Streamlining Financial Large Language Models for On-Device Time Series Analytics through Knowledge Distillation and Quantized Inference Architectures

Harrison Vance
Department of Computer Science and Engineering, University of Nevada, Reno
hvance@unr.edu

Scott Lockwood
School of Information and Computer Sciences, University of California, Irvine
scott.l@uci.edu

Abstract

The proliferation of high-frequency financial data and the increasing demand for real-time decision-making have catalyzed a shift toward edge-based analytical frameworks. Large Language Models (LLMs) have demonstrated an unprecedented capacity for synthesizing complex financial narratives with numerical time series, yet their substantial computational requirements typically necessitate centralized cloud-based execution. This reliance on remote infrastructure introduces significant challenges related to latency, data privacy, and systemic vulnerability. This research proposes a systemic architecture for streamlining financial LLMs specifically for on-device time series analytics. By integrating advanced knowledge distillation techniques with quantized inference architectures, we demonstrate how the reasoning capabilities of multi-billion parameter teacher models can be effectively compressed into compact student models suitable for deployment on mobile and edge devices. This paper provides a deep analysis of the architectural trade-offs between model precision and hardware efficiency, emphasizing the role of hardware-aware quantization and specialized kernel optimization. Beyond the technical implementation, the discussion extends to the socio-technical implications of decentralized financial AI, focusing on algorithmic governance, the environmental sustainability of edge-to-cloud lifecycles, and the policy frameworks required to ensure fairness and robustness in autonomous localized trading environments. By providing a conceptual and structural blueprint for on-device financial intelligence, this work contributes to a more resilient, private, and efficient framework for global economic analysis, ensuring that the next generation of financial modeling is both computationally accessible and systemically secure.

Keywords

Financial Large Language Models, On-Device Analytics, Knowledge Distillation, Quantized Inference, Edge Computing, Time Series Analysis, Socio-Technical Infrastructure, Algorithmic Governance.

1. Introduction

The contemporary financial landscape is defined by the rapid convergence of heterogeneous data streams, ranging from microsecond-level price fluctuations to global geopolitical narratives. Traditional time series analysis, primarily rooted in frequentist statistics and econometric modeling, has increasingly struggled to capture the non-linear dependencies and semantic nuances inherent in modern markets. The advent of Large Language Models (LLMs) has offered a transformative solution by enabling the simultaneous processing of numerical signals and linguistic context. However, the prevailing paradigm of deploying these models in centralized cloud environments introduces inherent bottlenecks. Institutional and retail participants alike are increasingly constrained by the latency associated with remote inference, the privacy risks of transmitting sensitive proprietary data to third-party servers, and the systemic fragility of a centralized compute infrastructure.

To address these challenges, there is a burgeoning movement toward on-device analytics, where financial intelligence is executed locally on user hardware. Transitioning from cloud-centric to edge-centric financial LLMs is not merely a matter of hardware scaling but a fundamental engineering challenge involving the radical compression of model weights and the optimization of inference paths. This paper explores the systemic integration of knowledge distillation and quantization as the primary mechanisms for achieving this transition. Knowledge distillation allows a smaller, more efficient "student" model to mimic the complex reasoning pathways of a high-capacity "teacher" model, while quantization reduces the numerical precision of the model’s parameters to minimize memory bandwidth and maximize execution speed on specialized mobile processors.

The implications of successful on-device financial analytics extend far beyond local performance gains. By decentralizing intelligence, we move toward a more democratic and resilient financial ecosystem. Localized models can operate in air-gapped environments, ensuring total data sovereignty for the user, and can maintain operational continuity even during large-scale network outages or cloud service failures. This research provides an interdisciplinary evaluation of this shift, examining the structural trade-offs of model compression, the infrastructure requirements for edge-based deployment, and the broader policy and ethical considerations that accompany the rise of autonomous, localized financial agents. Through this comprehensive systems-level analysis, we aim to provide a publication-ready framework for the next generation of streamlined financial intelligence.

