International Journal of Artificial Intelligence, Data Science, and Machine Learning
Grace Horizon Publication | Volume 6, Issue 4, 59-64, 2025
ISSN: 3050-9262 | https://doi.org/10.63282/3050-9262.IJAIDSML-V6I4P108

Original Article

Federated Learning Infrastructure for Privacy-Preserving Personalized Shopping in E-Commerce

Udit Agarwal[1], Aditya Gupta[2]
[1,2] *Independent Researcher, USA.*

Received On: 01/09/2025 ***Revised On: 05/10/2025*** ***Accepted On: 12/10/2025*** ***Published On: 01/11/2025***

Abstract *- The modern e-commerce landscape is defined by a fundamental tension between the strategic imperative for real-time, personalized user experiences and the dual constraints of stringent privacy regulations and high-latency computation. While personalized recommendations are critical for navigating information overload and driving engagement, traditional centralized architectures pose significant data privacy risks. Federated Learning (FL) has emerged as a foundational, privacy-preserving paradigm by training models on decentralized user data. However, standard FL implementations suffer from a severe communication bottleneck, introducing latency that is prohibitive for time-sensitive applications like product recommendations at checkout. This paper addresses this performance gap by introducing a new class of tooling exemplified by the Low-Latency Federated Learning (LoLaFL) framework. It details the architectural shift from conventional backpropagation-based FL to a novel forward-only, white-box framework. This approach dramatically reduces communication overhead and computational complexity, enabling the low-latency inference required to deliver secure, private, and instantaneous personalized recommendations at the most critical point of the customer journey.*

Keywords *- Federated Learning, Low-Latency AI, Personalized Recommendations, E-commerce, Privacy-Preserving Machine Learning, Real-Time Inference, LoLaFL.*

1. Introduction

1.1. The Dual Mandate in Modern E-commerce: Personalization and Privacy

The contemporary digital marketplace presents e-commerce platforms with a significant information overload problem. As the volume of available products and online content outpaces consumer capacity, Personalized Recommendation Systems (PRS) have become indispensable tools for filtering vast amounts of dynamically generated data to match user interests and behaviors.[1] These systems are no longer a luxury but a critical component for enhancing the shopping experience, driving sales growth, and improving customer engagement. However, the architectural foundation of traditional PRS creates a direct and pressing conflict with the growing global demand for user privacy. These systems have historically relied on centralized servers to aggregate, store, and process massive volumes of sensitive consumer data, such as purchase histories and browsing patterns.[1] This centralized approach, while effective for training powerful machine learning models, simultaneously creates a single, high-value target for data breaches and exposes user information to potential misuse, such as unauthorized sale to third-party companies.[1] This creates an inherent architectural conflict: the very mechanism of data aggregation that has traditionally been used to optimize model performance is the same one that generates the greatest privacy vulnerability. Resolving this requires more than incremental security improvements; it demands a fundamental paradigm shift in system design.

1.2. The Regulatory Landscape and the Shift towards Privacy-by-Design

This architectural tension is amplified by an increasingly stringent global regulatory environment. Legislative frameworks such as the General Data Protection Regulation (GDPR) in Europe and the California Consumer Privacy Act (CCPA) have established rigorous standards for data handling and user consent.[6] More recently, the European Union's Digital Markets Act (DMA) has introduced a proactive, *ex-ante* regulatory model targeting large online platforms designated as "gatekeepers".[8]

Unlike previous regulations that intervened after a violation occurred, the DMA establishes clear, upfront obligations and prohibitions to prevent harmful practices before they happen.[9] These rules explicitly forbid the misuse of data generated by business users to compete against them and mandate fair access and data portability, fundamentally altering how platforms can leverage user data.[9] Non-compliance carries severe financial penalties, with fines reaching up to 10% of a company's total global revenue.[8]

This regulatory pressure acts as a powerful catalyst for architectural innovation. It makes privacy-preserving frameworks a strategic necessity for maintaining market access and competitive advantage; particularly within the EU.[8] The adoption of a privacy-by-design architecture is no longer just a technical choice but a direct response to legal prohibitions that threaten core business models. This environment compels a transition away from reactive security measures toward systems where privacy is an integral and foundational component of their structure.

