure cannot be allowed to halt the inference pipeline. The system utilizes "logical replication," where critical reasoning paths are duplicated across different power and networking domains. If a node in the Deep Reasoning Cluster experiences a latency spike, the scheduler automatically redirects the semantic synthesis task to a redundant node with minimal state transfer overhead. This structural resilience ensures that the financial intelligence engine remains "always-on," providing the continuous stream of insights required for institutional risk management and autonomous trading.

4. Resource-Aware Distributed Scheduling Protocols

Resource-aware scheduling in the context of multi-modal finance is fundamentally about managing the "contention" for shared hardware resources. When a large-scale time series forecasting task and a multi-billion parameter LLM inference task arrive simultaneously at a compute node, the scheduler must decide how to partition the GPU kernels and CPU threads to ensure that neither modality breaches its latency budget. We introduce a "weighted-token-throughput" scheduling algorithm that prioritizes tasks based on their contribution to the final forecast's confidence interval. If the numerical signal is high-confidence but the linguistic context is ambiguous, the scheduler may de-prioritize the LLM task to save resources for other, more critical numerical streams.

A critical innovation in our scheduling framework is the use of "predictive resource pre-emption." By analyzing the historical arrival patterns of financial news and market data, the scheduler can anticipate periods of high computational demand—such as the opening of a major exchange or the release of a significant economic report. During these "high-alpha" periods, the scheduler pre-emptively clears lower-priority background tasks and warms up the LLM caches to ensure that the inference pipeline can handle the incoming data surge with zero cold-start latency. This proactive management of the compute fabric allows the system to maintain high throughput even during the most volatile market conditions.

Furthermore, the scheduler implements "quantization-aware task placement." Not all financial reasoning tasks require the full precision of an FP32 or FP16 model. Tasks such as high-level sentiment scanning can often be performed using INT8 or even INT4 quantized models without a significant loss in accuracy. Our scheduler identifies the "precision requirement" of each incoming query and routes it to the most energy-efficient hardware capable of meeting that requirement. This not only increases the overall throughput of the system but also contributes to the environmental sustainability of the infrastructure by reducing the total wattage consumed per inference.

5. Structural Trade-offs: Throughput vs. Latency vs. Precision

The engineering of a multi-modal financial system is a constant negotiation between three competing objectives: throughput (how many forecasts per second), latency (how fast is each forecast), and precision (how accurate is the reasoning). Increasing the precision of an

LLM—for example, by increasing the context window or utilizing a more complex attention mechanism—invariably increases latency. In a market where opportunity windows exist for only a few milliseconds, a highly precise but slow model is effectively useless. Conversely, a fast but shallow model may miss subtle causal drivers, leading to significant financial losses. Our system addresses this through a "dynamic precision scaling" mechanism.

In this paradigm, the system operates in a "base state" with optimized, quantized models that provide rapid, high-throughput inference. However, when the system detects a "divergence signal"—where the numerical time series contradicts the expected narrative—it automatically triggers a "deep reasoning" path. In this state, the scheduler reallocates resources to run a full-precision, high-context LLM to resolve the ambiguity. This "precision-on-demand" approach allows the system to maintain the high throughput necessary for general market monitoring while reserving deep, high-latency reasoning for the most complex and valuable financial events. This structural flexibility is essential for navigating the non-linear dynamics of modern finance.

Another critical trade-off is between "centralized coherence" and "decentralized agility." A centralized scheduler provides a globally optimal allocation of resources but introduces a single point of failure and significant networking overhead. A fully decentralized scheduler allows each node to make independent decisions, providing high agility but leading to "resource fragmentation" where some nodes are overloaded while others are idle. Our framework utilizes a "hierarchical scheduling" architecture, where a central coordinator sets high-level resource budgets and priorities, but individual cluster controllers manage the micro-second-level task placement. This hybrid approach balances the need for systemic stability with the requirement for localized responsiveness.

6. Hardware-Aware Orchestration and Sustainability

The environmental footprint of large-scale financial AI is a growing concern for both policymakers and institutional investors. The continuous operation of GPU clusters for LLM inference consumes immense amounts of electricity and water for cooling. A sustainable financial infrastructure must therefore move beyond simple performance optimization to focus on "energy-aware computing." Our proposed orchestration layer tracks the "performance-per-watt" of every node in the distributed system. By utilizing "carbon-aware scheduling," the system can shift non-urgent reasoning tasks—such as back-testing or long-term trend analysis—to data centers located in regions with a higher percentage of renewable energy on the grid.

