Fed-AdScale: A Distributed Edge-Cloud Infrastructure for Social Commerce via Privacy-Preserving LLM Inference and Trusted Execution

Miles Crawford
School of Engineering and Applied Sciences, Gonzaga University
m.crawford@gonzaga.edu

Marcus Whitmore
Department of Electrical Engineering and Computer Science, University of New Mexico
mwhitmore@unm.edu

Blake Callahan
Department of Management Information Systems, University of Delaware
blake.c@udel.edu

Abstract

The digital economy is currently witnessing a convergence of social media dynamics and electronic commerce, a phenomenon termed social commerce that relies heavily on personalized, context-aware interactions. However, the centralization of user data required for Large Language Model (LLM) inference creates significant privacy risks and regulatory hurdles. This paper proposes Fed-AdScale, a high-throughput distributed edge-cloud infrastructure designed to facilitate privacy-preserving LLM inference within social commerce ecosystems. Fed-AdScale leverages a multi-tier architecture that distributes computational workloads between resource-constrained edge devices and robust cloud environments using Trusted Execution Environments (TEEs) to ensure data confidentiality and model integrity. By implementing a novel hierarchical orchestration layer, the system manages the inherent trade-offs between inference latency, model accuracy, and communication overhead. We provide an exhaustive analysis of the system's structural design, emphasizing its robustness against adversarial attacks and its alignment with emerging global data governance policies. The discussion extends to the socio-technical implications of decentralized AI, including environmental sustainability through localized compute optimization and the promotion of algorithmic fairness via non-siloed data processing. By synthesizing advancements in hardware-assisted security and federated systems engineering, Fed-AdScale provides a scalable blueprint for the next generation of social commerce platforms that prioritize both user agency and commercial performance.

Keywords

Distributed Systems, Social Commerce, Federated Learning, Trusted Execution Environments, Edge Computing, Large Language Models, Privacy-Preserving AI.

1. Introduction

The rapid evolution of the digital marketplace has transitioned from static transactional websites to dynamic, socially-integrated ecosystems where purchase decisions are influenced by real-time interpersonal interactions and hyper-personalized content. Social commerce represents this intersection, demanding a level of predictive sophistication that traditional recommendation engines can no longer provide. The emergence of Large Language Models (LLMs) has offered a solution to this complexity, enabling platforms to understand nuanced user intent and generate persuasive, contextually relevant social advertisements. However, the deployment of LLMs at the scale required by global social commerce introduces a fundamental conflict between the hunger for personalized data and the imperative of user privacy. Centralized processing of highly sensitive social interactions not only risks catastrophic data breaches but also invites rigorous scrutiny from regulatory bodies enforcing frameworks like the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA).

Fed-AdScale is conceptualized as a socio-technical intervention designed to resolve this tension through structural innovation. Instead of mandating the transfer of raw user data to a central cloud for processing, Fed-AdScale distributes the inference and learning phases across a continuum of edge devices and secure cloud nodes. This approach is predicated on the belief that privacy is not merely a legal constraint but a technical requirement for building long-term trust in digital commerce. By utilizing Trusted Execution Environments (TEEs) at both the edge and the cloud gateway, Fed-AdScale ensures that even when data must leave a local device for complex reasoning tasks, it remains encrypted and inaccessible to the platform provider or third-party malicious actors. This architectural shift represents a departure from the "data lake" paradigm toward a "data stream" philosophy where intelligence is extracted locally and only non-sensitive insights are aggregated.

The significance of this research lies in its holistic approach to system-level discussion. We move beyond the narrow optimization of model weights to address the broader challenges of deployment, such as the energy cost of running LLMs on mobile hardware, the resilience of distributed networks against coordinated poisoning attacks, and the policy implications of decentralized governance. Fed-AdScale is designed to be high-throughput, catering to the millions of concurrent interactions typical of social media bursts, while maintaining a footprint that is sustainable for modern mobile infrastructure. Through this paper, we explore how Fed-AdScale acts as a foundational layer for a more equitable social commerce landscape, where user data is protected by hardware-level guarantees and commercial value is generated through collaborative, rather than extractive, intelligence.

2. Distributed Edge-Cloud Architecture and Hierarchical Design

The architectural core of Fed-AdScale is a three-tier hierarchy consisting of the local edge layer, the regional aggregator layer, and the global cloud core. At the local edge layer, user

devices perform initial, low-latency inference using distilled or quantized versions of LLMs. This ensures that the most immediate social interactions, such as text analysis for instant messaging within a shopping context, are handled without any external data transmission. However, as the complexity of the task increases—for instance, when generating a multi-modal social advertisement that aligns with a user's long-term behavioral profile—the edge device coordinates with the regional aggregator. This second tier consists of edge servers equipped with high-performance TEEs, such as Intel SGX or ARM TrustZone, which provide a "secure enclave" for more intensive computational tasks that exceed the power of a mobile processor but still require geographical proximity to minimize latency.