1.3. Federated Learning as a Foundational Paradigm for Private Recommendations

Federated Learning (FL) has emerged as the leading architectural solution to this personalization-versus-privacy dilemma. FL is a distributed machine learning paradigm that decouples the ability to perform machine learning from the need to store user data in a centralized cloud environment.[10] The approach inverts the traditional model by "bringing computation to the data" rather than the other way around.[10]

In an FL-based recommender system, a global model is distributed from a central server to client devices (e.g., smartphones or personal computers).[6] This model is then trained locally on each device using the user's private data, such as their interaction logs or purchase history.[11] Crucially, the raw data never leaves the device. Instead, only the resulting model updates—such as anonymized gradient vectors are encrypted and sent back to the central server for aggregation.[11] This decentralized process inherently protects user data, aligns with the principles of data minimization and purpose limitation found in regulations like the GDPR, and helps build user trust.[3]

1.4. Thesis: The Latency Gap and the Need for Specialized Tooling

While FL provides a robust solution to the privacy challenge, this paper argues that standard implementations are ill-suited for high-stakes, real-time e-commerce interactions, most notably personalized recommendations at the point of checkout. Traditional FL, particularly when applied to complex deep neural networks, suffers from a significant "communication bottleneck".[13] The process of transmitting high-dimensional model parameters from thousands or millions of devices, combined with the numerous iterative communication rounds required for model convergence, introduces unacceptable levels of latency.[14]

This latency gap renders conventional FL impractical for scenarios where a response is needed in milliseconds, not seconds or minutes. This paper posits that a new class of tooling is required to make privacy-preserving AI viable for these time-sensitive applications. It introduces the Low-Latency Federated Learning (LoLaFL) framework as a solution that fundamentally re-architects the learning process. By replacing iterative backpropagation with a deterministic, forward-only propagation method, LoLaFL overcomes the latency barrier, making privacy-preserving, real-time inference at checkout a feasible and strategically valuable reality.

2. Federated Architectures for Privacy-Preserving Recommendations

2.1. Core Principles and Workflow of Federated Recommender Systems (FedRS)

A Federated Recommender System (FedRS) applies the principles of FL to the task of generating personalized suggestions. The architecture is orchestrated by a central server that coordinates training rounds across a distributed network of client devices without accessing their raw data.[10] The workflow follows a distinct, iterative cycle composed of four key phases:

- Initialization and Distribution: The process begins on the central server, which initializes a global recommendation model (e.g., a neural network designed to predict user preferences). This initial model is then distributed to a selected subset of client devices to participate in a training round.[10]
- Local Training: Upon receiving the global model, each client device trains it using its own local and private data. For an e-commerce application, this data could include a user's click history, items added to a cart, or past purchases. This step ensures that sensitive information never leaves the user's device.[6]
- Update Computation and Transmission: After local training, each client computes a summary of the changes made to the model. This update, typically in the form of model weights or gradient vectors, encapsulates the learning from the local data without revealing the data itself. The update is then securely transmitted back to the server.[10]
- Secure Aggregation: The server collects the updates from the participating clients and aggregates them to produce an improved global model. The most common aggregation algorithm is Federated Averaging (FedAvg), which computes a weighted average of the client updates.[6]

This cycle is repeated, with the newly refined global model being sent out for subsequent rounds of training. Through this iterative process, the model collaboratively learns from the collective intelligence of all participating devices, leading to scalable and highly personalized recommendations while preserving user privacy by design.[3]

2.2. Enhancing Privacy Guarantees with Advanced Security Mechanisms

While data localization is the foundational privacy guarantee of FL, it is a necessary but insufficient condition for ensuring comprehensive data protection. Sophisticated adversaries could potentially infer sensitive information from the model updates transmitted to the server through so-called reconstruction attacks.[7] Therefore, a robust FedRS architecture must implement a multi-layered, defense-in-depth strategy that integrates additional privacy-preserving techniques (PPTs) to protect not just the stored data, but the information revealed during the learning process itself.

This multi-layered security imperative is achieved by combining data localization with cryptographic and statistical methods that protect the communication and aggregation phases of the FL workflow. Key techniques include:

- Differential Privacy: This is a statistical technique that provides a formal, mathematical guarantee of privacy. It involves adding a carefully calibrated amount of statistical noise to the model updates on the client device *before* they are transmitted to the server. This noise makes the contribution of any single user statistically indistinguishable from the rest, effectively masking individual data points and protecting against inference attacks.[7]

- Secure Multi-Party Computation (SMC) and Homomorphic Encryption (HE): These are advanced cryptographic techniques that protect data while it is being processed. With homomorphic encryption, model updates are encrypted on the client device, and the server can perform aggregation (e.g., summing the updates) directly on the encrypted data without ever decrypting it. This ensures that the individual client contributions remain hidden even from the central server coordinating the training.[7]
- Gradient Clipping: This is a client-side mechanism that helps mitigate information leakage from gradient updates. Before an update is sent, its vector norm is capped at a predefined threshold. This prevents unusually large gradients, which might correspond to unique or sensitive data points, from revealing information about the underlying training data.[17]

By layering these techniques, a FedRS can build a truly robust privacy framework. Data localization serves as the first line of defense, while PPTs like differential privacy and homomorphic encryption provide essential second and third layers that secure the collaborative training process itself.

