Accelerating Real-Time Intent Discovery in Digital Advertising via High-Throughput Systems Integrating Financial-Grade Time Series Forecasting and Large Language Models

Isabella Martinez
School of Information Systems, University of Maryland Baltimore County
imartinez@umbc.edu

Wesley Sinclair
Department of Systems and Industrial Engineering, University of Arizona
wesley.s@arizona.edu

Abstract

The efficacy of modern digital advertising infrastructures is increasingly dependent on the ability to perform sub-millisecond intent discovery within highly volatile auction environments. While traditional programmatic architectures rely on historical behavioral heuristics, they often fail to capture the transient, context-dependent shifts in consumer intent that drive conversion. This paper proposes a novel system architecture that integrates financial-grade time series forecasting with large language models to facilitate real-time intent discovery. By treating user interaction streams as high-frequency financial assets, we apply temporal learning pipelines designed for market microstructure to the advertising bidding process. This integration allows for the synthesis of quantitative engagement dynamics with the deep semantic reasoning of transformer-based architectures. Our research focuses on the system-level trade-offs between inferential depth and execution latency, emphasizing the necessity of hardware-aware distributed orchestration. We explore the structural requirements for a high-throughput pipeline that can ingest billions of events daily while maintaining causal consistency across geographically disparate edge nodes. Furthermore, the paper addresses critical socio-technical dimensions, including the governance of autonomous bidding agents, the sustainability of massive-scale transformer inference, and the ethical imperatives of algorithmic fairness in digital commerce. By aligning the precision of financial engineering with the contextual awareness of natural language processing, this framework provides a robust blueprint for the next generation of digital advertising infrastructures, ensuring that intent discovery is both statistically accurate and semantically grounded.

Keywords

Digital Advertising Infrastructure, Intent Discovery, Financial Machine Learning, Large Language Models, High-Throughput Systems, Distributed Orchestration, Socio-Technical Governance.

1. Introduction

The digital advertising landscape has evolved into a hyper-competitive, high-frequency ecosystem that mirrors the structural complexity of global financial markets. In this environment, the fundamental unit of value is the "intent signal"—a transient indicator of a consumer's potential to engage or convert. Real-time bidding (RTB) systems are currently tasked with processing millions of bid requests per second, each requiring a sophisticated evaluation of user context, historical behavior, and inventory quality. However, traditional programmatic architectures often suffer from a "contextual gap," where the systems can identify that a user is active but struggle to understand the why behind the activity. This limitation leads to significant inefficiencies in capital allocation and a degradation of the user experience. To bridge this gap, a systemic shift is required—one that moves away from static heuristics toward a dynamic, synergistic integration of temporal learning and semantic reasoning.

Accelerating real-time intent discovery requires an infrastructure capable of synthesizing two disparate data modalities: the high-velocity, low-dimensional numerical streams of user engagement and the high-dimensional, semantically dense narratives of linguistic context. Financial-grade time series forecasting provides the necessary tools to model engagement as a stochastic process, identifying momentum and volatility in user interest. Simultaneously, large language models (LLMs) offer the reasoning capabilities required to parse the semantic nuances of search queries, content consumption, and social interactions. The challenge lies in the orchestration of these components within a high-throughput distributed system that can satisfy the rigid latency constraints of real-time auctions.

This paper provides a comprehensive analysis of the system-level requirements for such an integrated infrastructure. We move beyond the optimization of individual model weights to examine the broader organizational and technical frameworks needed to deploy these systems at scale. This includes a deep dive into hardware-aware inference, distributed temporal synchronization, and the socio-technical governance of autonomous advertising agents. As intent discovery becomes increasingly automated, the implications for privacy, fairness, and environmental sustainability become central to the system's long-term viability. By synthesizing the rigors of financial engineering with the depth of computational linguistics, we propose a scalable architecture for a more intelligent and resilient digital advertising ecosystem.

2. Conceptual Convergence of Advertising and Financial Systems

The structural parallels between programmatic advertising and electronic financial markets are profound and informative for system design. Both domains are characterized by the rapid exchange of assets—impressions in advertising and securities in finance—within automated auction environments governed by strict temporal constraints. In finance, market microstructure analysis focuses on price discovery and liquidity provision; in advertising, the equivalent process is "intent discovery," where the system attempts to price an impression based on the probability of a specific consumer outcome. Recognizing this convergence

allows system architects to apply the robust methodologies developed for high-frequency trading (HFT) to the digital advertising pipeline.

