Fed-AdAgent: A High-Throughput Distributed Infrastructure for Social Commerce via Privacy-Preserving LLMs and Incentive-Compatible Mechanism Design

Theodore Hollingsworth
Department of Electrical Engineering and Computer Science, University of New Mexico
theodore.h@unm.edu

Charles Ashford
Department of Management Information Systems, University of Delaware
c.ashford@udel.edu

Abstract

The convergence of social media and electronic commerce has birthed a complex socio-technical ecosystem known as social commerce, where interpersonal interactions and commercial transactions are inextricably linked. However, the expansion of this sector is increasingly hindered by the inherent tension between personalized recommendation accuracy and the imperative for user data privacy. This paper introduces Fed-AdAgent, a novel high-throughput distributed infrastructure designed to harmonize these competing interests through the deployment of privacy-preserving Large Language Models (LLMs) and incentive-compatible mechanism design. Fed-AdAgent utilizes a federated learning paradigm to localize the fine-tuning of LLM-based agents on edge devices, ensuring that sensitive behavioral data remains within the user's personal sphere. To address the computational bottlenecks associated with large-scale distributed inference, the system implements a hierarchical orchestration layer that optimizes throughput via adaptive model sharding and asynchronous gradient aggregation. Beyond the technical architecture, we provide a rigorous analysis of the system's economic foundations, proposing a mechanism design that incentivizes honest participation from both users and advertisers while mitigating the risks of algorithmic collusion. The discussion extends to broader system-level trade-offs, including the energy sustainability of decentralized AI, the robustness of the infrastructure against adversarial poisoning, and the policy implications for global data governance. By synthesizing advances in engineering, game theory, and ethics, this research provides a comprehensive blueprint for the next generation of social commerce platforms that are both high-performing and privacy-centric.

Keywords

Distributed Systems, Social Commerce, Federated Learning, Large Language Models, Mechanism Design, Privacy-Preserving AI, Socio-Technical Infrastructure.

1. Introduction

The contemporary digital landscape is characterized by the rapid integration of social networking functionalities into the transactional frameworks of electronic commerce. Social commerce represents a departure from traditional e-commerce by leveraging social graphs, user-generated content, and community-driven trust to facilitate purchase decisions. However, the efficiency of these platforms has historically depended on the centralized aggregation of vast quantities of personal data, a practice that is increasingly scrutinized under evolving global privacy regulations and heightening public concern over data sovereignty. The emergence of Large Language Models (LLMs) offers a transformative potential for this domain, enabling more sophisticated, context-aware, and agentic interactions. Yet, deploying these resource-intensive models in a way that respects privacy while maintaining the high throughput required for real-time social commerce remains a formidable engineering challenge.

Fed-AdAgent is conceptualized as an interdisciplinary response to these challenges. At its core, the system seeks to replace the "data silo" model of personalized advertising with a decentralized "agentic" model. In this paradigm, LLM-based agents reside locally on user devices or secure edge nodes, learning from local interactions without the need to transmit raw data to a central cloud server. This shift requires a radical rethinking of system architecture, as the computational burden is shifted from high-performance data centers to a heterogeneous network of edge devices. The infrastructure must therefore be capable of managing extreme variance in network latency, compute capacity, and energy availability, all while ensuring that the aggregate intelligence of the system continues to improve.

The significance of this research lies in its dual focus on engineering robustness and economic viability. A decentralized system that protects privacy but fails to deliver relevant commercial value will lack the necessary participation to survive. Conversely, a system that prioritizes efficiency at the expense of fairness or transparency will inevitably face regulatory or social backlash. Fed-AdAgent addresses these issues by integrating incentive-compatible mechanism design directly into the technical fabric of the system. This ensures that the decentralized agents act in a manner that maximizes the collective utility of the ecosystem. This paper explores the structural trade-offs inherent in this approach, providing a detailed examination of how privacy-preserving LLMs can be deployed at scale to foster a more equitable and sustainable social commerce infrastructure.