Furthermore, hardware-aware orchestration involves the strategic use of "heterogeneous acceleration." While GPUs are the current standard for LLM training and inference, specialized AI accelerators and Neural Processing Units (NPUs) offer significantly better energy efficiency for specific transformer operations. Our system identifies the "op-code signature" of each reasoning task and offloads specific layers of the LLM to the hardware that can execute them with the lowest energy cost. This granular management of the hardware stack allows the financial intelligence engine to scale its capabilities without a proportional

increase in its carbon footprint, aligning corporate profitability with global sustainability goals.

Sustainability also extends to the "lifecycle of data." The storage and retrieval of massive multi-modal financial archives for model fine-tuning and retrieval-augmented generation (RAG) require significant disk and network resources. Our system implements a "tiered storage orchestration" protocol, where frequently accessed "hot" data remains in NVMe memory near the compute nodes, while historical "cold" data is moved to energy-efficient object storage. By minimizing the movement of data across the network and optimizing the storage hierarchy, the system reduces the overall energy overhead of the financial intelligence pipeline, ensuring its viability in an increasingly resource-constrained world.

7. Algorithmic Governance and Robustness

The deployment of autonomous multi-modal agents in finance introduces profound questions of algorithmic governance. If an LLM-enhanced forecasting pipeline triggers a massive liquidation that contributes to a market flash-crash, the responsibility must be traceable back to the system's design and logic. We argue for the implementation of "transparency-by-design" within the distributed scheduler. Every scheduling decision, model version, and data input must be recorded in a tamper-proof, high-performance audit log. This "forensic infrastructure" allows regulators and internal auditors to reconstruct the exact state of the system leading up to a significant market event, providing the accountability necessary for institutional trust.

Robustness in this context is not just about avoiding system crashes; it is about "semantic robustness." LLMs are known to be susceptible to "hallucinations" or the misinterpretation of subtle linguistic cues. In a financial setting, an LLM misinterpreting a "hawkish" central bank tone as "dovish" could have catastrophic consequences. Our framework mitigates this through a "cross-modal validation" layer. Before an LLM-derived insight is allowed to influence a high-stakes financial decision, it must be validated against the numerical time series. If the linguistic narrative and the numerical signal are fundamentally unaligned, the system flags the result for human intervention or triggers a secondary reasoning path. This "redundant logic" ensures that the intelligence remains grounded in empirical reality.

Governance also requires the management of "adversarial multi-modality." As these systems become public knowledge, bad actors may attempt to manipulate markets by injecting "semantic noise"—such as fake news or coordinated social media campaigns—designed to trick LLM-enhanced pipelines. A robust governance framework must include "informational sanitization" protocols that evaluate the credibility and provenance of linguistic data before it is ingested into the inference engine. By integrating adversarial detection into the distributed scheduler, we can ensure that the financial infrastructure remains resilient not only to technical failures but also to the sophisticated manipulation of information.

8. Socio-Technical Implications of Autonomous Financial Intelligence

The transition to autonomous, multi-modal financial intelligence is a socio-technical

transformation that redefines the relationship between capital, information, and human agency. As LLMs take over the task of narrative interpretation, the role of the human financial analyst shifts from "interpretation" to "oversight." This creates a "de-skilling" risk, where the next generation of financial professionals may lack the ability to perform independent critical analysis when the system fails. To counter this, we propose the integration of "human-in-the-loop" training within the systems themselves, where analysts are required to periodically "audit" the machine's reasoning process as part of their standard workflow.

The socio-technical perspective also highlights the risk of "algorithmic monoculture." If a small number of optimized, high-throughput LLM architectures come to dominate global financial forecasting, their shared biases and failure modes could lead to highly correlated market movements and systemic instability. To ensure a healthy and diverse financial ecosystem, policy frameworks should encourage "model diversity" and the use of open-source infrastructures. By preventing the concentration of reasoning power in a few proprietary "black boxes," we can maintain the market's ability to act as a decentralized information processor.

Finally, we must consider the "fairness" of access to multi-modal intelligence. The computational cost of running these high-throughput distributed systems is immense, creating a "compute divide" between large institutional players and smaller retail investors or emerging market participants. If the most advanced reasoning tools are only available to the wealthiest actors, it could lead to an even greater concentration of wealth and a degradation of market equity. Policy-makers must explore the creation of "public compute utilities" or shared financial intelligence infrastructures that provide democratized access to the tools of modern finance, ensuring that the benefits of the AI revolution are broadly distributed.