Financial-grade time series forecasting is particularly relevant for modeling user "interest momentum." User engagement is rarely a series of isolated events; it is a temporal sequence with varying degrees of autocorrelation and volatility. By treating a user's clickstream as a financial time series, we can utilize advanced forecasting techniques to identify regime shifts in consumer behavior. For instance, a sudden surge in search frequency for a specific category can be modeled as a "volatility breakout," signaling a transition from passive browsing to active intent. This quantitative foundation ensures that the bidding agent is not merely reacting to the last observed event but is anticipating the trajectory of user interest.

However, numerical dynamics alone cannot capture the full spectrum of intent. Linguistic context provides the causal grounding that numerical models often lack. A large language model can distinguish between a user searching for "apple" as a fruit and "apple" as a technology brand, a distinction that is vital for accurate pricing. The convergence of these fields requires a "cross-modal latent space" where temporal features and semantic embeddings are aligned. From a systems perspective, this necessitates a high-throughput fabric that can synchronize these signals without introducing catastrophic latency. The goal is to move from a frequentist model of "probability of click" to a reasoning model of "logic of intent," thereby increasing the robustness of the advertising infrastructure during periods of rapid market change.

3. High-Throughput System Architecture for Intent Synthesis

Designing a system capable of integrating LLMs with financial-grade forecasting for real-time intent discovery requires a radical departure from monolithic server architectures. We propose a tiered distributed infrastructure that partitions the cognitive load between "fast-path" temporal models and "deep-path" reasoning models. The fast-path consists of lightweight, high-velocity forecasting modules deployed at the network edge. These modules are responsible for the immediate processing of numerical streams, utilizing specialized hardware like Field-Programmable Gate Arrays (FPGAs) or Application-Specific Integrated Circuits (ASICs) to maintain sub-millisecond response times for the majority of auction requests.

The deep-path reasoning is handled by a distributed cluster of large language models. Given the computational intensity of transformer inference, these models are orchestrated through a "speculative reasoning" framework. In this approach, a smaller, faster "draft" model provides an initial semantic assessment of the user context, which is then verified or refined by a larger "target" model in the background. This tiered approach allows the system to provide a high-confidence intent signal even when the full reasoning pipeline has not completed. For the intent discovery system, this means that the bidding agent can act on a "pre-cognitive" signal from the fast-path while the deep-path provides the necessary context for long-term optimization.

The synchronization of these paths is managed by a hardware-aware orchestration layer that treats compute, memory, and network as a unified resource pool. High-throughput inference requires the minimization of data movement, necessitating a "compute-near-data" strategy where the model shards are co-located with the regional data ingestion points. Furthermore, the system must implement a "temporal consistency protocol" to ensure that the semantic insights from the LLM are correctly aligned with the corresponding numerical price action. This is achieved through a high-resolution global clock and a distributed log-structured merge-tree (LSM-tree) that allows for the efficient storage and retrieval of multi-modal events. By optimizing the architecture for the specific hardware constraints of each path, the system achieves a balance between analytical depth and operational velocity.

4. Structural Trade-offs in Real-Time Intent Discovery

The engineering of an intent discovery system is an exercise in managing fundamental structural trade-offs, primarily the tension between "inferential depth" and "execution throughput." Increasing the complexity of the semantic reasoning layer—for example, by using a larger transformer model or a longer context window—improves the precision of the intent discovery but increases the time required for a single inference step. In the world of real-time bidding, where an impression is lost if the bid is not received within fifty to one hundred milliseconds, excessive reasoning depth is a liability. Consequently, the system must implement a "dynamic reasoning depth" mechanism that adjusts the complexity of the inference based on the expected value of the auction.

Another critical trade-off exists between "centralized coherence" and "decentralized agility." A centralized system allows for a globally consistent model of user behavior but introduces significant network latency and represents a single point of failure. A fully decentralized system, with models running independently at every edge node, eliminates network latency but leads to "model drift" where different parts of the infrastructure develop divergent understandings of the same user. Our framework advocates for a "federated reasoning" approach, where the edge nodes perform local intent discovery while a central coordinator asynchronously aggregates these insights to update the global model. This allows the system to maintain high local agility while benefiting from the collective intelligence of the entire network.