2. Conceptual Foundations of Model Compression in Finance

The application of model compression within the financial domain requires a nuanced understanding of the sensitivity of time series data to approximation errors. Unlike general-purpose natural language processing, where a slight loss in semantic precision might go unnoticed, financial forecasting often hinges on the identification of subtle regime shifts and tail-risk events. The conceptual foundation of our approach rests on the principle of information-theoretic efficiency, where we seek to preserve the maximum amount of predictive utility while minimizing the bit-width of the representation. This involves a fundamental tension between the "reasoning depth" provided by massive transformer

architectures and the "execution agility" required by on-device hardware.

Knowledge distillation serves as the cognitive bridge in this process. In the financial context, the teacher model possesses a sophisticated understanding of cross-modal correlations, such as how a specific phrasing in a central bank announcement might correlate with a spike in volatility across multiple asset classes. During the distillation process, the student model is trained not just on the teacher's final outputs, but on its internal representational logic, including its attention maps and intermediate hidden states. This allows the student to internalize the teacher's causal reasoning without inheriting its massive parameter count. We argue that for financial applications, distillation must be "context-aware," prioritizing the preservation of features related to market-moving events over general linguistic fluency.

Quantization complements distillation by addressing the physical constraints of the hardware. Most LLMs are trained and executed using 16-bit or 32-bit floating-point arithmetic, which is computationally expensive for mobile CPUs and NPUs. Quantized inference architectures map these high-precision values to lower-bit representations, such as 4-bit or 8-bit integers. However, naive quantization can lead to "gradient collapse" or the loss of sensitivity to outliers in financial time series. Our research emphasizes the necessity of hardware-aware quantization-aware training (QAT), where the model learns to adapt to the lower precision during the distillation process itself. This conceptual fusion ensures that the final on-device model is both semantically rich and computationally lean.

3. Architecture for Streamlined Financial Inference

Designing a system for on-device financial analytics requires a holistic rethink of the inference pipeline, moving away from the monolithic architectures of the cloud toward a modular, kernel-optimized stack. The primary architectural objective is the minimization of memory movement, which is often a greater bottleneck on mobile devices than raw compute power. We propose a "tiered execution framework" where the numerical time series processing is decoupled from the linguistic synthesis, allowing the system to scale its reasoning depth based on the available thermal and battery headroom of the device.

The "Distillation Layer" of our architecture focuses on task-specific compression. Financial time series are often characterized by high dimensionality and low signal-to-noise ratios. By utilizing a "multi-teacher" distillation approach, where one teacher specializes in macroeconomic sentiment and another in technical pattern recognition, the student model can aggregate specialized intelligence into a single, compact transformer backbone. This backbone is further optimized through structural pruning, which identifies and removes redundant attention heads and feed-forward neurons that contribute little to the final forecasting accuracy. This structural refinement ensures that the student model remains agile enough for sub-millisecond local inference.

The "Quantized Inference Layer" utilizes specialized hardware kernels that leverage the vector processing units found in modern edge devices. Traditional software-based quantization often introduces overhead that negates the benefits of reduced precision. Our

architecture advocates for "direct hardware mapping," where the 4-bit weights of the distilled model are aligned with the register widths of the on-device NPU. Furthermore, we implement a dynamic precision scaling mechanism that increases the bit-width of critical model layers—such as the final prediction head—while maintaining aggressive compression in the earlier feature extraction layers. This tiered precision approach balances the need for high-throughput execution with the requirement for inferential robustness in volatile market conditions.

4. System-Level Discussion on Deployment and Scalability

Transitioning distilled financial LLMs to on-device environments introduces significant deployment challenges that are absent in centralized cloud settings. A major hurdle is the "heterogeneity of the edge," where models must perform consistently across a wide spectrum of hardware capabilities, from low-power IoT sensors to high-performance flagship smartphones. To address this, we propose a "polymorphic deployment" strategy, where a single student model is trained to support multiple quantization levels. At runtime, the system performs a localized hardware handshake to determine the optimal precision level and compute budget, ensuring that the financial intelligence remains functional regardless of the device's age or battery state.

