

A Study on Trustworthy Aggregation Mechanisms for Blockchain-Based Federated Learning

Zhixin Zhang *

Sydney Smart Technology College, Northeastern University, Qinhuangdao, Hebei, 066004, China

* Corresponding author Email: zzx200645@163.com

Abstract: To investigate how blockchain technology can improve the reliability, speed and legality of federated learning, this paper concentrates on the process of model combination in federated learning. By dealing with the problems of single vulnerable points and insufficient trust in centralized combination, the paper comprehensively studies the technical methods of typical solutions like BlockFL, FedChain and LBFL, and develops a four-dimensional evaluation system including decentralization, system performance, security and robustness, and privacy disclosure. Three main conflicts have been found: the contradiction between the immutability on chain and the principle of data minimization; the conflict between high-level confidentiality protection and scarce resources on edge devices; and the potential centralization risk under apparent decentralization. According to these results, this paper puts forward a lightweight cooperative design concept named "on-chain evidence storage and off-chain computation", which serves as a theoretical basis and practical approach for trustworthy federated learning in very sensitive situations.

Keywords: Federated Learning; Blockchain; Model Aggregation; Privacy Protection.

1. Introduction

Data has become the fourth factor of production in human society, but at the same time, issues surrounding data privacy and security have become increasingly prominent. As one of the key areas of technological focus, machine learning also faces a growing risk of data leaks during the data aggregation process. In response to this challenge, federated learning has emerged as an innovative privacy-preserving machine learning technique. This technology enables multiple participants to collaboratively train models without sharing their local raw data and has already demonstrated tremendous potential in highly sensitive fields such as healthcare, finance, and intelligent transportation. However, traditional federated learning models rely on a trusted central node for model aggregation and incentive distribution; if this central node is compromised or maliciously manipulated, the entire privacy protection system fails.

With the help of its distributed database, the autonomous execution of the smart contracts and the safe transactions, blockchain technology can offer a reliable base for the data collaboration in machine learning, featuring authentic operation logs and fair reward systems. Taking advantage of these features, researchers have recently studied the implementation of blockchain in strengthening the confidentiality mechanisms of machine learning. Kim et al. put forward the BlockFL method, in which they substituted the central servers with blockchain for the combination of the federated learning models and adopted the proof-of-work (PoW) consensus algorithm to check the modification of gradients[1]; Wang et al.'s FedChain system regarded the federated learning task as a 'proof of useful work'. By integrating the decentralized secret sharing with the zero-knowledge verification techniques, it guarantees both the model privacy and the enhanced consensus efficiency[2]; Qiao et al. developed the LBFL system, which applied the proof-of-contribution (PoC) consensus mechanism to evaluate the accuracy of the model updates as the consensus

weights, thus promoting the fair and incentive-based contributions and efficient summarization [3].

Although these methods have made certain advances in related areas, the majority of the current research concentrates on the execution of a single technical route, rather than carrying out a comprehensive assessment of the aggregation mechanisms concerning multiple dimensions such as privacy preservation, system efficiency and the level of decentralization.

To solve these problems, this paper concentrates on the trusted aggregation mechanisms based on blockchain for federated learning, establishes an evaluation system including four aspects, and comprehensively investigates the technical features and advantages/disadvantages of the common methods. On this basis, the paper discovers three main conflicts and puts forward the lightweight cooperative design principle of "on-chain evidence preservation and off-chain computation", offering theoretical support and feasible ways for the trustworthy federated learning in high sensitive situations.

2. Technical Foundations

2.1. A Multidimensional Definition of Trustworthiness

In federated learning, "trustworthy aggregation" not only ensures that the aggregated results are correct, but also makes the entire aggregation process verifiable and its behavior traceable, preventing any possible malicious influence from any participant. Differently from the traditional centralized distributed systems, "trust" in federated learning shows several multi-dimensional features, involving the correctness of results, the transparency of processes, and the fairness of incentives. Therefore, this section establishes a multidimensional conceptual framework, decomposing 'trusted aggregation' into four core dimensions: correctness, auditability, fairness, and collusion resistance. These definitions serve as the theoretical basis for the subsequent

analysis of system mechanisms."

Correctness requires that the aggregated result should exactly represent the legitimate updates made by all the trustworthy participants. But Byzantine attacks (for example, backdoor attacks, gradient alteration and uploading random noise) may destroy this characteristic. Therefore, measures like robust aggregation methods, anomaly detection and weighted aggregation based on contributions are necessary. For instance, Blanchard et al. proposed the "Krum" algorithm to select a reliable update vector in the presence of Byzantine workers [4]. Similarly, Yin et al. introduced "Byzantine-robust distributed learning" using median or trimmed mean [5]. The research of Bagdasaryan et al. shows that malicious clients can implant backdoors into the gradient updates without influencing the overall performance of the main task of the global model, which clearly indicates the necessity of guaranteeing the correctness of the aggregation process.[6]

The audit requirements in federated learning need to keep records of all the aggregation procedures, such as the participant identities, participation times and hashes of the uploaded data, for the verification and traceability by outsiders. This feature allows the concerned people to find out the source of the problem and distribute the responsibility when the model performance is different from the desired results.