3. Overcoming Latency Bottlenecks for Real-Time Inference

3.1. *The Communication Bottleneck in Traditional Federated Learning*

Despite its architectural strengths for privacy, traditional FL faces a critical performance limitation when applied to complex models in real-world settings: the communication bottleneck.[13] This issue is particularly acute when using deep neural networks for tasks like recommendation, as these models are computationally intensive and communication-heavy, making them ill-suited for applications that demand near-instantaneous inference.[14]

The bottleneck arises from the confluence of two primary factors:

- High-Dimensionality of Model Updates: Modern deep learning models can contain millions or even billions of parameters. In each round of traditional FL, the full set of these parameters, or their corresponding gradients, must be transmitted from each participating client device to the central server. This massive data payload consumes significant network bandwidth and introduces substantial transmission delays.[14]
- Numerous Iterative Communication Rounds: The training process for deep neural networks is typically based on iterative optimization algorithms like stochastic gradient descent. Achieving model convergence requires a large number of these back-and-forth communication rounds between the server and clients, with each round adding to the cumulative latency.[13]

The combined effect of these two factors results in high end-to-end latency, rendering traditional FL impractical for time-sensitive use cases. For an e-commerce platform, the goal of providing a personalized product recommendation during the few seconds a customer is at a checkout screen is simply not feasible with this architecture.

3.2. LoLaFL: A Forward-Only Propagation Framework for Low-Latency Inference

The Low-Latency Federated Learning (LoLaFL) framework is a novel approach designed specifically to dismantle this communication bottleneck.[14] It achieves this through a fundamental paradigm shift, moving away from the conventional "black-box" neural network trained via backpropagation to a "white-box" model built using a deterministic, forward-only process.

The key architectural innovations of LoLaFL are:

- Forward-Only Propagation: Unlike traditional models, LoLaFL models are constructed layer-by-layer in a single forward pass. This design completely eliminates the need for backpropagation, a computationally expensive and iterative process that is a primary source of complexity and latency in standard deep learning.[14]
- Deterministic Parameter Calculation: In a LoLaFL model, the parameters for each new layer are not learned through iterative optimization. Instead, they are calculated directly and deterministically from the features of the preceding layer using predefined mathematical formulae.[14]
- Layer-wise Transmission: The most significant advantage for latency reduction is that the model is built and transmitted one layer at a time. In each communication round, clients only need to compute and transmit the parameters for the *single, most recent layer*, rather than the entire model. This dramatically reduces the size of the data payload in each round, directly attacking the communication bottleneck.[14]

The following table provides a clear comparison of the architectural differences between traditional FL and the LoLaFL framework.

Table 1: Comparative Analysis of Traditional FL and LoLaFL Frameworks

Feature	Traditional FL	LoLaFL (Low-Latency Federated Learning)
Model Architecture	Black-box (parameters learned implicitly)	White-box (parameters calculated deterministically)
Propagation Method	Backpropagation	Forward-only Propagation

Training Objective	Minimize Loss Function	Learn Linear Discriminative Features
Data Transmitted per Round	Entire Model Parameters/Gradients	Single Layer Parameters / Covariance Matrices
Aggregation Algorithm	Linear (e.g., Federated Averaging)	Non-linear (Harmonic-mean-like)
Latency	High (due to model size and rounds)	Low (over 90% reduction)
Robustness to Non-IID Data	Challenged	More Robust

3.3. Analysis of Novel Non-Linear Aggregation Schemes

The unique, deterministic structure of LoLaFL renders standard aggregation algorithms like FedAvg, which is based on a simple linear arithmetic mean, suboptimal. Research has demonstrated that the optimal aggregation method for the global parameters in a LoLaFL system is non-linear and best approximated by a harmonic-mean-like (HM-like) function.[14]

Based on this finding, two novel non-linear aggregation schemes have been developed for the LoLaFL framework:

