A Decentralized Cloud-Edge Infrastructure for Social Commerce: Accelerating Privacy-Preserving LLM Inference via Trusted Execution Environments

Elias Vance
Department of Computer Science and Information Systems, Bradley University
evance@bradley.edu

Alan Davenport
Department of Electrical Engineering and Computer Science, University of New Mexico
alandavenport@unm.edu

Douglas Redmond
Department of Management Information Systems, University of Delaware
dredmond@udel.edu

Abstract

The convergence of social networking and electronic commerce, termed social commerce, has necessitated a paradigm shift in how personalized digital experiences are delivered. At the heart of this evolution is the deployment of Large Language Models (LLMs) to facilitate context-aware interactions, personalized recommendations, and automated customer service. However, the centralization of user data required for LLM inference poses significant privacy risks and creates substantial latency bottlenecks in high-throughput environments. This paper proposes a novel decentralized cloud-edge infrastructure designed specifically for the social commerce domain. By distributing LLM inference tasks across a continuum of edge nodes and cloud servers, the proposed architecture minimizes data movement and optimizes response times. Central to this infrastructure is the integration of hardware-assisted security through Trusted Execution Environments (TEEs), which ensure that sensitive user data remains encrypted and isolated even during the inference process. We provide an exhaustive system-level analysis of this framework, emphasizing the structural trade-offs between computational overhead and security guarantees. The discussion extends to broader socio-technical implications, including infrastructure governance, environmental sustainability of decentralized AI, and the policy challenges associated with hardware-verified privacy. Our findings suggest that a tiered, TEE-augmented approach not only enhances user privacy but also provides the necessary scalability for the next generation of global social commerce platforms.

Keywords

Distributed Systems, Social Commerce, Trusted Execution Environments, Edge Computing, Large Language Models, Privacy-Preserving AI, Socio-Technical Infrastructure.

1. Introduction

The contemporary digital landscape is defined by the seamless integration of social interaction and commercial transaction. Social commerce has transitioned from a niche marketing strategy to a dominant economic force, leveraging the collective intelligence of social graphs to drive personalized consumption. This shift has been accelerated by the arrival of Large Language Models (LLMs), which allow platforms to parse complex natural language, predict consumer sentiment, and generate persuasive, context-aware content. Yet, the traditional deployment of these models relies on massive, centralized data centers that ingest vast quantities of private user data. This centralization creates a fundamental tension between the desire for hyper-personalization and the imperative for data sovereignty. As global regulations like the GDPR and CCPA become more stringent, the cost of data centralization—both in terms of legal liability and consumer trust—has become a significant barrier to innovation.

A decentralized cloud-edge infrastructure offers a compelling solution to these challenges by shifting the locus of computation closer to the data source. In a social commerce context, this means performing LLM inference on edge devices—such as smartphones, localized edge servers, or smart gateways—rather than transmitting raw behavioral data to a distant cloud. However, edge devices are notoriously resource-constrained, often lacking the memory and compute power necessary to run multi-billion parameter models in real-time. Furthermore, decentralization introduces new security vulnerabilities, as edge nodes are physically accessible and may be compromised by malicious actors. Therefore, the acceleration of LLM inference in a decentralized setting must be paired with robust, hardware-level security measures to maintain the integrity of the privacy-preserving promise.

This paper introduces an integrated architectural framework that leverages Trusted Execution Environments (TEEs) to bridge the gap between performance and privacy. TEEs, such as Intel SGX or ARM TrustZone, provide a secure enclave where code and data can be processed in isolation from the rest of the system. By deploying LLM shards within these enclaves, we can guarantee that neither the platform provider nor an unauthorized intruder can access the underlying user data during inference. The following sections provide a detailed examination of the structural design of this infrastructure, the trade-offs inherent in distributed AI deployment, and the socio-technical governance required to ensure a fair and sustainable social commerce ecosystem.

2. Architectural Design of a Tiered Cloud-Edge Continuum

The proposed infrastructure for social commerce is structured as a tiered continuum that transcends the binary distinction between cloud and edge. At the lowest tier, mobile edge devices perform initial semantic filtering and lightweight inference using distilled LLMs. These models are optimized for low-latency interactions, such as instant message auto-completion or basic product categorization. Because these devices are closest to the user, they handle the most sensitive raw data, which never leaves the local TEE. This "privacy-at-the-source" approach ensures that the primary behavioral signals are processed in a zero-trust environment, significantly reducing the attack surface of the overall platform.