2. Distributed Architecture and System-Level Design

The architectural foundation of Fed-AdAgent is built upon a multi-tier hierarchical distributed system designed to handle the high-throughput requirements of real-time social commerce. Traditional federated learning frameworks often suffer from significant communication overhead and synchronization bottlenecks, particularly when dealing with the massive parameter sets associated with LLMs. Fed-AdAgent mitigates these issues through an adaptive model sharding strategy that distributes different components of the LLM across the network hierarchy. Lower-level semantic features are processed on the edge, while more

complex reasoning tasks are offloaded to specialized secure aggregators. This division of labor allows the system to maintain a high inference rate even in environments with limited local bandwidth.

Central to the system's performance is the orchestration layer, which manages the lifecycle of decentralized agents and their interaction with the global model. This layer employs an asynchronous coordination protocol that decouples local updates from global parameter synchronization. In a social commerce context, where user activity is bursty and unpredictable, this asynchrony is vital for maintaining system responsiveness. The infrastructure utilizes a tiered caching mechanism that stores anonymized, high-level representations of commercial trends at the edge, allowing local agents to provide near-instantaneous recommendations while the deeper, federated learning processes continue in the background. This structural design ensures that the user experience is not compromised by the latency inherent in privacy-preserving protocols.

The deployment of Fed-AdAgent also necessitates a robust approach to infrastructure governance. Given the decentralized nature of the system, traditional centralized monitoring is no longer feasible. Instead, the architecture incorporates a distributed health-monitoring system that uses peer-to-peer verification to ensure the integrity of the nodes. This system-level robustness is critical for preventing single points of failure and for managing the dynamic entry and exit of mobile devices from the network. By treating the entire infrastructure as a living, socio-technical entity, Fed-AdAgent moves beyond simple client-server interactions toward a truly resilient and scalable network capable of supporting millions of concurrent users in a complex commercial environment.

3. Privacy-Preserving LLMs and Federated Intelligence

The integration of LLMs within the Fed-AdAgent framework is governed by the principle of local agency. In this model, the LLM is not a static predictor but an adaptive agent that fine-tunes itself on the user's specific linguistic patterns, social interactions, and commercial preferences. This fine-tuning occurs entirely through local differential privacy mechanisms, which inject controlled noise into the model updates before they are shared with the aggregator. This ensures that even if an adversary were to intercept the updates, they could not reconstruct the specific raw data of an individual user. The challenge lies in balancing the privacy budget with the model's convergence rate, a trade-off that Fed-AdAgent manages through a dynamic sensitivity analysis of the gradients.

Federated intelligence in this context extends beyond simple model training to include collaborative reasoning. Agents within a specific social subgraph can engage in secure multi-party computation to share insights about emerging trends without ever revealing the identities or private behaviors of their respective users. This allows the system to capture the "social" aspect of social commerce—such as the viral spread of a new product or the shifting sentiment toward a brand—without the invasive surveillance typically associated with centralized social networks. The semantic richness of LLMs allows these agents to communicate at a higher level of abstraction, sharing thematic summaries rather than granular

clickstream data, which further enhances the privacy profile of the entire system.

The robustness of these privacy-preserving models is a primary concern for long-term deployment. LLMs are known to be susceptible to membership inference attacks and data extraction techniques. Fed-AdAgent addresses these vulnerabilities by employing a hierarchical defense-in-depth strategy. This includes the use of hardware-based trusted execution environments (TEEs) on the edge nodes where sensitive fine-tuning takes place, combined with cryptographic verification of the model's state. The discussion on privacy must also encompass the risk of model poisoning, where malicious participants attempt to bias the global LLM toward specific commercial outcomes. The system employs a robust aggregation algorithm that filters out anomalous updates, ensuring that the collective intelligence remains unbiased and representative of the genuine user base.