The integration between these tiers is managed by a dynamic orchestration engine that evaluates the structural trade-offs of every inference request. This engine considers variables such as the current network bandwidth, the battery status of the edge device, and the sensitivity of the data involved. If the data is deemed highly personal, the system mandates local processing or encryption within a local TEE. If the reasoning requires the vast knowledge base of a full-scale LLM, the orchestrator facilitates a secure "handshake" with the cloud core. In the cloud, Fed-AdScale maintains a large-scale model that is updated via federated learning protocols. Crucially, the cloud core never sees the raw inputs; it only receives encrypted representations or gradient updates that have been processed through a differential privacy layer. This hierarchical design allows Fed-AdScale to achieve a throughput that rivals centralized systems while maintaining a strictly decentralized trust model.

A key system-level innovation within this architecture is the use of "model sharding" across the edge-cloud continuum. Unlike traditional deployments where a model is either entirely on the cloud or entirely on the edge, Fed-AdScale splits the LLM into semantic components. The early layers of the transformer architecture, which handle general linguistic patterns, may reside on the edge, while the deeper layers, which specialize in commercial reasoning and cross-domain synthesis, are called upon from the secure cloud nodes. This division reduces the memory pressure on edge devices and minimizes the data overhead of each request. By treating the LLM as a distributed resource rather than a monolithic file, Fed-AdScale provides a flexible infrastructure that can adapt to the heterogeneous capabilities of the modern mobile ecosystem, from low-end smartphones to high-performance workstations.

3. Hardware-Assisted Security and Trusted Execution Environments

The reliance on Trusted Execution Environments (TEEs) is the cornerstone of Fed-AdScale's privacy-preserving capabilities. In the context of social commerce, where advertising agents must process private chat logs, browsing histories, and location data, traditional software-based encryption is often insufficient because the data must be decrypted in memory during the inference process. TEEs solve this "data-in-use" problem by creating a hardware-isolated region of the processor where data can be decrypted and processed while remaining invisible to the operating system, hypervisor, or even a root-level attacker. Fed-AdScale utilizes these enclaves to host the most sensitive parts of the LLM inference pipeline, ensuring that the "thoughts" of the advertising agent are never exposed to the

underlying infrastructure.

Deploying LLMs within TEEs, however, presents significant engineering challenges related to memory constraints and side-channel vulnerabilities. Standard TEEs often have limited Enclave Page Cache (EPC) sizes, which are far too small for the multi-billion parameter sets of modern LLMs. Fed-AdScale addresses this through a paged-inference strategy, where only the active layers of the model are swapped into the secure memory at any given time. This requires a highly optimized memory management unit that can predict the next required shard of the model to minimize the latency penalty of swapping. Furthermore, the system implements rigorous defenses against cache-timing attacks by normalizing the execution time of different inference paths, ensuring that an observer cannot infer the content of a user's query by measuring how long the processor takes to respond.

The use of TEEs also facilitates a robust verification mechanism known as remote attestation. Before an edge device offloads a computation to a regional or cloud node, it receives a cryptographic proof that the node is running the exact version of the Fed-AdScale software that has been audited and approved. This establishes a chain of trust that is rooted in hardware rather than the promises of a service provider. In a social commerce environment where multiple third-party vendors might provide specialized advertising agents, this attestation process ensures that no unauthorized or malicious model can gain access to the user's secure data stream. This creates a "zero-trust" infrastructure where security is a mechanical property of the system rather than a policy-based hope.

4. Infrastructure Robustness and Adversarial Defense

A distributed infrastructure for social commerce must be resilient not only to hardware failures but also to sophisticated adversarial interventions. Fed-AdScale faces unique threats such as model poisoning, where a malicious actor provides fraudulent gradient updates during the federated learning process to bias the LLM's advertising output toward certain products or political ideologies. To counter this, the system incorporates a robust aggregation layer that uses statistical outlier detection and Byzantine-resilient algorithms to identify and prune malicious updates before they can affect the global model. By maintaining a history of node reliability, the system can dynamically de-prioritize or quarantine nodes that exhibit suspicious behavior, ensuring that the collective intelligence remains uncorrupted.

Furthermore, the system is designed to withstand "inference attacks" where an adversary queries the model repeatedly to reconstruct the private data used for training. Fed-AdScale implements a strict privacy-budgeting mechanism based on Renyi Differential Privacy. Every time a node contributes to the global model or responds to an inference request, a certain amount of "privacy noise" is added, and the cumulative privacy loss is tracked. Once a user's privacy budget for a given period is exhausted, the system throttles further high-resolution inference for that user, reverting to a generalized, non-personalized mode. This ensures that the long-term privacy of the user is never compromised by the short-term demands of commercial personalization.

