Fin-LLM-Inference: A High-Throughput Distributed System for Real-Time Financial Time Series Forecasting via Heterogeneous LLM-Augmented Reasoning Pipelines

Brandon Prescott
Department of Electrical Engineering and Computer Science, University of New Mexico
b.prescott@unm.edu

Christopher Sinclair
Department of Management Information Systems, University of Delaware
christopher.s@udel.edu

Abstract

The integration of Large Language Models (LLMs) into financial time series forecasting represents a transformative shift from purely frequentist econometric models to context-aware reasoning systems. However, the high-throughput requirements of modern capital markets create a significant tension with the computational latency inherent in transformer-based architectures. This paper introduces Fin-LLM-Inference, a high-throughput distributed system designed for real-time financial forecasting using heterogeneous LLM-augmented reasoning pipelines. We propose a multi-tiered architecture that strategically partitions reasoning tasks between optimized edge-based distilled models and robust cloud-based reasoning engines. By aligning hardware-aware optimizations with the unique non-stationarity of financial data, the system achieves a balance between predictive depth and execution speed. Our analysis focuses on the system-level trade-offs involving inference latency, model consistency, and architectural robustness. Furthermore, we examine the socio-technical implications of deploying such systems, including algorithmic governance, environmental sustainability in massive-scale AI clusters, and the policy challenges associated with automated financial decision-making. We argue that the future of financial intelligence lies in the seamless coordination of heterogeneous compute resources that can interpret both microstructure signals and macroeconomic narratives. This research provides a comprehensive blueprint for the next generation of resilient, scalable, and fair financial AI infrastructure, concluding with a forward-looking discussion on the regulatory landscape for autonomous financial agents.

Keywords

Distributed Systems, Financial Time Series, Large Language Models, Heterogeneous Computing, High-Throughput Inference, Socio-Technical Infrastructure, Algorithmic Governance.

1. Introduction

The digital transformation of global financial markets has reached a critical juncture where the velocity of data generation often outpaces the cognitive capacity of traditional analytical frameworks. For decades, financial time series forecasting was dominated by autoregressive models and stochastic differential equations that treated market movements as purely numerical phenomena. While these methods were effective in stable regimes, they consistently failed during black-swan events or periods where the underlying market narrative underwent a fundamental shift. The emergence of Large Language Models has introduced a new dimension to this field: the ability to perform qualitative reasoning over unstructured data, such as news cycles, central bank communications, and social sentiment, and to synthesize these insights with quantitative market microstructure data.

However, the deployment of LLMs in a real-time financial environment is fraught with systemic challenges. Financial markets operate on millisecond scales, whereas LLM inference, even with modern hardware acceleration, often incurs latencies that render the insights obsolete by the time they are generated. This "latency-utility gap" necessitates a fundamental rethinking of the underlying distributed systems architecture. It is no longer sufficient to treat the LLM as a static oracle; instead, it must be integrated into a high-throughput pipeline that can dynamically allocate reasoning tasks across a heterogeneous array of compute resources. This paper proposes the Fin-LLM-Inference framework as a solution to this architectural bottleneck, emphasizing the coordination between edge-based speed and cloud-based analytical depth.

The motivations for this research extend beyond mere technical optimization. As financial systems become more autonomous, the socio-technical implications of their design—ranging from environmental sustainability to algorithmic fairness—become paramount. A system that optimizes only for throughput while ignoring the robustness of its reasoning or the fairness of its outcomes risks introducing systemic instability into the global economy. Consequently, this paper provides a deep analytical discussion on the governance and policy requirements for high-throughput financial AI. By treating the system as a socio-technical infrastructure rather than a purely engineering artifact, we aim to establish a more resilient foundation for the future of financial intelligence.

2. The Architectural Paradigm of Heterogeneous Reasoning

Traditional AI systems in finance often rely on a monolithic approach, where a single model is expected to handle all aspects of the forecasting task. In contrast, the Fin-LLM-Inference system adopts a heterogeneous architectural paradigm. This approach recognizes that not all financial reasoning tasks are created equal. Identifying a trend in high-frequency tick data requires rapid, low-latency execution that is best suited for distilled, specialized models. Conversely, interpreting the long-term implications of a change in interest rate policy requires the deep, multi-hop reasoning capabilities of a full-scale LLM. The core of our system is a dynamic orchestration layer that partitions these tasks based on their semantic complexity and temporal requirements.