To accomplish this objective, researchers have put forward several technical methods. As noted by Atzei et al., the inherent immutability and transparency of blockchain provide a solid foundation for building auditable and traceable systems in various domains [7]. In their investigation about the sharing of healthcare data, Xia et al. indicated that recording the operation logs of the data (such as hash values and timestamps) using blockchain is an efficient way to achieve "non-repudiation" and "traceability," thus guaranteeing the system's auditability [8]. Moreover, secure logs in Trusted Execution Environments (TEEs) and zero-knowledge proofs (ZKPs) are utilized to check the computational processes without revealing the data.

Fairness indicates that the contributions made by each participant to the total model should be fairly evaluated, acknowledged and rewarded, thus avoiding "free-riding" or ignoring the efforts of those with high contributions.

Several studies have explored contribution evaluation methods, such as using Shapley Value to quantify each participant's marginal contribution to the global model performance [9]. Kang et al. put forward a joint optimization technique which integrates a reputation system with deep reinforcement learning to guarantee the fair allocation of rewards by evaluating the contributions of each member, thus solving the incentive problem in federated learning [10]. The above mentioned approach of measuring the contributions is the main principle for ensuring system fairness.

Collusion resistance means the capability of a system to preserve its fundamental functions and guarantee the reliability and security of the aggregated results, even if several malicious participants cooperate. In their research on secure aggregation, Bonawitz et al. presented a protocol based on secure multi-party computation (MPC) which can tolerate a certain number of malicious nodes and ensures that only the valid updates from the honest participants are correctly combined. This can prevent the malicious nodes from cooperating to alter the global model parameters, thus providing a strong theoretical and technical basis for the anti-collusion performance of federated learning.[11]

The four dimensions defined above constitute the conceptual framework for evaluating trustworthy aggregation. However, transforming these abstract concepts into practical system capabilities requires specific technical support. To achieve correctness, auditability, fairness, and collusion resistance simultaneously, we must rely on the unique properties of blockchain technology and advanced cryptographic primitives. Therefore, the following section will further discuss the technical foundations, specifically analyzing the intrinsic relationship and trade-offs between trustworthiness and privacy protection technologies, which lays the groundwork for the mechanism analysis in Chapter 3.

2.2. The Differences and Connections Between Trustworthiness and Privacy Protection

In federated learning, privacy protection and trustworthy aggregation are usually considered together, but they have different main objectives: privacy protection mainly concentrates on 'confidentiality', which aims to avoid semi-honest nodes from obtaining raw data by means like gradient inversion, thus focusing on the data security; whereas trustworthy aggregation stresses on 'integrity' and 'robustness', trying to prevent malicious nodes from causing the model convergence failure by Byzantine attacks (e.g., gradient tampering), and concentrates on the behavioral and process reliability.

In regard to the technical implementation, it is like a dialectical relationship of "foundation and superstructure". On one hand, the privacy protection is important for trusted aggregation; aggregation of plain text without any privacy assurance may result in data leakage [8]. On the other hand, trusted aggregation can preserve the privacy value; however, if the aggregation process is not trusted, the computed results under privacy protection may lose their validity due to the influence from malicious nodes. Nevertheless, there exists a contradiction between them: in order to guarantee the trustworthiness of the process (e.g., by maintaining the evidence of a blockchain and performing verifiable computation), usually more data visibility or redundancy is required, which is contrary to the 'data minimization' principle in privacy protection. This aspect should be considered in the future evaluation systems.

3. Analysis of Typical Solution Mechanisms

3.1. BlockFL: A Proof-of-Work Consensus Verification Mechanism Based on a Public Blockchain

Kim et al. introduced BlockFL in 2019, which is one of the early approaches to incorporate blockchain technology into the federated learning aggregation procedure [1]. The main idea is to use blockchain rather than the conventional central servers, thereby enabling decentralized verification and aggregation of the model without depending on a credible third party.

In the BlockFL architecture, system entities mainly include devices and miners. The aggregation procedure is as follows:

Firstly, each participant sends the gradient updates to their respective miner nodes after finishing the local model training, who are responsible for collecting and verifying the correctness of the gradient data. Subsequently, the miner which can mine the block by means of a Proof-of-Work (PoW)

consensus algorithm will include the valid gradients in a new block and perform the aggregation process via a smart contract to update the global model. Finally, the devices of all participants obtain the latest block from the blockchain and recompute the global model locally, thereby completing one cycle.