Robustness and redundancy also present a significant trade-off against cost and energy efficiency. To ensure five-nines of availability, the system must replicate its temporal and semantic pipelines across multiple geographic zones. This redundancy doubles the computational footprint and energy consumption of the infrastructure. In an era where sustainability is a core corporate and societal value, this "redundancy-sustainability" trade-off must be managed through intelligent power scaling and hardware-aware load balancing. By routing traffic to data centers powered by renewable energy during peak production hours, the system can maintain high robustness while minimizing its environmental impact. These structural choices define the resilience and efficiency of the intent discovery pipeline in a volatile digital economy.

5. Deployment, Sustainability, and Infrastructure Resilience

The deployment of an integrated intent discovery system at a global scale requires a focus on infrastructure resilience and sustainability that goes beyond traditional software engineering. The massive computational footprint of continuous LLM inference and high-frequency time series forecasting leads to significant energy demands. To address this, our framework advocates for a "sustainable-by-design" architecture. This involves the use of "green scheduling" algorithms that prioritize compute nodes based on their real-time carbon intensity. Furthermore, we emphasize the deployment of "energy-proportional" hardware that can scale its power consumption based on the volume of bid requests, ensuring that the system remains efficient during low-traffic periods.

Infrastructure resilience is also a function of "adversarial robustness." Digital advertising systems are frequent targets for malicious actors who attempt to manipulate the intent discovery process through "bot-driven volatility" or "semantic poisoning." A cross-modal system is inherently more resilient to these attacks because it can verify the linguistic context against the numerical engagement patterns. If a sudden surge of "intent" is identified by the LLM but is not accompanied by any meaningful change in engagement momentum, the system can flag the signal as synthetic. This "cross-modal verification" provides a critical layer of defense against fraud, ensuring that advertising capital is directed toward genuine human intent.

Furthermore, the infrastructure must support "non-disruptive evolution." As new transformer architectures are developed and new market dynamics emerge, the system must allow for the hot-swapping of models without requiring a full system reboot. This is achieved through a microservices-based architecture where the temporal and semantic encoders are containerized and orchestrated through a decentralized scheduler. This modularity also facilitates "regional optimization," where different reasoning models can be deployed in different geographic markets to account for local language nuances and regulatory requirements. By building a flexible and modular infrastructure, the intent discovery system can adapt to the evolving complexities of the global advertising landscape while maintaining high operational uptime.

6. Algorithmic Governance and the Ethics of Intent Discovery

As intent discovery systems become more autonomous and semantically aware, the question of algorithmic governance becomes a central socio-technical concern. Traditional advertising oversight focuses on "transparency" and "choice," but these concepts are difficult to apply to a distributed swarm of reasoning agents. We propose a "governance-through-architecture" approach, where ethical constraints are embedded directly into the system’s objective function. This involves the implementation of "fairness-aware" bidding agents that are explicitly programmed to avoid discriminatory pricing or targeting based on protected characteristics, even if such patterns appear statistically profitable.

Interpretability is also a critical component of governance. Unlike frequentist models, which are often criticized as "black boxes," a reasoning-augmented intent discovery system can be designed to provide "causal justifications" for its decisions. By utilizing attention-map

visualization and natural language generation, the system can explain why a particular user was classified as having "high intent" for a specific product. This "Explainable AI" (XAI) capability is essential for building trust with consumers and for providing regulators with the information they need to audit the system for bias or manipulation. In the event of an unusual bidding event, these justifications allow for a rapid "post-mortem" analysis of the agent's logic.

Finally, we must address the "privacy-utility" trade-off. Deep intent discovery requires access to granular user data, which often conflicts with emerging privacy regulations like the GDPR and CCPA. Our framework utilizes "privacy-preserving computation," such as federated learning and differential privacy, to ensure that the system can discover intent without ever seeing raw, identifiable user information. By performing the majority of the reasoning at the edge and only transmitting encrypted, aggregated embeddings to the central coordinator, the system maintains a high degree of utility while respecting individual privacy. This alignment of technical capability with ethical responsibility is essential for the social legitimacy of the advertising infrastructure.

7. Global Policy Implications and the Regulatory Landscape

The rise of high-throughput, intent-aware advertising agents necessitates a fundamental shift in global regulatory policy. Existing frameworks are largely designed for "static" privacy and "offline" consumer protection; they are insufficient for a market driven by AI-generated reasoning. Policy-makers must now consider the "integrity of the informational environment" as a core component of economic stability. If an intent discovery system is flawed, it can lead to "informational bubbles" or the systemic exclusion of certain groups from economic opportunities. Regulators must therefore develop standards for the "stress-testing of reasoning," ensuring that systems are robust to a wide range of hypothetical market shocks.