9. Policy, Ethics, and the Future of Financial Agents

The future of financial intelligence lies in the evolution of "autonomous agentic swarms"—networks of distributed LLMs and time series models that can negotiate, trade, and manage risk with minimal human intervention. Governing these swarms requires a fundamental shift from regulating "entities" to regulating "interactions." Policy frameworks must focus on the "behavioral signatures" of autonomous agents, setting clear boundaries for market conduct and ensuring that agents do not inadvertently engage in collusive or manipulative behaviors that undermine market integrity. This requires a new level of technical literacy among financial regulators and the development of "automated compliance" tools that can monitor agent activity in real-time.

Ethics in this domain must address the "moral status" of automated financial decisions. When a system optimizes for profit at the expense of social welfare—for example, by triggering a debt crisis in a vulnerable economy through aggressive short-selling—the ethical failure is a system design failure. We advocate for the inclusion of "pro-social objective functions" within the distributed scheduler. Instead of merely optimizing for throughput or alpha, the system could be required to optimize for "market stability" or "liquidity provision" during periods of stress. By encoding ethical values directly into the resource-aware protocols, we can ensure

that financial intelligence serves the broader goals of human society.

In conclusion, the optimization of multi-modal financial intelligence is a multidisciplinary endeavor that requires the seamless integration of high-performance systems engineering, advanced machine learning, and rigorous socio-technical governance. By developing resource-aware distributed scheduling protocols that can manage the complex interdependencies of numbers and narratives, we can build a financial infrastructure that is both technologically superior and socially responsible. The path forward is one of continuous adaptation, where the resilience of our systems is matched by the depth of our ethical commitments. As we stand at the threshold of the autonomous finance era, our task is to ensure that the "bridge" between modalities remains a stable foundation for global prosperity.

10. Conclusion

This research has demonstrated that the future of institutional financial intelligence depends on the successful integration of Large Language Models into time series inference pipelines through resource-aware distributed scheduling. We have proposed a tiered infrastructure and a hardware-aware orchestration framework that addresses the massive computational and temporal challenges inherent in multi-modal synthesis. By navigating the structural trade-offs between precision and latency, our proposed system ensures that deep semantic reasoning can coexist with the high-velocity requirements of global markets.

Beyond the technical optimizations, we have emphasized that the robustness, sustainability, and fairness of these systems are essential for their social legitimacy and systemic stability. The deployment of autonomous financial agents requires a new paradigm of algorithmic governance and policy intervention that prioritizes transparency, accountability, and the democratization of access to intelligence. As the boundaries between linguistic narrative and numerical signal continue to blur, the infrastructures we build today will determine the resilience and equity of the global financial ecosystem for decades to come. The synergy of human oversight and machine reasoning, facilitated by optimized distributed systems, remains our most powerful tool for navigating the complexities of the modern economic world.

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. Fu, L., Chen, X., Gao, K., Huang, X., & Tong, K. (2025, October). Memory-Augmented Knowledge Fusion with Safety-Aware Decoding for Domain-Adaptive Question Answering. In 2025 6th International Conference on Machine Learning and Computer Application (ICMLCA) (pp. 1-6). IEEE.

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

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

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

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

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

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

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

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

27. 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.

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

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

30. Mo, F., Haddadi, H., Katiyar, K., Ansari, R., & Chuah, C. N. (2021). PPFL: Privacy-preserving federated learning with trusted execution environments. Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services, 94-108.

31. Wang, J., et al. (2021). A field guide to federated optimization. arXiv preprint arXiv:2107.06917.

32. Rothchild, D., et al. (2020). FetchSGD: Communication-efficient federated learning with sketching. Proceedings of the 37th International Conference on Machine Learning.

33. Kairouz, P., et al. (2021). Advances and open problems in federated learning. Foundations and Trends in Machine Learning, 14(1-2), 1-210.

34. Yi, X. (2026). Privacy-Enhanced Ad Targeting for Social E-Commerce: A Federated Learning Framework with Zero-Knowledge Verification for Creator Monetization. Frontiers in Business and Finance, 3(1), 102-113.

35. Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology (TIST), 10(2), 1-19.

36. Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., & Chandra, V. (2018). Federated learning with non-iid data. arXiv preprint arXiv:1806.00582.

37. Zuboff, S. (2019). The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. PublicAffairs.