The essential combination mechanism of BlockFL involves summing up the local updates from each node by a weighted average, as depicted in Equation (1):

$$w^{(\ell)} = w^{(\ell-1)} + \sum_{i=1}^{N_D} \frac{N_i}{N_S} (w_i^{(\ell)} - w^{(\ell-1)}) \quad (1)$$

Unlike traditional FL, computing in BlockFL is performed locally and independently by each device, rather than being computed by a central server and then distributed.

Another major contribution of BlockFL lies in the fact that it is the first to establish an end-to-end latency analysis model for this architecture. Kim et al. not only identified the additional latency introduced by the blockchain (communication, computation, and consensus) but also derived a closed-form solution for the optimal block generation rate (λ^*), as shown in Equation (2):

$$\lambda^* \approx 2 \left(T_{bp,\bar{o}}^{(\ell)} \left[1 + \sqrt{1 + 4N_M \left(1 + \frac{T_{wait}}{T_{bp,\bar{o}}^{(\ell)}} \right)} \right] \right)^{-1} \quad (2)$$

This formula quantifies the impact of block propagation delay T_{bp} and the number of miners N_M on system performance, providing a theoretical basis for balancing “consensus efficiency” and “fork risk”.

Research by Kim et al. found that there is a convex-optimized solution for the block generation rate. If the generation rate is too fast, it leads to frequent “forking,” wasting computing power and increasing latency; conversely, if the generation rate is too slow, the utilization of miners’ computing resources decreases.

Different from traditional FL, which relies on a central server, BlockFL’s global model updates are computed locally by each device. Consequently, the failure of a single miner or device will not cause the entire network’s model to collapse, making the system more robust.

Despite theoretically achieving decentralization and auditability, BlockFL has significant shortcomings in terms of privacy and efficiency. On the one hand, this approach does not incorporate differential privacy or homomorphic encryption mechanisms; gradients are recorded on-chain in plaintext, posing a risk that the original data could be inferred; on the other hand, the introduction of the proof-of-work consensus mechanism incurs extremely high computational and time overheads, and uploading the full gradients to the blockchain also results in storage costs that grow linearly with the number of nodes.

In summary, BlockFL is more suitable for specific scenarios with low throughput requirements but extremely high demands for resistance to single points of failure and auditability; it is difficult to apply directly to real-time training tasks on large-scale edge computing devices.

3.2. FedChain: Consensus Protocol Based on Federated Learning and Proof-of-Useful Work

In contrast to the “on-chain aggregation” method of BlockFL, FedChain presents an alternate way of integration: considering the federated learning task as the consensus mechanism of the blockchain. Suggested by Wang et al. from

Tsinghua University in 2023, this method innovates by substituting the “energy-consuming hash calculations” of the traditional proof-of-work with “effective calculations,” that is, competing for the privilege of recording transactions by performing federated learning tasks.[2]

The FedChain system model consists of three main components: task publishers (who publish FL tasks), miners (who carry out the training) and model validators. Firstly, FedChain uses a decentralized Proof-of-Useful Work (PoUW), enabling miners or mining pools to demonstrate their computing power by solving federated learning problems, thus to some degree overcoming the energy consumption problems brought by the conventional Proof-of-Work (PoW). Secondly, FedChain puts forward a mining pool summarization method. The nodes are classified into various mining pools according to the historical communication delay. The nodes in one pool compete by dealing with federated learning tasks, which not only decreases the consensus delay but also improves efficiency by reducing the number of broadcasts over the whole network.

Within the FedChain mining pool, a multi-level privacy safeguard has been set up to avoid the malicious actions of centralized servers or the disclosure of model parameters. It uses a decentralized structure without a central parameter server and a Ring-All-Reduce architecture based on secret sharing. For the concealment of parameters, each participant splits their model parameters and inserts random noises before transmitting and aggregating them in the ring topology. Thus, during the aggregation process, no single node can overhear the raw parameters of other nodes, which results in differential privacy protection during the training.

To solve the problem of non-independent and identically distributed (Non-IID) data, which is often encountered in federated learning, FedChain presents a new model optimization algorithm based on data distribution to enhance the global model’s performance. Different from the conventional FedAvg algorithm that only considers the contribution according to the amount of data, this method innovatively uses the Kullback-Leibler (KL) divergence to measure the difference between the local data distribution and the global example data distribution. During the model combination stage, the system adjusts the aggregation weights of each node in a dynamic way according to the calculated KL divergence, giving larger weights to those nodes whose data distributions are nearer to the global distribution. This weighting scheme, depending on the similarity of the data distributions, can effectively prevent the decrease of the model performance due to data imbalance, guaranteeing that the global model can achieve high convergence accuracy and stability even in complicated non-independent and identically distributed (Non-IID) situations.