The intermediate tier consists of regional edge aggregators or "fog nodes." These are high-performance servers located within proximity to major urban centers or telecommunications hubs. These nodes host larger, more capable model shards that can handle complex reasoning tasks, such as multi-modal sentiment analysis or personalized social feed curation. The regional tier acts as a buffer, aggregating anonymized updates from multiple edge devices and performing collective inference that informs the social commerce graph. The use of TEEs at this level is critical, as regional nodes handle data from thousands of users. Hardware-enforced isolation ensures that even if a regional server is physically or logically attacked, the multi-tenant data remains protected within secure enclaves.

The final tier is the global cloud core, which serves as the central orchestration and training layer. Unlike traditional models, the cloud core in our infrastructure does not handle raw user data. Instead, it manages model versioning, global parameter updates through federated learning, and high-level commercial analytics. The cloud core receives only encrypted, non-invertible embeddings or gradient updates from the lower tiers. This hierarchical distribution of labor optimizes throughput by matching the complexity of the LLM task to the available compute resources at each tier. It also creates a robust fallback mechanism; if a regional node fails, the edge devices can either revert to local inference or route through an alternative secure gateway, ensuring the continuous availability of social commerce services.

3. Accelerating LLM Inference via Trusted Execution Environments

Accelerating LLM inference within a TEE requires a sophisticated understanding of both hardware constraints and model architecture. TEEs traditionally suffer from limited memory capacity—often referred to as the "enclave memory wall"—which is particularly problematic for LLMs that require gigabytes of VRAM. To address this, our infrastructure employs a model sharding and paging strategy. By breaking the transformer blocks into smaller, executable units, the system can load and unload model weights into the secure enclave dynamically. This paging is synchronized with the inference stream, allowing the system to maintain a high-throughput rate despite the overhead of constant encryption and decryption at the enclave boundary.

The computational overhead of TEE-based inference is further mitigated through hardware-aware model quantization and pruning. By reducing the precision of the LLM weights and removing redundant neurons, we can shrink the model footprint enough to fit within the secure memory of modern edge processors. Furthermore, the infrastructure utilizes specialized AI accelerators that are integrated with the TEE's security controller. This allows for encrypted data to be passed directly from the secure memory to the GPU or NPU without being exposed to the insecure host operating system. This synergy between hardware-assisted security and AI acceleration is the key to making privacy-preserving LLMs viable for the real-time demands of social commerce.

Beyond simple performance metrics, the use of TEEs provides a unique form of "computational truthfulness" in the social commerce ecosystem. In a decentralized

environment, advertisers and platform operators need assurance that the LLM is behaving according to the agreed-upon policy—for instance, that it is not unfairly biasing certain products or engaging in predatory targeting. Through remote attestation, the TEE can provide a cryptographic proof to all stakeholders that the specific, audited model is indeed the one being executed. This transparency builds trust among users, brands, and regulators, transforming the infrastructure from a "black box" into a verifiable socio-technical system.

4. System-Level Trade-offs and Robustness Analysis

Designing a decentralized infrastructure for LLMs involves navigating complex trade-offs between latency, accuracy, and security. A purely local inference approach maximizes privacy and minimizes latency but often results in lower accuracy due to the use of highly compressed student models. Conversely, offloading all complex reasoning to the cloud maximizes accuracy but introduces significant latency and privacy risks. Our tiered architecture allows for dynamic load balancing where the system determines the optimal inference location based on the sensitivity of the request and the current network conditions. This elasticity is essential for maintaining robustness in the face of fluctuating user traffic and heterogeneous device capabilities.

The robustness of the decentralized infrastructure is also tested by adversarial attacks, such as model poisoning or membership inference. In a social commerce setting, an adversary might attempt to inject malicious data into the federated learning loop to bias the global model toward specific commercial outcomes. Fed-AdScale mitigates this by implementing a decentralized auditing layer within the TEEs. Local nodes verify the integrity of the gradient updates before they are sent to the regional aggregators, ensuring that only "honest" updates contribute to the global model. This distributed defense mechanism makes the infrastructure significantly more resilient to coordinated attacks than a centralized database.