Another critical policy dimension is the regulation of "synthetic engagement." As AI models become capable of generating highly persuasive content, there is a risk of a "recursive feedback loop" where AI-driven advertising influences the very AI-driven intent discovery systems that monitor the market. Regulators must establish clear "watermarking" requirements for AI-generated advertising and implement strict penalties for systems that utilize synthetic engagement to inflate inventory value. Furthermore, international cooperation is essential to manage "cross-border informational arbitrage," where a system exploited in one jurisdiction could cause instability in another.

The regulatory framework should also incentivize the development of "pro-social advertising." Instead of merely optimizing for short-term conversion, the intent discovery system could be required to incorporate "long-term consumer welfare" and "market diversity" as core components of its objective function. A cross-modal system is uniquely positioned to handle such multi-objective tasks, as it can parse complex regulatory guidelines and incorporate them into its reasoning process. By aligning the system's "narrative logic" with public policy goals, we can transform autonomous advertising agents from potential sources of manipulation into tools for sustainable and equitable economic growth.

8. Socio-Technical Perspectives on the Future of Digital Commerce

The transition toward semantically aware intent discovery is not merely a technical event but a socio-technical evolution that redefines the human-machine partnership in digital commerce. We are moving away from a world where humans define the targeting rules and machines execute the bids, toward a world of "delegated reasoning." In this future, the role of the marketing professional is to act as an "intent curator," guiding the machine's reasoning process and ensuring it remains aligned with brand values and societal expectations. This requires a new set of interdisciplinary skills, bridging the gap between computational linguistics, systems engineering, and economic theory.

This evolution also impacts the "structure of consumer trust." Historically, trust in advertising was built on the reputation of brands and the transparency of their claims. In the era of autonomous intent discovery, trust will be built on the "verifiability of logic." If an autonomous system can demonstrate consistent, explainable, and ethical reasoning, it will gain the trust of consumers who are increasingly wary of automated manipulation. This shift from "reputation-based trust" to "logic-based trust" has the potential to make digital commerce more transparent and less prone to the "irrational exuberance" that often accompanies speculative bubbles.

Finally, we must consider the "long-term cognitive impact" of delegated intent discovery. If we rely on machines to interpret our desires and serve our needs, we must ensure that we do not lose our own capacity for spontaneous discovery and critical evaluation. The design of the advertising infrastructure must therefore include "human-centric feedback loops" where the system's interpretations are regularly challenged and refined by diverse groups of human experts. By treating the digital advertising system as a socio-technical organism rather than a purely algorithmic one, we can ensure that the "bridge" between numbers and narratives serves to strengthen the fabric of our global society.

9. Conclusion

This paper has proposed a unified system-level framework for accelerating real-time intent discovery in digital advertising by integrating financial-grade time series forecasting and large language models. Through a tiered distributed architecture, we have demonstrated how advertising infrastructures can synchronize high-velocity engagement streams with the deep semantic context of consumer narratives. Our analysis of structural trade-offs, deployment resilience, and algorithmic governance provides a comprehensive roadmap for building the next generation of digital advertising intelligence.

The journey toward context-aware, autonomous advertising is fraught with technical and ethical challenges, but it also offers a unique opportunity to create a more efficient and equitable global economy. By grounding the "logic of numbers" in the "logic of stories," we can build systems that are not only faster and more accurate but also more robust to the non-stationarities of human behavior. As we continue to develop these cross-modal bridges, the focus must remain on the socio-technical dimensions of our work—ensuring that our

infrastructures are sustainable, fair, and transparent. The future of digital advertising is not just about the clicks we predict, but the intent we understand and the values we uphold in the process.

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. Acemoglu, D., & Restrepo, P. (2019). Automation and new tasks: How technology displaces and creates labor. Journal of Economic Perspectives, 33(2), 3–30.

3. Baltussen, G., van Vliet, P., & van Vliet, S. (2021). The cross-section of stock returns before 1926 (and beyond). Journal of Financial Economics, 141(3), 1146–1163.

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

5. 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.

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

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

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., 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.

16. Kirilenko, A. S., Kyle, A. S., Samadi, M., & Tuzun, T. (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. Liu, T. (2026). Leakage-Safe Benchmark Design for Market-Stress Early Warning: An Economically Credible Evaluation.

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

20. 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.

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., 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, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information

Processing Systems, 30.

28. Varian, H. R. (2007). Position auctions. International Journal of Industrial Organization, 25(6), 1163–1178.

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

30. 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.

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

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

33. 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.

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