During the model verification, FedChain has established a validation approach based on zero-knowledge proofs (zkCNN) to resolve the issue of “how to assess the model accuracy without revealing the model parameters.” With this method, the miners can give proofs to the verifiers showing that their models have attained a certain accuracy on a specific data set without disclosing any detailed parameter information. The system guarantees that only valid and high-quality models can obtain block-creation rewards in order to prevent dishonest nodes from submitting false or low-quality models. At the same time, it satisfies the requirement of privacy protection in a blockchain environment. Transforming federated learning into a “proof-of-useful-

work" strategy, FedChain can make proper use of the computational resources. Although the application of zero-knowledge proofs and circular full reduction makes each computation round more complicated, the mining pool aggregation mechanism can effectively decrease the communication delay, offering an important reference model for building an efficient, secure and environmentally friendly decentralized machine learning network.

3.3. LBFL: Lightweight Committee Consensus Mechanism Based on Proof of Contribution

BlockFL and FedChain have problems like large amount of computational resources and high communication delay. To solve these possible difficulties, Qiao et al. put forward the LBFL (Lightweight Blockchain-based Federated Learning) structure in 2024 [3]. The main purpose of this method is to build a lightweight federated learning system which can coordinate efficiency and equity. Its innovative point is the development of a committee-based consensus mechanism based on Proof of Contribution (PoC), which can decrease the system's computational and communication load while guaranteeing the trustworthiness of model combination. Its consensus mechanism is founded on Proof of Contribution (PoC).

In contrast to the traditional blockchains which depend on the computational power (PoW) or staking (PoS) for record-keeping, LBFL considers the quality of the model updates, an important part of the federated learning, as the basis for reaching consensus. To evaluate the contributions, LBFL proposes a contribution assessment algorithm which evaluates the degree to which each participant node enhances the overall model performance. Instead of fighting for the opportunity to record transactions by meaningless hash collisions, the system assigns consensus weights according to the quality of the gradients uploaded by the nodes, such as the improvements in accuracy and the representativeness of data distribution. Regarding the fair rewards, this method realizes the principle of "more work, more reward". Nodes with greater contributions have a higher possibility of being chosen to record transactions and receive more rewards, thus preventing the "free-riding" behavior and guaranteeing that the model updates included in the aggregation are very effective.

To solve the network congestion due to the participation of full nodes in the consensus, LBFL uses a committee structure: Firstly, the system chooses a group of reputable nodes according to their contributions and reputation scores to constitute a "committee". During the combination process, only the members of the committee are allowed to verify and package the blocks. This method restricts the broadcasting range to within the committee, thus decreasing the redundant communication. With respect to its lightweight feature, the committee mechanism of LBFL is less demanding in latency and energy consumption for reaching consensus compared to the network-wide PoW competition in BlockFL, making it more appropriate for the edge computing environment with limited resources.

With respect to security and stability, LBFL maintains the security of the system without compromising on lightweight performance. In regard to Byzantine fault tolerance, by means of rigorous assessment of contributions and admission procedures for the committee, malicious nodes (like those sending random noise or backdoor models) cannot gain

enough contribution points to join the committee and hence are excluded from the main aggregation process. Similarly to BlockFL, LBFL also makes use of the immutability of the blockchain to keep records of the aggregation process and the behaviors of the nodes, thus ensuring the traceability of the model updates and the openness of the system.

LBFL indicates an important trend in which blockchain facilitates federated learning for "real-world application." It efficiently resolves the issue of high energy consumption in Proof-of-Work (PoW) methods and the long latency in full-node consensus.

4. An Evaluation Framework for Systems Integrating Federated Learning and Blockchain

In order to solve the problem of unsystematic evaluation criteria in the previous studies, this paper establishes a four-dimensional evaluation system - derived from the multi-dimensional concept of "trusted aggregation" and including the technical features of blockchain and federated learning - which covers decentralization, system efficiency, security and stability, and privacy detection. The purpose of this system is to evaluate the trade-offs between trust, performance and resource usage under various integration methods comprehensively.

4.1. Design of Evaluation Dimensions

To carry out a quantitative study on the performance disparities of different methods, this paper has established the following four fundamental assessment aspects:

4.1.1. Degree of Decentralization

This indicator evaluates the dependence of a system on central nodes and the extent of power dispersion.

On the one hand, in the architectural topology, the system should be judged to see if it uses a full-node peer-to-peer (P2P) network, a committee-based method or a centralized parameter server; on the other hand, at the consensus mechanism level, it is required to investigate the distribution of the right to record transactions. For instance, Proof of Work (PoW) and Proof of Stake (PoS) usually have a high degree of decentralization, whereas systems depending on committees or reputation scores, although enhancing efficiency, may involve certain centralization risks.