- HM-like Aggregation: This scheme directly implements the harmonic-mean-like principle to aggregate the layer parameters sent by the clients. It provides a mathematically sound method for combining the deterministically calculated parameters into a robust global layer.
- CM-based Aggregation: This second scheme introduces a further optimization to reduce latency. It leverages the low-rank structures often present in high-dimensional feature data. Instead of transmitting the layer parameters themselves, clients compute and transmit low-rank-approximated covariance matrices of their local features. This acts as a form of compression, and the server can reconstruct the aggregated global parameters from these matrices. The use of techniques like Singular Value Decomposition (SVD) for this approximation provides an additional layer of latency reduction.[14]

3.4. Performance Implications for Time-Sensitive Applications

The architectural redesign embodied by LoLaFL and its non-linear aggregation schemes yields radical improvements in performance, making it highly suitable for time-sensitive applications. Quantitative analysis shows that, compared to traditional FL frameworks, LoLaFL delivers:

- Drastic Latency Reduction: The HM-like aggregation scheme achieves a latency reduction of over 91%, while the more advanced CM-based scheme reduces latency by over 98%.[14]
- Accelerated Convergence: LoLaFL converges to an effective model up to ten times faster when measured in terms of the number of communication rounds required.[14]
- Comparable Accuracy and Enhanced Robustness: These significant performance gains are achieved while maintaining model accuracy comparable to that of traditional deep learning-based FL. Furthermore, LoLaFL demonstrates greater robustness to non-IID (non-independent and identically distributed) data, a common and difficult challenge in real-world federated settings where user data varies significantly from device to device.[14]
- Holistic Evaluation: In addition to latency and convergence, privacy guarantees are quantified through differential privacy parameters (ε, δ), and recommendation quality is evaluated using ranking metrics such as Precision@K and NDCG, providing a comprehensive assessment of privacy-preserving personalized recommendations.

This shift introduces a new performance trade-off paradigm for system optimization. In traditional FL, latency is primarily a function of the model's size and complexity; improving a model by making it larger directly increases latency. In contrast, the latency and computational complexity of LoLaFL are primarily determined by the dimensionality of the input data (d) and the number of classes (J) in the dataset.[14] This means that for many recommendation tasks with moderate data dimensionality, LoLaFL offers a substantial performance advantage. Consequently, the engineering focus for optimization shifts from model compression techniques (like quantization) to feature selection and dimensionality reduction methods aimed at minimizing 'd'.

4. Conclusion

4.1. Synthesizing Privacy and Performance for the Next Generation of E-commerce

The evolution of e-commerce is shaped by the powerful and often conflicting demands of deep personalization and stringent data privacy. While Federated Learning provides a robust and elegant architectural foundation for delivering privacy-preserving recommendations, its standard implementation fails to meet the critical low-latency requirements of real-time customer interactions, such as those at the checkout. The resulting performance gap has, until now, left a crucial business need unmet. The LoLaFL framework and its associated tooling effectively resolve this conflict by fundamentally re-engineering the federated training process. By abandoning the iterative, communication-heavy backpropagation method in favor of a deterministic, forward-only, and layer-wise training architecture, LoLaFL dismantles the communication bottleneck that plagues traditional FL. This approach drastically reduces latency and accelerates convergence by orders of magnitude, all without compromising the accuracy of the final model or the privacy of the underlying user data.

4.2. The Strategic Value of Low-Latency, Federated Tooling

The LoLaFL framework should be viewed not merely as a technical improvement but as a strategic enabler for the next generation of e-commerce. By providing a tool that is both privacy-compliant by design and fast enough for real-time inference, it empowers e-commerce platforms to achieve several critical business objectives simultaneously. First, it allows platforms to confidently deploy sophisticated personalization features in an increasingly strict regulatory environment, mitigating the risks associated with frameworks like the DMA and GDPR. Second, it enables the enhancement of the customer experience at the most critical point of the sales funnel—the checkout—where speed, relevance, and a frictionless process are paramount to converting a sale and fostering loyalty. Finally, by solving the latency problem, LoLaFL unlocks a host of new opportunities for on-device AI applications that were previously infeasible. The adoption of such specialized, low-latency federated tooling represents the next logical step in the evolution of intelligent, user-centric, and trustworthy digital commerce.

References

[1] 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.

[2] Konečný, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., & Bacon, D. (2016). Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*.

[3] McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics* (pp. 1273-1282). PMLR.