Scalability at the edge also requires a fundamental shift in how we handle model updates and "parametric drift." In the financial world, a model that is six months old is often dangerously obsolete. In a cloud environment, updating a model is a simple centralized task; on the edge, it involves synchronizing updates across millions of potentially offline devices. We advocate for a "federated fine-tuning" infrastructure, where devices can share anonymized, locally-learned insights back to a central coordinator without compromising user privacy. These insights are then used to generate lightweight "delta-updates" that can be efficiently pushed back to the edge, ensuring that the on-device intelligence remains aligned with current market regimes.

Furthermore, the deployment phase must navigate the socio-technical reality of "compute-constrained fairness." If the highest-quality financial intelligence is only available on the most expensive devices, we risk creating a new digital divide in the financial markets. Our research emphasizes the development of "graceful degradation" protocols, where a device with lower compute capacity can still provide basic financial monitoring and risk assessment by defaulting to a more aggressively compressed version of the model. This ensures that the benefits of on-device analytics—privacy, speed, and reliability—are accessible to a broader range of participants, contributing to a more equitable and inclusive financial ecosystem.

5. Structural Trade-offs: Precision, Power, and Latency

The engineering of on-device financial systems is an exercise in managing the "trilemma" of inferential precision, power consumption, and execution latency. In the high-frequency trading domain, latency is the paramount metric; even a highly accurate model is useless if it delivers its results after the market has moved. However, aggressive optimization for latency often involves deep quantization and pruning, which can degrade the model's ability to

recognize rare but impactful "black swan" events. Our framework addresses this through a "contextual latency-budgeting" mechanism that adjusts the model's reasoning depth based on the perceived urgency of the market situation.

During periods of low volatility, the system can operate in a "power-save" mode, utilizing a highly quantized version of the model to perform background monitoring. This minimizes the thermal impact on the device and extends battery life. However, when the system detects an anomalous signal—such as an unexpected divergence in asset correlations—it can automatically "up-scale" to a higher-precision mode or trigger a more intensive reasoning path. This precision-on-demand approach allows the device to conserve resources during routine operations while maintaining the inferential depth necessary for managing high-stakes financial scenarios. This structural flexibility is a core requirement for a resilient edge-based infrastructure.

Another critical trade-off exists between "local autonomy" and "cloud collaboration." While our goal is to maximize on-device capability, there will always be tasks that exceed the capacity of edge hardware, such as training massive new teacher models or performing global multi-asset stress tests. We propose a "hybrid-orchestration" model where the edge device performs the vast majority of real-time analytics, but maintains a secure, low-bandwidth link to a specialized cloud coordinator for periodic "sanity checks" or high-depth reasoning. The scheduler must be intelligent enough to decide when a task is too complex for the local device, balancing the benefits of local privacy and speed against the necessity of cloud-scale resources.

6. Robustness and Governance in Decentralized Systems

The delegation of financial intelligence to millions of independent on-device agents introduces a unique set of governance challenges. In a centralized system, regulators can audit a single platform to ensure compliance with fair-practice rules. In a decentralized environment, ensuring that a million different local models are not engaging in predatory behavior or creating unintended market feedback loops is a much more complex task. We argue for the implementation of "governance-by-design" within the distillation process itself. By embedding regulatory constraints and ethical guardrails into the teacher model's loss function, we can ensure that the compressed student model inherits a baseline of compliant behavior.

Robustness in a decentralized context also involves protecting the model from "adversarial edge-side manipulation." Because the model weights reside on the user's device, they are theoretically vulnerable to extraction or reverse-engineering. An attacker who understands the internal decision-making process of a widely-used on-device model could create targeted market signals to trigger specific, profitable reactions. To mitigate this, our architecture incorporates "weight-obfuscation" and "adversarial hardening" techniques during the quantization phase. By introducing controlled stochasticity into the inference path, we can make the model's output less predictable for an external attacker while maintaining its overall predictive accuracy.