4.1.2. System Efficiency

The emphasis of this aspect is on the resource consumption, including computation, communication and storage, produced by the solution, which is crucial for its implementation on edge devices.

In connection with the communication load, this factor considers the amount of model parameters or gradients exchanged whether online or off-line - transmitting all the gradients online requires a large amount of bandwidth, whereas a model which combines off-line computation with online data storage is more economical; The computational delay consists of the local training time, the time required to reach an agreement (such as the collision process in PoW) and the verification time (for example, the overhead of producing and verifying zero-knowledge proofs); furthermore, the storage cost is an important basis for assessing this matter. As the blockchain ledgers grow in proportion to the time, the nodes should have enough storage space; therefore, the evaluation should distinguish the storage requirements of full nodes from those of light nodes.

4.1.3. Security and Robustness.

This section is intended to assess the system's ability to resist malicious attacks and abnormal behaviors, including three features: Byzantine fault tolerance, resistance to combination, and data integrity.

With respect to Byzantine fault tolerance, the system should be able to recognize and eliminate those nodes which submit malicious gradients (e.g., backdoor attacks or label-flipping attacks); concerning the resistance against coalition, the system should preserve the accuracy of the summed outcomes even if several malicious nodes collaborate; and for the maintenance of data consistency, the system utilizes the immutability of blockchain to guarantee that the model updates remain unaltered during transmission and storage.

4.1.4. Privacy and Auditability.

The system is now assessed regarding its capability to follow up on user activities while ensuring data privacy, particularly in privacy protection and auditability.

Regarding the privacy protection capability, this factor evaluates the use of techniques like differential privacy, homomorphic encryption or secure multi-party computation to avoid gradient leakage, emphasizing the distinctions between "on-chain plaintext" and "on-chain ciphertext". The traceability aspect demands that the system stores all operation logs on the blockchain (including information about participants, timestamps and model hashes) in order to ensure that responsibility can be identified when there is a conflict or model error.

4.2. Horizontal Comparison of Typical Solutions

Based on the evaluation framework described above, this paper conducts a comparative analysis of three typical approaches—BlockFL, FedChain, and LBFL—with the specific results shown in Table 1.

Table 1. Horizontal Comparison of Typical Blockchain-Enabled Federated Learning Solutions

Evaluation Dimensions	Metrics	BlockFL	FedChain	LBFL	Quantitative Evidence / Complexity (From Literature)
Degree of Decentralization	Consensus Mechanism	PoW	PoUW	PoC	BlockFL: PoW (Hash Puzzle) FedChain: PoUW (FL Task) LBFL: PoC (Committee)
	Network Architecture	Full-node	Mining Pool	Committee	BlockFL: N nodes FedChain: P pools LBFL: C committee members
System Efficiency	Communication Overhead	High	Medium	Low	BlockFL: $O(N \cdot M)$ (Full gradients) FedChain: Pool-based Aggregation LBFL: $O(1)$ (Metadata only)
	Computational Complexity	Medium	High	Low	BlockFL: Hashing (SHA-256) FedChain: ZKP Gen (High Cost) LBFL: Lightweight Verification
Security and Robustness	Resistance to Attacks	Weak	Strong	Moderate	BlockFL: No verification logic FedChain: ZKP Verification LBFL: Reputation/Role-based
	Resistance to Collusion	Weak	Strong	Moderate	BlockFL: Vulnerable to 51% FedChain: ZKP + Secret Sharing LBFL: Role-switching
Privacy and Auditability	Privacy Protection	None	Strong	Moderate	BlockFL: Plaintext on-chain FedChain: Secret Sharing + ZKP LBFL: Off-chain / Metadata
	Auditability	High	High	High	All use Blockchain (Immutable Ledger)

Note: Complexity and overhead evaluations are based on the theoretical analysis and experimental results reported in [1-3]. Specifically, LBFL achieves $O(1)$ storage complexity [3], while FedChain incurs high computational costs due to ZKP generation [2].

As a novel approach, BlockFL possesses complete decentralization and good auditability; however, its efficiency limitations are also quite obvious. The necessity of transmitting all gradients to the blockchain and the application of PoW consensus mechanism hinder its adaptation to large-scale real-time training and the absence of any privacy protection measures is a serious defect.

FedChain has the advantages of privacy and security. It can perform "useful computation" and ensure strong privacy by using zero-knowledge proofs (ZKPs) and ring-based full reduction. However, the considerable amount of computational work, such as the lengthy procedure of producing ZKPs, makes it unsuitable for use on devices with

restricted resources.

LBFL emphasizes lightweight structure and efficiency. It decreases the communication and computation load by using PoC consensus and a committee mechanism, suitable for edge computing systems. Nevertheless, the committee mechanism has reduced some decentralization and its security depends mainly on the reliability of the contribution evaluation algorithm.