[4] Park, J., et al. (2025). Federated Recommender System with Data Valuation for E-commerce Platform. *arXiv preprint*.

[5] Sun, Z., Xu, Y., Liu, Y., He, W., Jiang, Y., Wu, F., & Cui, L. (2023). A Survey on Federated Recommendation Systems. *IEEE Transactions on Neural Networks and Learning Systems*.

[6] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N.,... & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (pp. 5998-6008).

[7] Yang, Q., Liu, Y., Cheng, Y., Kang, Y., Chen, T., & Yu, H. (2019). Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST), 10*(2), 1-19.

[8] Federated Learning-based Personalized Recommendation Systems: An Overview on Security and Privacy Challenges - CyberSecDome, accessed October 29, 2025, https://cybersecdome.eu/wp-content/uploads/2024/01/IEEE-Transactions-on-Consumer-El ectronics-Federated-Learning-based-Personalized-Recommendati-2023.pdf

[9] (PDF) E-commerce Personalized Recommendations: a Deep Neural Collaborative Filtering Approach - ResearchGate, accessed October 29, 2025,

[10] https://www.researchgate.net/publication/377081826_E-commerce_Personalized_Recommendations_a_Deep_Neural_Collaborative_Filtering_Approach

[11] (PDF) Federated Learning on Recommender Systems - ResearchGate, accessed October 29, 2025,

[12] https://www.researchgate.net/publication/388088244_Federated_Learning_on_Recommen der_Systems

[13] Federated Learning on Recommender Systems - IEEE Computer Society, accessed October 29, 2025,

[14] https://www.computer.org/csdl/proceedings article/bigdata/2024/10825895/23yjUjOpcOY

[15] (PDF) A Survey on Federated Recommendation Systems, accessed October 29, 2025,

[16] https://www.researchgate.net/publication/366821201_A_Survey_on_Federated_Recomme ndation_Systems

[17] Recommendation Systems Using Federated Learning - Meegle, accessed October 29, 2025,

[18] https://www.meegle.com/en_us/topics/recommendation-algorithms/recommendation-syste ms-using-federated-learning

[19] Analysis of Privacy Preservation Enhancements in Federated Learning Frameworks - Shaping the Future of IoT with Edge Intelligence - NCBI, accessed October 29, 2025, https://www.ncbi.nlm.nih.gov/books/NBK602365/

[20] Digital Markets Act Summary: EU DMA Law Explained - Usercentrics, accessed October 29, 2025,

[21] https://usercentrics.com/knowledge-hub/digital-markets-act-dma-impacts-user-privacy-and-consent-management/

[22] The Digital Markets Act: Shaping Fair Competition in the Digital Age, accessed October 29, 2025, https://business.trustedshops.com/blog/digital-markets-act

[23] Federated Learning: The Decentralized Revolution Transforming AI While Preserving Privacy | by Nicolasseverino | Oct, 2025 | Medium, accessed October 29, 2025,

[24] https://medium.com/@nicolasseverino/federated-learning-the-decentralized-revolution-transforming-ai-while-preserving-privacy-2e0a0122d8b8

[25] How is federated learning used in personalized recommendations?, accessed October 29, 2025,

[26] https://milvus.io/ai-quick-reference/how-is-federated-learning-used-in-personalized-recom mendations

[27] Federated Learning: A Privacy-Preserving Approach to ... - Netguru, accessed October 29, 2025, https://www.netguru.com/blog/federated-learning

[28] Low-Latency Collaborative Predictive Maintenance: Over-the-Air Federated Learning in Noisy Industrial Environments - MDPI, accessed October 29, 2025, https://www.mdpi.com/1424-8220/23/18/7840

[29] LoLaFL: Low-Latency Federated Learning via Forward-only ..., accessed October 29, 2025, https://arxiv.org/abs/2412.14668

[30] Privacy-Preserving Federated Learning - Hasso-Plattner-Institut, accessed October 29, 2025, https://hpi.de/arnrich/research-areas/privacy-preserving-

federated-learning.html

[31] (PDF) Federated Learning Architectures for Privacy-Preserving Artificial Intelligence Applications on Edge Devices - ResearchGate, accessed October 29, 2025,

[32] https://www.researchgate.net/publication/392749199_Federated_Learning_Architectures_f or_Privacy-Preserving_Artificial_Intelligence_Applications_on_Edge_Devices

[33] Federated Learning for Cybersecurity: A Privacy-Preserving Approach, accessed October 29, 2025, https://www.mdpi.com/2076-3417/15/12/6878