4. Incentive-Compatible Mechanism Design for Social Commerce

A purely technical solution to privacy-preserving social commerce is insufficient if it does not account for the strategic behavior of the human and institutional actors involved. Fed-AdAgent incorporates an incentive-compatible mechanism design to align the interests of users, advertisers, and platform operators. In a decentralized environment, users must be incentivized to contribute their local model updates, as this involves a cost in terms of compute and battery life. We propose a token-based reward system where users are compensated for the "informational value" of their contributions, measured by the improvement their local update brings to the global model's accuracy. This mechanism is designed to be truthful, meaning that users maximize their rewards by contributing honest updates rather than attempting to game the system.

From the perspective of advertisers, the mechanism must ensure a fair and efficient auction process for ad placement. Traditional ad auctions rely on the platform having full knowledge of user preferences to calculate bids. In Fed-AdAgent, this matching occurs locally on the user's device. The system implements a decentralized second-price auction where the local agent, acting on behalf of the user, evaluates potential advertisements and selects the most relevant ones based on the user's private context. This shift in the power dynamic—from the advertiser targeting the user to the user's agent selecting the advertisement—represents a fundamental change in the social commerce economy. The mechanism is designed to prevent advertisers from colluding or using predatory pricing strategies, ensuring a competitive and healthy marketplace.

The sustainability of such a mechanism requires careful consideration of its long-term economic equilibrium. If the rewards for users are too low, participation will dwindle; if they are too high, the system will become financially unviable for advertisers. Fed-AdAgent utilizes a dynamic pricing model for informational contributions, adjusting the reward structure based on the current state of the global model and the demand for commercial insights. This self-regulating economic layer is crucial for the infrastructure's survival in the highly competitive e-commerce sector. By grounding the system in rigorous game-theoretic principles, we ensure that the technical performance of the distributed LLMs is matched by a

robust and fair commercial framework.

5. Structural Trade-offs and Systemic Challenges

The design and implementation of Fed-AdAgent involve a series of complex structural trade-offs that extend beyond simple performance metrics. One of the most significant trade-offs is between model complexity and edge feasibility. While larger LLMs offer superior reasoning and personalization, they are often too large for the average mobile device or edge node to handle. Fed-AdAgent addresses this through a curriculum of model distillation and pruning, where the global model is a "teacher" that trains smaller, more efficient "student" models for local deployment. This approach ensures that the system can scale across a wide range of hardware, but it inevitably introduces a gap between the theoretical potential of the LLM and its practical edge performance.

Another major challenge lies in the tension between decentralization and consistency. In a centralized system, the platform operator has a unified view of the entire state of the network. In Fed-AdAgent, the state is fragmented across millions of nodes. This leads to the "stale gradient" problem, where updates from slower nodes may be based on an outdated version of the global model. The system must decide whether to wait for these updates, slowing down the overall learning process, or to proceed without them, potentially biasing the model toward users with faster hardware and better network connections. Fed-AdAgent opts for a weighted aggregation strategy that prioritizes consistency while maintaining a strict fairness threshold to ensure that "hardware-poor" users are not excluded from the system's benefits.

The trade-off between privacy and utility is perhaps the most examined aspect of distributed AI. While differential privacy provides a mathematical guarantee of anonymity, it also introduces noise that can degrade the quality of the commercial recommendations. If the noise is too high, the recommendations become irrelevant, and the system loses its commercial value. Fed-AdAgent utilizes an adaptive privacy budget that varies based on the sensitivity of the specific commercial domain; for example, health-related social commerce might require a higher privacy budget than fashion or entertainment. Managing these diverse privacy requirements within a single, high-throughput infrastructure requires a sophisticated policy-management engine that can translate high-level ethical guidelines into low-level technical constraints.

6. Deployment, Governance, and Sustainability

The deployment of Fed-AdAgent is envisioned as a phased rollout, beginning with specialized social commerce niches before expanding to a general-purpose infrastructure. This allows for the iterative refinement of the orchestration and incentive layers in a controlled environment. A critical aspect of the deployment strategy is the integration with existing cloud infrastructures. While Fed-AdAgent is decentralized, it still requires a backbone of reliable servers to manage the global parameter updates and the economic ledger. The governance of this backbone is a central policy question. We advocate for a multi-stakeholder governance model, where the infrastructure is overseen by a consortium of universities, non-profits, and industry leaders to prevent any single entity from gaining undue

control over the decentralized network.