This tiered structure allows for a "fast-path" and a "slow-path" for financial intelligence. The

fast-path is deployed at the edge, utilizing hardware-accelerated processors to perform real-time feature extraction and local reasoning. The slow-path resides in a distributed cloud environment, where larger model ensembles perform exhaustive context synthesis. The synchronization of these tiers is handled through a speculative reasoning protocol, where the edge models generate initial forecasts that are subsequently verified or refined by the cloud-based engines. This reduces the average latency of the system while maintaining a high upper bound on reasoning quality. The trade-off here involves the overhead of maintaining consistency between the heterogeneous tiers, a challenge we address through a versioned state management system.

The structural robustness of this heterogeneous pipeline is further enhanced by its ability to handle non-stationary data. Financial markets are notoriously prone to regime shifts, where the statistical properties of the data change abruptly. A monolithic model often struggles to adapt to these shifts in real-time. By utilizing a heterogeneous pipeline, our system can swap out specialized edge models as the market regime changes, while the cloud-based reasoning layer provides a stable anchor of general economic knowledge. This modularity ensures that the system remains performant across a wide range of market conditions, providing a more reliable forecasting mechanism for institutional deployments.

3. Distributed Infrastructure and Hardware-Aware Optimization

The physical deployment of Fin-LLM-Inference requires a highly specialized distributed infrastructure that is aware of the underlying hardware topography. In the context of high-throughput finance, the bottleneck is often the memory bandwidth and the serialization latency of the model weights. To mitigate these issues, we utilize a hardware-aware model sharding strategy that distributes the LLM layers across multiple GPU clusters based on their interconnect bandwidth. By aligning the model's computational graph with the physical network of the data center, the system minimizes the "all-reduce" overhead that typically slows down large-scale inference.

Beyond the data center, the integration of edge compute nodes—located in proximity to financial exchanges—is a critical component of the infrastructure. These edge nodes act as the first line of reasoning, performing high-speed linguistic and numerical alignment. We utilize aggressive quantization and model pruning to fit the necessary reasoning capabilities into the constrained thermal and power envelopes of edge hardware. This edge-to-cloud continuum is managed by a centralized control plane that monitors the health and latency of every node in real-time. If an edge node experiences a spike in latency or a hardware failure, the control plane automatically re-routes the reasoning task to a neighboring node or escalates it to the cloud, ensuring high availability.

Sustainability is a primary consideration in this infrastructure design. The energy consumption of continuous LLM inference at scale is a significant concern for both cost and environmental impact. Fin-LLM-Inference addresses this by implementing a "carbon-aware" scheduling policy. The system modulates its reasoning intensity based on the immediate commercial value of the forecast and the availability of renewable energy at the data center

location. During periods of low volatility, the system can revert to more energy-efficient models, only spinning up the massive inference clusters when the market complexity demands it. This approach demonstrates that high-throughput intelligence can be reconciled with environmental stewardship through intelligent system-level management.

4. Real-Time Data Pipelines and Narrative Synthesis

Financial time series are no longer just strings of numbers; they are embedded in a complex web of global narratives. The Fin-LLM-Inference system treats narrative data as a primary input modality, equivalent to price and volume signals. This requires a real-time data pipeline that can ingest, de-noise, and vectorize vast quantities of unstructured text from global news feeds, social media, and regulatory filings. The challenge here is one of semantic alignment: the system must be able to map a geopolitical event described in a news report to a specific set of anomalies in the market microstructure. We achieve this through a cross-modal embedding space that allows the LLM to "see" the relationship between text and numbers.

The reasoning pipeline performs what we term "narrative synthesis," where multiple conflicting sources of information are reconciled into a coherent market outlook. For instance, a bullish news cycle might be contradicted by bearish signals in the order book. Traditional models would often ignore the text or treat it as a simple sentiment score. Fin-LLM-Inference uses its reasoning engine to ask why the signals are conflicting. Is the news cycle lagging the market? Is there a hidden liquidity drain? This multi-step reasoning provides a much deeper level of intelligence than simple correlation-based forecasting. It allows the system to detect not just the movement, but the causal driver of the movement.