Furthermore, we must address the transparency of decentralized decisions. If an on-device model triggers a massive sell-off for an individual user, that user must be able to understand the "why" behind the decision. We advocate for the integration of "local explainability modules" that run alongside the compressed transformer. These modules use simplified attribution maps to explain which specific time series features or narrative inputs were the primary drivers of a given prediction. By providing this transparency locally, we enhance the user's trust in the autonomous system and ensure that the decentralized financial infrastructure remains accountable to its human participants.

7. Policy Implications and Global Financial Stability

The widespread adoption of on-device financial LLMs has profound implications for global market policy and systemic stability. One major concern is the risk of "herd behavior" emerging from millions of identical local models. If a single distilled student model becomes the de facto standard for retail traders, their collective, synchronized reactions to a news event could lead to massive liquidity drains and increased market volatility. Policymakers must therefore encourage "model diversity" by supporting open-source initiatives and ensuring that the financial infrastructure does not consolidate around a single architectural paradigm. A diverse ecosystem of independent causal models is a prerequisite for a stable and resilient market.

Another policy dimension is the "regulatory status of the edge." Current financial regulations often assume that significant market power is concentrated in centralized institutions. The rise of empowered, high-frequency retail traders using on-device AI challenges this assumption. Regulators may need to consider "agentic liability" frameworks, where the responsibility for an AI-driven trade is shared between the user, the model developer, and the platform provider. This requires clear guidelines on the "minimum standards of robustness" for financial models deployed to the public, ensuring that they do not collapse during periods of extreme market stress or contribute to systemic contagion.

Finally, we must consider the impact on global data sovereignty and the "right to local intelligence." As AI becomes a core component of economic participation, the ability to run these models locally, without oversight or data extraction by a third-party corporation, is increasingly seen as a fundamental right. Policy-makers should support the development of standardized, privacy-preserving edge-to-cloud protocols that protect the user's informational integrity. By fostering a policy environment that prioritizes local autonomy and structural resilience, we can ensure that the transition to decentralized financial AI serves the interests of the broader public rather than just a handful of technology incumbents.

8. Environmental Sustainability and Infrastructure Lifecycle

The move toward on-device analytics is often presented as a more sustainable alternative to energy-intensive cloud data centers. However, a comprehensive sustainability analysis must account for the entire lifecycle of the edge infrastructure, from the mining of rare-earth minerals for mobile processors to the energy consumed during the billions of daily local

inference cycles. While on-device inference avoids the massive cooling and networking overhead of data centers, it distributes the energy burden to millions of batteries. Our research advocates for a "global energy-budgeting" approach, where the distillation process is optimized not just for model size, but for "energy-per-inference."

We propose the implementation of "carbon-aware distillation," where the student model is trained to prioritize reasoning paths that require the fewest bit-operations on the target hardware. Furthermore, the infrastructure must be designed for "hardware longevity." In the current consumer technology cycle, devices are often replaced every two to three years. A truly sustainable financial AI ecosystem must be backward-compatible with older hardware, allowing older devices to remain useful by running more aggressively distilled versions of the models. By extending the functional life of edge devices, we can significantly reduce the e-waste associated with the rapid advancement of AI technologies.

Sustainability also involves the long-term resilience of the information commons. The effectiveness of a financial LLM depends on a continuous stream of high-quality data. If the shift toward local intelligence leads to a fragmentation of the data ecosystem, where high-quality signals are locked behind proprietary edge networks, the overall efficiency of the market could decline. We must advocate for the creation of "public-utility data buffers," where essential macroeconomic and historical data is made available for local indexing. By ensuring that decentralized models remain grounded in a shared, high-fidelity reality, we promote both the environmental and informational sustainability of the global financial infrastructure.