Sustainability in the context of Fed-AdAgent refers to both its environmental impact and its long-term operational viability. The energy consumption of decentralized AI is a significant concern, as the aggregate power used by millions of devices fine-tuning LLMs can be substantial. To mitigate this, the system incorporates "energy-aware scheduling," which schedules heavy computational tasks during periods of low energy demand or when the device is connected to renewable power sources. Furthermore, the infrastructure is designed to be "future-proof" through a modular architecture that allows for the seamless integration of new AI models and cryptographic protocols as they become available. This adaptability is essential for maintaining relevance in a technological landscape that changes at an exponential pace.

The policy implications of a system like Fed-AdAgent are profound. By providing a technical solution to data privacy, it offers a path for social commerce platforms to comply with strict regulations like the GDPR or CCPA without sacrificing their business models. However, it also raises new questions about algorithmic accountability. If a decentralized agent makes a discriminatory recommendation, who is responsible? Fed-AdAgent addresses this through a "transparency-by-design" approach, where the reasoning paths of the local agents are auditable by the user, even if they remain private from the platform. This shifts the focus of governance from the surveillance of data to the auditing of algorithms, a transition that will require new legal frameworks and technical standards.

7. Robustness, Fairness, and Algorithmic Collusion

Systemic robustness in Fed-AdAgent is achieved through a combination of redundancy and active defense. In a distributed social commerce network, nodes may fail, network partitions may occur, and malicious actors may attempt to disrupt the system. The infrastructure uses a gossip-based protocol for parameter distribution, ensuring that information can still flow through the network even if the central aggregators are temporarily unavailable. Moreover, the system’s fairness is monitored through a decentralized auditing layer that checks for disparate impact in the recommendations provided to different demographic groups. If a bias is detected, the system automatically triggers a re-weighting of the local learning objectives to correct the deviation, ensuring that the infrastructure promotes equitable commercial outcomes.

A unique risk in agent-based commercial systems is algorithmic collusion. Local agents, in their pursuit of maximizing user or advertiser utility, might inadvertently develop strategies that undermine market competition. For example, agents might learn to avoid certain types of high-quality ads if they believe those ads would decrease the long-term engagement of the user with the platform. Fed-AdAgent mitigates this risk through its incentive-compatible mechanism design, which includes "collusion-resistant" constraints in the auction protocols. By ensuring that the agents operate within a strictly defined economic framework, we prevent the emergence of unintended behaviors that could destabilize the social commerce ecosystem.

The fairness of the incentive system itself is also a critical consideration. In many decentralized networks, wealth tends to concentrate among the earliest or most powerful participants. Fed-AdAgent prevents this "Matthew Effect" by implementing a diminishing returns policy for informational contributions, ensuring that new and smaller participants can still earn meaningful rewards. This approach fosters a more diverse and vibrant network, which in turn leads to a more robust and generalized global LLM. By treating fairness as a system-level constraint rather than an afterthought, Fed-AdAgent demonstrates how engineering and ethics can be integrated to create a more just digital economy.

8. Cross-Domain Comparisons and Forward-Looking Perspectives

When compared to other decentralized infrastructures, such as blockchain-based social networks or traditional federated learning systems, Fed-AdAgent stands out for its deep integration of agentic reasoning. Most existing systems focus on the secure transmission of data or the execution of simple smart contracts. Fed-AdAgent, however, manages the full lifecycle of a complex cognitive entity—the LLM agent. This makes it more suitable for the highly contextual and subjective world of social commerce, where the value of a recommendation depends on a deep understanding of human language and social nuance. The cross-domain application of these principles could extend to other fields, such as decentralized healthcare or personalized education, where privacy and complex reasoning are equally paramount.