In conclusion, the current methods have a "tragedy of the triangles" regarding decentralization, efficiency, security and privacy. BlockFL gives up efficiency and privacy for decentralization; FedChain sacrifices efficiency to achieve security and privacy; and LBFL reduces certain

decentralization to improve efficiency. This offers practical support for the discussions in the following chapters on the main conflicts and hints at revised design principles.

5. Identification and Analysis of the Core Conflict

By using the four-dimensional evaluation framework described in Chapter 4, we can analyze the three representative methods—BlockFL, FedChain and LBFL—and discover that although blockchain technology improves the trustworthiness of federated learning to some degree, the existing hybrid architectures still encounter the problem of conflicting multi-dimensional measures in their effort to achieve "high trustworthiness". This situation is not only due to the limitations in technical implementation, but also results from the inherent contradictions in the system's basic design. In this chapter, the data governance, resource limitations and power relations will be considered together with the technical implementation and incentive systems to clarify the three main conflicts which occur when blockchain is applied to federated learning.

5.1. The Tension between on-chain Immutability and the Principle of Data Minimization

The main features of blockchain are the unchangeability and total openness of data, while privacy-preserving federated learning follows the principle of data minimization, i.e., reducing both the data exposure and the storage amount. These two methods are essentially contradictory in terms of their technical execution.

On one side, from the technical viewpoint, there exists a contradiction between the requirements of evidence preservation and the privacy breaches. In order to ensure "auditability" in the model aggregation process, it is usually necessary to upload the gradient updates or their hash values to the blockchain for evidence preservation. The BlockFL method is an illustration of this strategy, as it sends out the complete gradient vectors to the public blockchain. Nevertheless, it has been found that high-dimensional gradients inherently include confidential information from the original data. If the gradients are uploaded to the blockchain in plain text or with weak encryption, which meets the requirements of auditability and traceability, it will directly violate the "data minimization" principle of privacy protection and make the system susceptible to gradient inversion attacks.

Moreover, there is a contradiction between the verification rewards and the privacy expenditures in the incentive system. In the blockchain network, the miners or validator nodes will perform the verification work fairly only when they receive appropriate rewards. To promote the nodes to confirm the accuracy of the model updates, more information should be provided to them. This leads to a main incentive problem: in order to encourage the nodes to conduct thorough verification, some privacy has to be sacrificed; however, implementing strict privacy techniques (such as those used in FedChain which incorporates zero-knowledge proofs) would increase the computing costs due to the complicated verification process. If the rewards are not sufficiently high, the nodes might not be motivated to participate in the verification.

5.2. The Tension Between Strong Privacy Protection And Limited Resources On Edge Devices

The main application areas of federated learning usually include many resource-limited edge devices like mobile phones and Internet of Things (IoT) devices; however, strong trust and privacy protections generally require high computational and communication expenses. The current approaches usually ignore the physical limitations that edge devices have to deal with and the economic feasibility of the nodes themselves when constructing efficient secure systems. Therefore, it is hard to accomplish the objective of "high trust" under the practical conditions of a "low cost" environment.

A thorough privacy protection and trust verification need complicated cryptographic processes. For instance, the zkCNN verification method in FedChain, which allows model quality assessment without disclosing parameters, includes the creation and verification of proofs containing many polynomial commitments and elliptic curve computations. This may impose a heavy load on the edge nodes with low computational ability and may become a limitation for the system performance.

From an incentive mechanism viewpoint, the nodes are regarded as rational economic entities which are motivated to take part in federated learning because of the equilibrium between advantages and disadvantages. The adoption of advanced privacy and high-trust techniques, like ZKPs, substantially raises the computational expenses of the nodes, including electric bills, depreciation of the equipment and communication costs. If the system provides insufficient incentives, such as token rewards, to offset these extra expenses, the rational nodes will decide to withdraw or decrease their participation, resulting in a "negative selection" situation where "the higher the privacy protection, the less willing they are to join."

5.3. The Risk of Implicit Centralization Under Nominal Decentralization

Although the original intent of using blockchain to enable federated learning was to eliminate single points of failure and achieve full decentralization, in the practical design of efficient architectures, it is often unavoidable to introduce centralization as a necessary compromise. A major challenge in architectural design is how to use technical means to curb excessive concentration of power while maintaining fair incentives.

To address the efficiency issues caused by network-wide broadcasts, solutions such as LBFL employ a "committee consensus" mechanism, in which only a subset of high-reputation nodes are selected to participate in record-keeping. While this design significantly improves system throughput, it effectively concentrates power in the hands of a small number of committee nodes. From a technical implementation perspective, this is an inevitable trade-off—sacrificing partial decentralization for efficiency—resulting in a system architecture characterized by "nominal decentralization but de facto polycentrism."