Furthermore, we must consider the "energy-privacy trade-off." The computational cost of running TEE-based inference on millions of edge devices is substantial. Encryption, memory paging, and the constant state-switching required for secure enclaves consume more power than traditional inference. In a world increasingly focused on green computing, the infrastructure must be optimized for sustainability. We propose a "frugal-inference" policy where high-intensity secure computation is only triggered when a specific privacy threshold is met. For low-sensitivity tasks, the system can revert to more energy-efficient, non-TEE paths, provided that the user has consented to a lower level of isolation. This nuanced approach to energy management is a critical component of the infrastructure's long-term viability.

5. Infrastructure Governance and Socio-Technical Alignment

The deployment of a decentralized AI infrastructure for social commerce is as much a governance challenge as it is a technical one. In a centralized system, the platform owner holds all the power and responsibility. In a decentralized, TEE-based system, authority is fragmented across hardware manufacturers, model developers, regional node operators, and the users themselves. Establishing a clear governance framework is essential for resolving disputes, managing updates, and ensuring accountability. We advocate for a

"consortium-based" governance model where key stakeholders participate in a decentralized autonomous organization (DAO) that oversees the infrastructure's technical standards and ethical guidelines.

A primary goal of this governance is to ensure algorithmic fairness. Social commerce platforms are often accused of creating "filter bubbles" or perpetuating biases through their recommendation engines. By decentralizing the inference process, we have an opportunity to embed fairness checks directly into the local TEEs. Local agents can be programmed to audit the LLM's output for signs of discriminatory bias before the content is ever presented to the user. This "on-device auditing" shifts the responsibility of fairness from a central, often opaque, authority to a transparent, hardware-verified process. However, this requires a standardized set of fairness metrics that can be interpreted by the LLM agents across the entire continuum.

Furthermore, the governance framework must address the "digital divide" created by high-end hardware requirements. TEE-based LLM inference requires modern, expensive processors, which may exclude users in developing regions or lower-income demographics from accessing the most secure features of the social commerce platform. To promote equity, the infrastructure should support "delegated secure computation," where users with older devices can lease secure enclave time from a trusted regional provider. The policy implications of this are significant; it requires a new type of regulatory oversight that focuses on "hardware-as-a-service" and ensures that security is a universal right rather than a luxury good.

6. Deployment, Scalability, and Global Policy Implications

The global deployment of a decentralized cloud-edge infrastructure for social commerce faces significant logistical and regulatory hurdles. Different regions have vastly different standards for data residency and privacy protection. The beauty of a decentralized TEE-based approach is its inherent adaptability to these localized requirements. Because the data is processed locally and the cloud only receives anonymized insights, the infrastructure can naturally comply with data localization laws without requiring a complete redesign for every country. This "compliance-by-design" is a major strategic advantage for global social commerce players who wish to enter highly regulated markets.

Scalability is achieved through the modular nature of the edge-cloud continuum. As the user base grows, the platform operator can simply deploy more regional aggregator nodes or incentivize users to contribute their own edge compute power to the network. This "horizontal scaling" is far more cost-effective than building increasingly massive centralized data centers. Moreover, the use of LLMs allows the infrastructure to handle a diverse range of languages and cultural contexts through a single, unified architecture. The model can be fine-tuned locally for specific dialects or social norms, ensuring that the social commerce experience is truly global in scope but local in feel.

From a policy perspective, the rise of hardware-verified privacy may lead to a shift in how

data protection is enforced. Instead of auditing a company's internal data practices—which is a slow and often reactive process—regulators could instead audit the "attestation reports" produced by the TEEs. If a platform can prove through a cryptographic chain of trust that it never had access to raw user data, it may be granted a "safe harbor" from certain privacy liabilities. This proactive approach to regulation could accelerate the adoption of privacy-enhancing technologies across the entire digital economy, not just within social commerce. However, it also places an enormous amount of power in the hands of a few hardware manufacturers, necessitating a broader policy discussion on the sovereignty of the underlying silicon.

7. Sustainability and the Environmental Footprint of Decentralized AI

The transition toward a decentralized AI infrastructure necessitates a rigorous evaluation of its environmental impact. While decentralization can reduce the energy requirements for data transmission and cooling in centralized facilities, the aggregate energy consumption of millions of edge devices performing LLM inference is potentially massive. To address this, we propose an "eco-centric" deployment strategy. This involves the use of "carbon-aware" task scheduling, where heavy LLM reasoning is scheduled during periods of high renewable energy availability in the user's local grid. Furthermore, the regional nodes can be designed to use waste heat for local industrial or residential purposes, creating a "circular energy" model for the social commerce infrastructure.