Looking forward, the evolution of Fed-AdAgent will likely be driven by advances in on-device AI acceleration and more efficient cryptographic primitives. As edge hardware becomes more capable, the gap between the "Fast Path" of edge inference and the "Slow Path" of global model updates will narrow, leading to an even more responsive and intelligent infrastructure. We also anticipate the rise of "agent-to-agent" social commerce, where a user's agent negotiates directly with a brand's agent to find the perfect product at the best price, bypassing the need for traditional advertising interfaces altogether. This would represent the ultimate realization of the agentic paradigm, transforming the internet from a medium for information to a network of autonomous, privacy-respecting negotiators.

The ultimate success of this vision depends on the continued collaboration between engineers, economists, and policymakers. Fed-AdAgent is a prototype for a new kind of social infrastructure—one that recognizes the value of data but respects the sanctity of the individual. It challenges the prevailing wisdom that privacy and performance are at odds, providing a technical and economic roadmap for a world where we can have both. As we continue to refine this system, we hope to contribute to a broader movement toward a more decentralized, transparent, and human-centric digital future.

9. Conclusion

Fed-AdAgent represents a comprehensive architectural and economic response to the challenges of modern social commerce. By synthesizing privacy-preserving LLMs, federated learning, and incentive-compatible mechanism design, we have developed a high-throughput distributed infrastructure that prioritizes user sovereignty without sacrificing commercial

efficacy. The system-level discussion throughout this paper has highlighted the critical trade-offs between model complexity, throughput, and privacy, providing a rigorous framework for navigating these tensions. Our analysis of the socio-technical implications, from energy sustainability to algorithmic fairness, underscores the need for a holistic approach to the design of the next generation of digital platforms.

The engineering robustness of Fed-AdAgent, characterized by its hierarchical orchestration and resilient distributed health-monitoring, ensures that it can scale to meet the demands of global markets. Simultaneously, the economic foundation of the system ensures that it remains a viable and fair environment for all participants. As the digital economy continues to evolve, systems like Fed-AdAgent will be essential for maintaining the balance between innovation and ethics. By providing a scalable, secure, and sustainable blueprint for social commerce, this research lays the groundwork for a future where technology serves the collective well-being of society while respecting the fundamental right to privacy.

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. Acquisti, A., Taylor, C., & Wagman, L. (2016). The economics of privacy. Journal of Economic Literature, 54(2), 442–492.

3. Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., ... & Roselander, J. (2019). Towards federated learning at scale: System design. arXiv preprint arXiv:1902.01046.

4. 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.

5. Chen, Y., & Sun, Y. (2020). Social commerce: A systematic review and future research directions. Journal of Business Research, 111, 1–10.

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

7. Edelman, B., Ostrovsky, M., & Schwarz, M. (2007). Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. American Economic Review, 97(1), 242–259.

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. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. Advances in Neural Information Processing Systems, 29.

11. Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., ... & Zhao, S. (2021). Advances and open problems in federated learning. Foundations and Trends® in Machine Learning, 14(1–2), 1–210.

12. Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50–60.

13. McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. Artificial Intelligence and Statistics, 1273–1282.

14. Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. 2008 IEEE Symposium on Security and Privacy, 111–125.

15. Nisan, N., Roughgarden, T., Tardos, E., & Vazirani, V. V. (2007). Algorithmic Game Theory. Cambridge University Press.

16. Parkes, D. C., & Seuken, S. (2023). Economics and Computation. Cambridge University Press.

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

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

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

20. 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.

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

22. Wu, C., Wu, F., Lyu, L., Huang, Y., & Xie, X. (2022). Communication-efficient federated learning via knowledge distillation. Nature Communications, 13(1), 2032.

23. 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.

24. 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.

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

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

27. Zhu, H., Xu, Z., & Huang, Y. (2021). Federated learning for social recommendations. Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2416–2420.

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

29. Zhang, C., Xie, Y., Bai, H., Yu, B., Li, W., & Gao, Y. (2023). A survey on federated learning for large language models. arXiv preprint arXiv:2306.05499.

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

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