Under incentive mechanisms based on Proof of Contribution (PoC) or Proof of Work (PoW/PoUW), the system tends to reward nodes that contribute more computing power or higher-quality data. This "more work, more reward" principle of fair incentives can trigger the "Matthew Effect" over the long term: Resource-rich nodes (such as large mining

pools or data giants) will accumulate an ever-increasing number of tokens and reputation, thereby gaining an absolute advantage in committee elections or competition for accounting rights. Ultimately, ordinary nodes will be marginalized, and the system will essentially degenerate into an oligarchic structure controlled by a few supernodes—a development that runs counter to the spirit of equal collaboration advocated by federated learning.

6. Principles of Lightweight Collaborative Design

In order to deal with the three main conflicts mentioned before, this chapter puts forward the lightweight cooperative design concept of “on-chain record keeping and off-chain computation.” It is intended to attain an optimum equilibrium among system efficiency, privacy preservation and decentralization – while maintaining credibility – by means of hierarchical structure and functional separation.

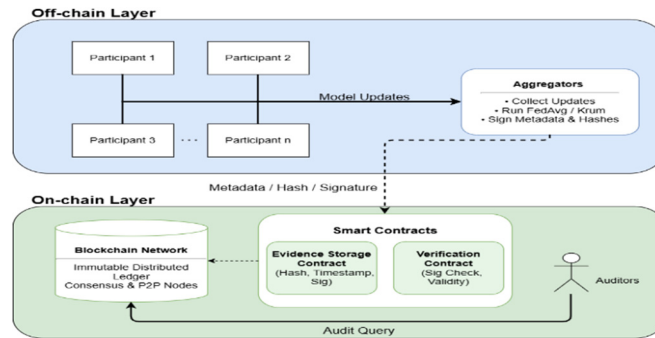


Figure 1. System Architecture of On-chain Evidence Storage and Off-chain Computation

6.1. On-chain evidence storage: Building a Lightweight Trust Anchor

The main advantage of blockchain is its immutability and traceability, not its storage and calculation of large volumes of data. Hence, it should be regarded as the “trust center” of the system, being utilized only for recording important metadata and operation hashes rather than raw gradients or model parameters.

The design rule is that only metadata, including the identities of participants, timestamps, hash values of model updates and contribution scores, should be uploaded to the chain. This method can fulfill the auditability needs, guaranteeing the traceability of actions and the determination of responsibility, and at the same time comply with the data minimization principle to avoid privacy problems caused by storing gradients in clear text on the chain.

The consensus method needs to be improved. The high-power-consumption and low-speed PoW consensus should be substituted by other efficient consensus schemes such as the committee mechanism in LBFL or PoUW in FedChain. Considering the reputation and contribution to choose the validator nodes, these schemes can significantly reduce the communication and computational load while ensuring security and stability, therefore decreasing the chance of “hidden centralization.”

6.2. Off-chain computation: Enabling Highly Efficient Privacy-Preserving Aggregation

To guarantee security, complicated model training, privacy-preserving computation and reliable aggregation should be carried out completely off-line.

On one side, privacy-preserving computations need to be processed off-line, that is, protecting gradients by means of techniques like differential privacy, homomorphic encryption or secure multi-party computation. By executing these high-computational operations off-line, we can prevent the use of resource-consuming cryptographic methods, such as zero-knowledge proofs, on the edge devices with limited resources, thus solving the contradiction between strong privacy and

resource limitations.

Furthermore, off-line aggregation should be carried out. This includes executing robust aggregation algorithms like Krum and mean pruning off-line to guarantee Byzantine fault tolerance for the uploaded model updates and to eliminate malicious or inferior updates. Only the verified and improved aggregation results, or their hash values, should be submitted on-line for the final verification and data storage.

7. Summary and Outlook

7.1. Core Contributions

This paper systematically investigates the trustworthy aggregation mechanisms for blockchain-based federated learning. The core contributions of this study are summarized as follows:

Establishment of an Evaluation Framework. We proposed a four-dimensional evaluation system covering decentralization, system efficiency, security/robustness, and privacy/auditability to assess aggregation mechanisms comprehensively.

Comparative Analysis. We conducted an in-depth analysis of three typical solutions (BlockFL, FedChain, and LBFL), identifying their technical trade-offs and performance bottlenecks.

Identification of Core Conflicts. We revealed three inherent conflicts in current architectures: the tension between on-chain immutability and data minimization, the contradiction between strong privacy and limited edge resources, and the risk of implicit centralization.

Proposal of Design Principles. We introduced a lightweight cooperative design concept, “on-chain evidence storage and off-chain computation,” to resolve these conflicts and provide a theoretical basis for trustworthy FL in sensitive scenarios.