Robustness in the data pipeline is maintained through an adversarial filtering layer. Financial information streams are often intentionally manipulated by actors seeking to trigger specific algorithmic responses. Our system employs a secondary LLM agent that acts as a "sanity checker," evaluating the truthfulness and consistency of the incoming narrative stream. By comparing new information against a historical knowledge base and cross-referencing multiple independent sources, the system can filter out high-entropy noise or malicious disinformation. This layer of semantic security is essential for ensuring that the forecasting system remains reliable in an increasingly contested information environment.

5. Algorithmic Governance and the Transparency-Depth Trade-off

As the reasoning pipelines in financial AI become more complex, the challenge of algorithmic governance becomes more acute. Regulators and institutional compliance teams require transparency into how decisions are made, yet the inner workings of a multi-billion parameter LLM are notoriously opaque. Fin-LLM-Inference addresses this through a "transparency-by-design" approach, where the system generates a human-readable reasoning trace for every significant forecast. This reasoning trace explains the logic the model used to arrive at its conclusion, citing specific data points from both the microstructure and narrative streams.

This creates a structural trade-off between reasoning depth and interpretability. Often, the

most accurate forecasts are the result of highly non-linear associations that are difficult to express in natural language. Conversely, forcing the model to be highly interpretable can sometimes limit its predictive power. Our system manages this trade-off by maintaining two parallel reasoning paths. The primary path optimizes for predictive accuracy, while the secondary path, running in the background, attempts to provide a faithful approximation of that logic in human-understandable terms. This "dual-reasoning" architecture ensures that the system can meet the performance requirements of the market while satisfying the transparency requirements of the regulators.

Governance also extends to the management of model bias. LLMs trained on historical financial data may inadvertently inherit the biases of past market participants or reflect systemic inequities in the global financial system. To prevent these biases from being amplified by our high-throughput pipeline, we implement a continuous fairness auditing layer. This layer tests the model's reasoning against a set of synthetic scenarios designed to identify discriminatory patterns or unfair market advantages. By making fairness an objective function of the system's deployment, we ensure that the Fin-LLM-Inference framework contributes to a more equitable and stable financial ecosystem.

6. System Deployment, Scaling, and Resilience

The transition from a laboratory prototype to a production-scale financial infrastructure is a journey fraught with technical and operational risks. Deployment of the Fin-LLM-Inference system follows a "blue-green" strategy, where new model versions and architectural updates are first tested in a shadow environment that replicates real-time market conditions without executing actual trades. This allows the system to "prove" its robustness and accuracy before being given control over capital. The scaling of the system is handled through a decentralized orchestration model, where additional compute nodes can be added to the network without requiring a complete system reboot.

Resilience in a distributed financial system is not just about avoiding crashes; it is about maintaining a "graceful degradation" of service during periods of extreme stress. In a market crisis, the volume of data can increase by orders of magnitude, and the latency of the cloud interconnects can become unpredictable. Fin-LLM-Inference is designed to handle these conditions by automatically shedding non-essential reasoning tasks and falling back to its most efficient edge-based models. This ensures that even in the worst-case scenario, the system provides a continuous, albeit simplified, stream of intelligence. This resilience is a critical property for maintaining market stability, as a sudden failure of major AI agents could trigger a liquidity vacuum.

The deployment phase also involves the physical and logical security of the infrastructure. Financial AI systems are high-value targets for cyberattacks. We utilize a "zero-trust" security model where every communication between nodes, whether at the edge or in the cloud, is encrypted and verified using hardware-level security tokens. Furthermore, the model weights themselves are protected using a distributed "watermarking" technique that can detect unauthorized copying or tampering. By treating security as a fundamental part of the system's

architecture rather than an afterthought, we create a robust defense against the evolving threat landscape of the digital financial world.