Robustness also extends to the physical and network reliability of the system. Social commerce is inherently global, meaning Fed-AdScale must operate across regions with varying levels of internet stability. The infrastructure utilizes a peer-to-peer (P2P) fallback mechanism where, in the event of a cloud core outage, regional aggregators can form a mesh network to continue providing localized inference services. This "graceful degradation" ensures that the commerce platform remains functional even under duress. By prioritizing a decentralized and self-healing network topology, Fed-AdScale provides a level of availability that is essential for the continuous, 24/7 nature of modern global trade.

5. Sustainability, Energy Efficiency, and Computational Trade-offs

The environmental impact of Large Language Models is an increasingly critical concern in systems research. Centralized AI training and inference consume vast amounts of electricity and require massive cooling infrastructures. Fed-AdScale addresses the sustainability challenge by optimizing the computational "locality" of each task. By performing the majority of initial processing on the edge, the system reduces the carbon footprint associated with long-distance data transmission and centralized server cooling. Mobile processors, while less powerful than data center GPUs, are often more energy-efficient for small-scale, quantized inference tasks. Fed-AdScale leverages this by using hardware-aware quantization, which reduces the precision of model weights to the minimum level necessary for commercial accuracy, thereby saving significant energy on every inference cycle.

Structural trade-offs are also visible in the system’s approach to "lazy" versus "eager" computation. In a social commerce context, many user interactions are exploratory and do not lead to immediate transactions. Fed-AdScale uses a predictive scheduling algorithm that only triggers full-scale LLM reasoning when a high-intent commercial signal is detected. For routine social browsing, the system uses "shallow" models that require minimal power. This adaptive complexity ensures that the high-throughput capabilities of the infrastructure are reserved for high-value interactions, drastically reducing the aggregate energy consumption of the network. This focus on "frugal AI" aligns with the emerging institutional requirements for corporate social responsibility and environmental stewardship in the tech sector.

Moreover, the sustainability of Fed-AdScale is tied to the longevity of the hardware it supports. By providing a scalable infrastructure that can run distilled models on older devices, the system prevents "forced obsolescence," where users are required to upgrade their hardware simply to access the latest AI features. This democratizes access to high-quality social commerce tools and reduces the electronic waste generated by the rapid turnover of consumer electronics. The forward-looking perspective of Fed-AdScale is one where intelligence is a shared, efficient, and accessible resource, rather than a luxury reserved for those with the most expensive hardware.

6. Fairness, Bias, and Algorithmic Governance

A socio-technical system for social commerce must be scrutinized for its impact on social equity and algorithmic fairness. LLMs are prone to inheriting the biases present in their training data, which can lead to discriminatory advertising practices where certain

demographic groups are systematically excluded from high-value opportunities or targeted with predatory commercial content. Fed-AdScale addresses this through a decentralized auditing framework. Because the inference happens locally or in secure enclaves, the system can run "fairness checks" in real-time. These checks compare the distribution of ads across different user profiles at the edge without needing to collect the raw data centrally. If a significant deviation is detected, the system can adjust the local model's weights to ensure a more equitable distribution of commercial content.

The governance of Fed-AdScale is predicated on the idea of "participatory infrastructure." Rather than a single corporation dictating the rules of the social commerce ecosystem, the decentralized nature of the system allows for a more distributed form of policy enforcement. Global data governance policies can be encoded as "smart contracts" within the regional aggregators, ensuring that the system automatically complies with local laws regarding consumer protection and data usage. This allows for a modular approach to governance, where the same underlying infrastructure can adapt to the diverse legal landscapes of different countries, from the privacy-centric regulations of Europe to the more market-oriented policies of North America.

Furthermore, the system promotes fairness by breaking the "data monopolies" held by large tech incumbents. In the Fed-AdScale model, commercial intelligence is generated through a collaborative federated process where small businesses and individual sellers can benefit from the aggregate insights of the network without needing to own the data themselves. This levels the playing field in the social commerce sector, fostering a more competitive and diverse marketplace. By embedding fairness and transparency into the hardware-level architecture of the system, Fed-AdScale ensures that algorithmic governance is not an optional feature but a core structural property.

7. Policy Implications and Global Data Governance

The deployment of Fed-AdScale has profound implications for global data policy and the legal definition of data ownership. As the infrastructure moves data processing to the edge and secures cloud interactions with TEEs, the traditional "custodianship" model of data protection becomes obsolete. If a company never possesses the raw data of its users, the legal liabilities associated with data loss are fundamentally transformed. This shift supports the concept of "data sovereignty," where users maintain control over their digital identities and only share the minimal amount of information necessary for a specific transaction. Policy-makers can leverage the technical guarantees of Fed-AdScale to craft more effective regulations that focus on hardware-verified privacy rather than after-the-clock legal audits.