7.2. Applicability of the Proposed Design Principles

The proposed principle of “on-chain evidence storage and off-chain computation” is particularly suitable for consortium

blockchains or private blockchains where node identities are partially trusted, but data privacy is paramount (e.g., cross-silo federated learning in healthcare or finance). In these scenarios, the high auditability of the blockchain ensures compliance, while off-chain computation alleviates the burden on resource-constrained devices. However, this design may face challenges in fully permissionless public blockchains with massive, anonymous nodes, where the coordination cost for off-chain secure aggregation could become a bottleneck. Furthermore, for scenarios requiring real-time, millisecond-level inference (e.g., autonomous driving), the latency introduced by off-chain cryptographic protocols (like MPC or HE) needs to be carefully optimized.

7.3. Limitations and Future Work

Due to the limitation of space and experimental environment, this paper mainly conducts qualitative analysis and theoretical derivation of the mechanism.

First, the proposed "on-chain evidence storage and off-chain computation" framework is currently validated through theoretical derivation and qualitative analysis, lacking a fully functional prototype deployed in large-scale real-world edge networks.

Second, the comparison of the three typical solutions relies on theoretical metrics, which highlights architectural differences but lacks quantitative performance benchmarks (e.g., exact latency in milliseconds, specific energy consumption) under identical experimental settings.

To address these limitations, future research should focus on three main directions. Priority should be given to building a federated learning prototype system based on a consortium blockchain to verify the feasibility of the proposed architecture. Subsequently, comparative experiments need to be conducted to quantitatively measure performance gains in terms of throughput, latency, and resource consumption. Finally, further investigation into lightweight homomorphic encryption or zero-knowledge proof algorithms, specifically optimized for edge devices, is recommended to further reduce computational complexity while maintaining security.

References

- [1] Kim, H., Park, J., Bennis, M., & Kim, S. L. (2019). Blockchain-based on-device federated learning. *IEEE Communications Letters*, 24(6), 1279–1283. <https://doi.org/10.1109/LCOMM.2019.2913140>.
- [2] Wang, P. (2023). Fedchain: An efficient and secure consensus protocol based on proof of useful federated learning for blockchain. *arXiv preprint arXiv:2308.15095*. <https://doi.org/10.48550/arXiv.2308.15095>.
- [3] Qiao, S., Jiang, Y., Han, N., Hua, W., Lin, Y., Min, S., & Wu, X. (2024). LBFL: A lightweight blockchain-based federated learning framework with proof-of-contribution committee consensus. *IEEE Transactions on Big Data*, 11(4), 1745–1759. <https://doi.org/10.1109/TBDATA.2024.3372164>.
- [4] Blanchard, P., El Mhamdi, E. M., Guerraoui, R., & Stainer, J. (2017). Machine learning with adversaries: Byzantine tolerant gradient descent. *Advances in Neural Information Processing Systems*, 30. <https://doi.org/10.5555/3294771.3294854>.
- [5] Yin, D., Chen, Y., Kannan, R., & Bartlett, P. (2018). Byzantine-robust distributed learning: Towards optimal statistical rates. In *International Conference on Machine Learning* (pp. 5650–5659). PMLR.
- [6] Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., & Shmatikov, V. (2020). How to backdoor federated learning. In *International Conference on Artificial Intelligence and Statistics* (pp. 2938–2948). PMLR. <https://doi.org/10.5555/3403017.3404067>.
- [7] Atzei, N., Bartoletti, M., & Cimoli, T. (2017). A survey of attacks on ethereum smart contracts (SoK). In *International Conference on Principles of Security and Trust* (pp. 164–186). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-54455-6_8.
- [8] Xia, Q., Sifah, E. B., Asamoah, K. O., Gao, J., Xia, H., & Du, X. (2017). MedShare: Trust-less medical data sharing among cloud service providers via blockchain. *IEEE Access*, 5, 14757–14767. <https://doi.org/10.1109/ACCESS.2017.2730843>.
- [9] Jia, R., Dao, D., Wang, B., Hubis, F. A., Gurel, N. M., Li, B., Zhang, C., & Song, D. (2019). Efficient task-specific data valuation for nearest neighbor algorithms. *arXiv preprint arXiv:1908.08619*. <https://doi.org/10.48550/arXiv.1908.08619>.
- [10] Kang, J., Xiong, Z., Niyato, D., Xie, S., & Zhang, J. (2019). Incentive mechanism for reliable federated learning: A joint optimization approach to combining reputation and contract theory. *IEEE Internet of Things Journal*, 6(6), 10700–10714. <https://doi.org/10.1109/JIOT.2019.2935249>.
- [11] Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., & Seth, K. (2017). Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (pp. 1175–1191). <https://doi.org/10.1145/3133956.3133982>.