7. Policy Implications and the Future of Autonomous Finance

The widespread adoption of high-throughput reasoning systems like Fin-LLM-Inference has significant implications for global financial policy. As autonomous agents become the primary drivers of market activity, the traditional rules of market conduct—designed for human participants—may become obsolete. For instance, the concept of "intent" in market manipulation is difficult to apply to an LLM that is simply optimizing for a complex reward function. This necessitates a shift from conduct-based regulation to infrastructure-based regulation. Policy-makers must focus on the design standards, data diets, and stress-testing requirements of the AI systems themselves.

There is also a risk of "systemic synchronization," where multiple institutions deploying similar LLM architectures react to the same narrative signals in the same way, leading to flash crashes or extreme volatility. To mitigate this, policy frameworks should encourage diversity in the AI ecosystem. Our heterogeneous architecture facilitates this by allowing for the integration of models from different providers and training datasets. Regulators might also consider "circuit breakers" for AI reasoning, where systems are required to pause or reduce their activity if they detect a lack of epistemic diversity in the market. This proactive approach to policy can help harness the benefits of AI while protecting the systemic stability of the economy.

Furthermore, the environmental cost of financial AI will likely lead to new regulations on the "carbon intensity" of automated trading. Institutions may be required to report the energy usage associated with their AI infrastructures and offset their emissions. Fin-LLM-Inference's "carbon-aware" scheduling is a direct response to this anticipated policy shift. By demonstrating that high-performance AI can be environmentally responsible, we hope to set a new standard for the industry. The future of autonomous finance is not just about who has the fastest model, but who can build the most sustainable, transparent, and fair system.

8. Socio-Technical Perspectives on LLMs in Global Capital Markets

The integration of LLMs into finance is a socio-technical event that alters the relationship between information, technology, and society. At its core, the Fin-LLM-Inference system is a tool for the "industrialization of reasoning." It takes the traditionally human task of interpreting the world and scales it to the speed of the digital exchange. This shift has profound implications for the labor market in finance, as the demand for traditional analysts may decrease while the demand for "reasoning engineers"—those who can design and govern these complex pipelines—will skyrocket.

Moreover, the global nature of the Fin-LLM-Inference infrastructure means that it must operate across diverse cultural and linguistic contexts. A narrative in an emerging market might be interpreted differently than one in a developed economy. Our system's ability to handle heterogeneous models allows it to incorporate "local" reasoning agents that understand

the specific nuances of different regions. This global-to-local synthesis is essential for ensuring that financial AI does not become a tool for "informational colonialism," where the logic of a few dominant economies is imposed on the rest of the world.

Finally, we must consider the "trust infrastructure" that underlies our system. For the Fin-LLM-Inference system to be successful, it must be trusted not just by its operators, but by the public and the regulators. This trust is built through a consistent track record of performance, transparency, and fairness. By prioritizing these values in the system's design, we are not just building a better forecasting tool; we are contributing to the long-term health and legitimacy of the global financial system. The socio-technical journey of LLMs in finance is just beginning, and the choices we make today in the design of our infrastructures will shape the economic landscape for decades to come.

9. Conclusion

This paper has presented Fin-LLM-Inference, a high-throughput distributed system designed to bridge the gap between real-time financial data and the deep reasoning capabilities of Large Language Models. Through a tiered, heterogeneous architecture that coordinates edge and cloud compute resources, we have demonstrated a path toward ultra-low-latency financial intelligence that does not sacrifice analytical depth. Our discussion has emphasized that the success of such a system is dependent not only on its engineering brilliance but also on its alignment with the principles of governance, sustainability, and fairness.

The Fin-LLM-Inference framework serves as a scalable blueprint for the next generation of financial infrastructure. It challenges the industry to move beyond monolithic models and simple sentiment scores, toward a more nuanced and causally-grounded understanding of market dynamics. As we move closer to a future of fully autonomous finance, the need for resilient, transparent, and socially responsible AI systems has never been greater. By treating the system as a complex socio-technical entity, we can ensure that the rise of machine intelligence in capital markets leads to a more stable, efficient, and equitable 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. 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.

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

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

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

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

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

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

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

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

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

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. 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.

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

33. Rothchild, D., et al. (2020). FetchSGD: Communication-efficient federated learning with

sketching. Proceedings of the 37th International Conference on Machine Learning.

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