The cross-border nature of social commerce also introduces challenges related to "data residency" laws, which require that the data of citizens be stored and processed within national borders. Fed-AdScale's hierarchical design naturally supports these requirements by utilizing regional aggregators that act as local data anchors. A user's private data never needs to cross an international border to be processed; only the anonymous, encrypted gradients are sent to the global cloud core for model improvement. This provides a technical pathway for

global social commerce platforms to operate in highly regulated markets without needing to build completely separate, siloed infrastructures for each country.

However, the decentralization of AI also raises new questions about accountability. If an autonomous advertising agent generates content that violates local advertising standards or consumer protection laws, the distributed nature of the system can make it difficult to assign blame. Fed-AdScale addresses this by maintaining a hardware-attested log of all model versions and policy configurations used in the inference process. This "black box" for AI decisions provides a tamper-proof record that can be used by regulators in the event of a dispute. This illustrates the system's role as a bridge between technical innovation and legal oversight, providing the tools necessary for a transparent and accountable digital economy.

8. Case Illustrations and Cross-Domain Applications

To understand the practical utility of Fed-AdScale, we can consider the case of a cross-border social commerce event, such as a "flash sale" orchestrated through an international messaging app. During such an event, the system must handle a sudden surge of millions of queries. Fed-AdScale's high-throughput architecture manages this by offloading the majority of the linguistic analysis to the edge, while using the regional TEE-equipped servers to handle the complex task of matching the social context of the conversation with the available product inventory. The use of secure enclaves ensures that even during a high-pressure commercial event, the users' private discussions about the products remain strictly confidential.

The principles of Fed-AdScale are also applicable beyond the domain of social commerce. In the field of personalized medicine, for example, a similar infrastructure could be used to provide real-time health recommendations based on data from wearable devices. The sensitive medical data would stay on the device or within a regional secure enclave, while a global model for disease prediction is updated through federated learning. Similarly, in the smart city context, Fed-AdScale could manage the massive streams of data from urban sensors to optimize traffic flow or energy distribution while protecting the privacy of individual citizens. These cross-domain comparisons highlight the versatility of the edge-cloud TEE framework as a general solution for the challenges of the "data-intensive" society.

The forward-looking perspective of this research involves the transition toward "autonomous commercial agents" that act as fiduciaries for the user. In such a future, a user's local LLM agent would not only receive advertisements but would actively negotiate with seller agents on the user's behalf, finding the best deals and managing transactions within a secure, decentralized environment. Fed-AdScale provides the necessary infrastructure for this transition, ensuring that as commerce becomes more autonomous, it also remains more private and secure. By establishing a robust, scalable, and secure foundation today, we are paving the way for the sophisticated and human-centric commercial ecosystems of tomorrow.

9. Conclusion

Fed-AdScale represents a significant departure from the centralized, data-extractive models that have dominated the first era of social commerce. By proposing a distributed edge-cloud

infrastructure that utilizes hardware-assisted security and federated learning, we have demonstrated that it is possible to achieve high-throughput, hyper-personalized advertising without compromising user privacy. The structural design of the system, with its hierarchical orchestration and paged-inference strategies, provides a realistic and scalable solution to the computational challenges of deploying Large Language Models at the edge. Throughout this paper, we have explored the complex interplay between engineering constraints, economic incentives, and social values, emphasizing that the success of a digital infrastructure depends as much on its governance and fairness as it does on its technical performance.

The robustness and sustainability of Fed-AdScale make it an ideal candidate for long-term deployment in the global digital economy. Its ability to adapt to varying hardware capabilities and regulatory environments ensures its relevance in a rapidly changing technological landscape. Furthermore, the inclusion of fairness-aware auditing and adversarial defenses ensures that the system promotes an equitable and secure marketplace for all participants. As we move further into the age of artificial intelligence, the principles of decentralization and hardware-verified trust will become increasingly vital. Fed-AdScale stands as a testament to the power of interdisciplinary systems research in creating a future where technology empowers the individual while fostering a vibrant and innovative commercial society.

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. Anati, I., Gueron, S., Johnson, S., & Scarlata, V. (2013). Innovative instructions and software model for isolated execution. Proceedings of the 2nd International Workshop on Hardware and Architectural Support for Security and Privacy, 10(1).

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. Costan, V., & Devadas, S. (2016). Intel SGX explained. Cryptology ePrint Archive.

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. 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. Chen, X. (2024, November). Cloud Storage User Behavior Analysis and Dynamic Replica Strategy Optimization Based on Improved RFM and Fuzzy Clustering. In International Conference on Cognitive based Information Processing and Applications (pp. 425-434). Singapore: Springer Nature Singapore.

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. Yi, X. (2026). A Federated and Differentially Private Incentive–Marketing Framework for Privacy-Preserving Cross-Channel Measurement in AI-Powered Digital Commerce.

25. 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.

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

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.

32. 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.