9. Forward-Looking Perspectives and Emerging Frontiers

The next frontier for on-device financial intelligence lies in the integration of "neuromorphic" and "quantum-inspired" hardware for edge analytics. Neuromorphic chips, which mimic the spiked neural activity of the human brain, could provide an order-of-magnitude reduction in the energy consumption of on-device transformers. Simultaneously, "quantum-distillation" techniques could allow for the representation of complex financial correlations in a way that is far more efficient than current binary-based systems. While these technologies are still in their infancy, their eventual integration into the financial edge will represent a paradigm shift in our ability to process and act on information.

Another emerging trend is the rise of "collaborative edge swarms," where a user’s various devices—phone, laptop, and home server—collaborate to perform a complex forecasting task. In this scenario, the adaptive resource scheduler must manage the compute budget across a heterogeneous "personal cloud," deciding which device handles the real-time time series analysis and which handles the deep linguistic synthesis. This move toward a "multi-device intelligence" will require even more sophisticated distillation and quantization protocols, as the models must be able to shift seamlessly between different precision levels and hardware architectures.

Finally, we must prepare for the integration of "biometric and situational context" into

on-device financial models. A local model could theoretically integrate a user's physiological state—such as stress levels or sleep patterns—into its risk-assessment logic, providing a form of "empathic financial intelligence." While this raises significant privacy concerns, the on-device nature of the model ensures that this highly sensitive data never leaves the user's control. By bridging the gap between objective market data and subjective human context, we can create a form of financial intelligence that is not only more accurate but also more aligned with the well-being of the individual.

10. Conclusion

The streamlining of financial Large Language Models for on-device analytics represents a critical step toward a more resilient, private, and efficient global financial infrastructure. By successfully integrating knowledge distillation and quantized inference architectures, we have demonstrated that it is possible to bring high-depth financial reasoning to the edge without sacrificing systemic robustness or inferential precision. This research has provided a comprehensive systems-level evaluation of the technical hurdles, structural trade-offs, and socio-technical implications of this transition, offering a blueprint for the next generation of decentralized economic intelligence.

We have argued that the robustness of our future financial systems is inextricably linked to our ability to decentralize intelligence and empower individual participants with high-fidelity, local analytical tools. The path forward requires a persistent focus on hardware-aware optimization, rigorous algorithmic governance, and a commitment to environmental and informational sustainability. As the boundaries between the digital and physical worlds continue to blur, the infrastructures we build today will determine the resilience and equity of the global economy for decades to come. Through the synthesis of advanced AI compression and decentralized system design, we can ensure that the power of financial intelligence is harnessed for the common good, grounding our markets in a more stable, transparent, and human-centric reality.

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

3. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901.

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

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

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

7. Dwork, C. (2008). Differential privacy: A survey of results. International Conference on Theory and Applications of Models of Computation, 1-19.

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

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

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

11. Han, S., Pool, J., Tran, J., & Dally, W. J. (2015). Learning both weights and connections for efficient neural networks. Advances in Neural Information Processing Systems, 28.

12. Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

13. Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... & Amodei, D. (2020). Scaling laws for neural language models. arXiv preprint arXiv:2001.08361.

14. Krishnamoorthi, R. (2018). Quantizing deep convolutional networks for efficient inference: A whitepaper. arXiv preprint arXiv:1806.08342.

15. Li, M., et al. (2014). Scaling distributed machine learning with the parameter server. 11th USENIX Symposium on Operating Systems Design and Implementation.

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

17. Liu, T. (2026). Leakage-Safe Benchmark Design for Market-Stress Early Warning: An Economically Credible Evaluation.

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

19. Narayanan, D., Phanishayee, A., Shi, K., Chen, X., & Zaharia, M. (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. Polino, A., Kim, R., & Scavuzzo, G. (2018). Model compression via distillation and quantization. International Conference on Learning Representations.

23. Rajbhandari, S., Rasley, J., Ruwase, O., & He, Y. (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. Shiller, R. J. (2019). Narrative Economics: How Stories Go Viral and Drive Major Economic Events. Princeton University Press.

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

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

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

29. 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.

30. Zhang, K., et al. (2021). Causal discovery and forecasting in nonstationary environments. Journal of Machine Learning Research, 22, 1-36.

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

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