The sustainability of the infrastructure is also tied to the lifecycle of the hardware. The rapid pace of AI innovation often leads to "forced obsolescence," where older devices are unable to run the latest privacy-preserving models. Our architecture mitigates this through a "dynamic distillation" pipeline. When the global model is updated, the cloud core generates a range of student models tailored to different hardware generations. This ensures that even older devices can participate in the secure social commerce ecosystem, extending the useful life of the hardware and reducing electronic waste. This focus on "long-life" infrastructure is a core tenet of our socio-technical design.

Finally, we must consider the "informational sustainability" of the social commerce ecosystem. In an era of AI-generated content, there is a risk of the social graph becoming flooded with low-quality or manipulative messages. By using LLMs as "semantic gatekeepers" at the edge, the infrastructure can help maintain the quality and authenticity of the social interactions. The local agent can verify the factual accuracy of a claim or detect the markers of a bot-driven influence campaign before the content is shared. This preservation of the informational commons is essential for the long-term health of the social commerce market.

8. Cross-Domain Comparisons and Forward-Looking Perspectives

When compared to other decentralized architectures, such as blockchain-based social networks, the TEE-augmented cloud-edge model offers superior performance and scalability for AI-intensive tasks. Blockchain is excellent for maintaining a transparent ledger of transactions but is notoriously inefficient for the high-throughput, low-latency requirements

of LLM inference. By using TEEs for the compute-heavy parts of the system and a lightweight distributed ledger only for the final commercial settlements, we can achieve the best of both worlds: the cognitive power of modern AI and the trust-less transparency of decentralization.

Looking forward, we anticipate the emergence of "personal AI fiduciaries"—agents that live entirely on the user's device and act as their advocate in the social commerce marketplace. These fiduciaries would not only manage the user's privacy but would also proactively negotiate with brand agents to find the best products at the lowest prices. In this "agent-to-agent" economy, the infrastructure's role shifts from a content delivery network to a secure negotiation platform. The cloud-edge continuum, secured by TEEs, provides the necessary "neutral ground" where these autonomous agents can interact without the risk of exploitation by a centralized intermediary.

The ultimate goal of this research is to provide a blueprint for a more human-centric digital economy. By decentralizing the power of LLMs and securing them with hardware-level guarantees, we can create a social commerce infrastructure that values the individual as much as the transaction. This requires a sustained commitment to interdisciplinary research, bridging the gap between computer science, economics, and law. As we move toward a world where AI is ubiquitous, the principles of decentralization and hardware-verified trust will be the foundation upon which a more equitable and resilient society is built.

9. Conclusion

This paper has proposed a decentralized cloud-edge infrastructure for social commerce that prioritizes user privacy through the acceleration of LLM inference within Trusted Execution Environments. We have detailed a tiered architectural framework that balances the computational demands of large-scale AI with the resource constraints of edge devices. By shifting inference to the edge and using TEEs to isolate sensitive data, we have demonstrated a technical path toward hyper-personalized social commerce that respects data sovereignty and complies with global privacy regulations.

The socio-technical analysis provided throughout the paper has highlighted the critical importance of infrastructure governance, environmental sustainability, and algorithmic fairness. We have argued that the transition toward decentralized AI must be accompanied by a new approach to policy and regulation—one that leverages hardware-verified trust to create a more transparent and accountable marketplace. As social commerce continues to expand, the integration of TEE-based LLM inference will be a key differentiator for platforms that seek to build long-term trust with their users.

Ultimately, the Fed-AdScale infrastructure represents a significant step forward in the evolution of distributed systems. It challenges the prevailing "data-extractive" model of AI and offers a more sustainable and equitable alternative. By continuing to refine these architectural patterns and fostering a global dialogue on the ethics of decentralized AI, we can ensure that the next generation of digital platforms is both high-performing and

fundamentally human-centric.

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. 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).

4. 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.

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

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

8. 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.

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

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

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

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

13. 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.

14. 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.

15. 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.

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

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

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

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

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

21. 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.

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

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

24. 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.

25. Yi, X. (2026). Privacy-Enhanced Ad Targeting for Social E-Commerce: A Federated Learning Framework with Zero-Knowledge Verification for Creator Monetization. Frontiers in Business and Finance, 3(1), 102-113.

26. 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.

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

28. 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.

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

30. 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.

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

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

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.