

Research on AI Trusted Traceability Technology for Data Element Circulation

Guangdi Zhang, Yi Yang, Na Ni, Yue Tan, Peng Wang

Beijing Key Laboratory of Future City Artificial Intelligence Measurement Technology, the NO.203 Institute of the Second Academy of China Aerospace Science & Industry Corp, Beijing, China

Abstract: This article studies AI trustworthy traceability technology for data element circulation. The system has sorted out key technologies such as blockchain certification, digital watermarking, federated learning (FL) auditing, and trusted execution environment (TEE), and compared and analyzed them from the dimensions of applicability, security, and scalability; Furthermore, a multi-layer fusion trustworthy traceability framework is proposed, which integrates mechanisms such as data identification, model watermarking, on chain authentication, and smart contracts, and constructs a four layer collaborative architecture covering the data layer, model layer, circulation layer, and application layer, achieving trustworthy traceability of data throughout its entire lifecycle from production, circulation to application. This framework establishes a digital identity for data through a dual identification of "data fingerprint+digital watermark", utilizes blockchain and smart contracts to achieve transparent and controllable circulation processes, relies on TEE to ensure trustworthy execution in the inference stage, and supports responsibility tracing through blood relationship graphs and audit logs. Finally, the application prospects of the framework in scenarios such as data transactions, model services, industry data sharing, and AIGC (AI Generated Content) governance were discussed, and open challenges such as performance, interoperability, and privacy traceability balance were pointed out, providing technical references for the trustworthy circulation of data elements.

Keywords: Data element circulation; Trustworthy traceability; Blockchain certification; Digital watermark; Federated Learning; Trusted Execution Environment.

1. Introduction

The circulation of data elements faces four major practical difficulties: difficulties in property rights confirmation, pricing, circulation, and traceability. Especially in the context of deep embedding of AI technology into the entire lifecycle of data, the collection, annotation, cleaning, training, inference, and application of data are highly coupled. The "black box" characteristics of AI models themselves make problems such as opaque data sources, untraceable usage paths, and difficult to define responsibility increasingly prominent.

Blockchain and distributed ledger technology (DLT) are widely used in data storage and transaction record management because of its non-tampering and decentralized characteristics [1]; Digital watermarking and data fingerprinting technology can embed unique identification for data sets and AI models to track illegal copying and embezzlement [2]; Privacy-preserving computing technologies such as federated learning (FL) and secure multi-party computation (MPC) ensure that data is "usable but not visible" while providing a certain degree of operational auditability [3]. However, most of the above technologies focus on single dimensional traceability requirements and lack systematic integration for the composite scenario of "AI+data circulation". Although blockchain solutions can ensure that records cannot be tampered with, they are difficult to cope with the efficient on chain and on chain storage pressure of massive fine-grained data; Although model watermarking technology can achieve model source tracing, it lacks full traceability coverage at the data level; Although privacy computing solutions can protect data privacy, they are still weak in terms of behavior tracking and responsibility determination after circulation.

Based on the above background and challenges, this article focuses on the core issue of AI trustworthy traceability technology for data element circulation. This article systematically outlines the key technology spectrum that supports AI trustworthy traceability, covering core technologies such as blockchain certification, digital watermarking, FL auditing, and trusted execution environment (TEE), and conducts comparative analysis from the dimensions of applicability, security, and scalability; Subsequently, a multi-layer fusion trustworthy traceability framework for data element circulation scenarios was proposed, which integrates mechanisms such as data identification, model watermarking, on chain authentication, and smart contracts, covering the entire lifecycle of data from production to circulation and application; Finally, the application prospects of the framework in practical data trading and AI service markets were discussed, and the open issues faced by existing technologies in terms of performance, interoperability, privacy traceability balance, and new challenges of AIGC (AI Generated Content) were pointed out, providing direction for future research.

2. Key technology analysis

2.1. Blockchain certification technology

Blockchain certification utilizes the tamper proof nature and timestamp mechanism of distributed ledgers to record key events such as data generation, circulation, and transactions in a structured transaction format on the blockchain [4]. The specific blockchain workflow is shown in Figure 1. Its core advantage lies in its ability to solidify records through consensus among multiple parties. Once the data event is confirmed by the consensus mechanism, no single participant can tamper with the historical records afterwards, providing

verifiable objective basis for defining data ownership and tracing the flow trajectory. In the scenario of data element circulation, this technology is suitable for links that require joint confirmation from multiple parties and have high requirements for record consistency. However, this technology faces two limitations. Firstly, there is a contradiction between the transparency of on chain data and privacy protection requirements. Secondly, throughput bottlenecks restrict its large-scale application in high-frequency transaction scenarios.

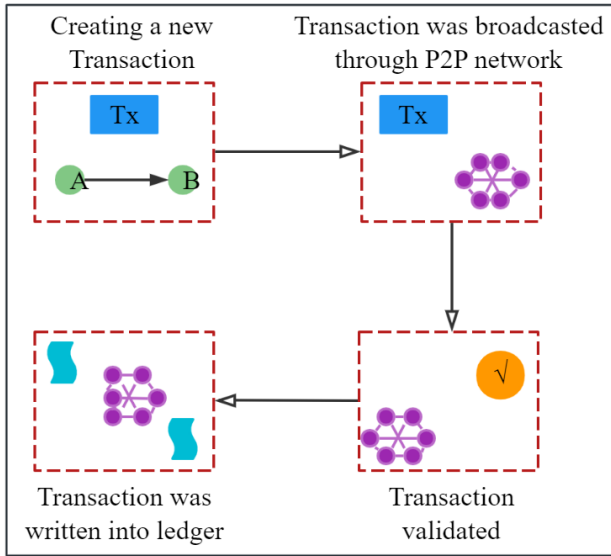


Figure 1. Blockchain workflow

2.2. Digital watermarking technology

Digital watermarking achieves source tagging and version tracking of content by embedding imperceptible identification information in the data carrier. In the AI trustworthy traceability system, the value of digital watermarking is reflected in two aspects: for training data, the watermark can carry source identification during the data flow process, supporting downstream applications to verify the source of the data; For model assets, model watermarking

can mark the ownership and authorization status of the model, providing evidence for model theft and infringement [5]. The technical difficulty lies in the trade-off between robustness and invisibility. Watermarks need to be able to be extracted after common processing such as compression and format conversion, while not significantly reducing the availability of the original data. The embedding and extraction of digital watermarks have relatively low computational costs and have good potential for large-scale deployment. However, their security depends on the design of specific embedding algorithms, and there is a risk of being removed or compromised by watermark removal or overlay attacks.

2.3. FL audit technology

FL audit addresses the credibility issues in multi-party collaborative training scenarios by recording, verifying, and analyzing the local computing processes and gradient update interactions of each participant, ensuring the integrity of the FL process and the compliance of the participants [6]. Its core mechanism includes signing or hashing gradient updates to prevent malicious tampering, identifying malicious behaviors such as poisoning attacks or member inference attacks through anomaly detection, and reconstructing the training process in case of disputes using audit logs. This technology is suitable for privacy protection scenarios where data does not leave the domain and multi-party joint modeling is required. It is an important link connecting privacy computing and trustworthy traceability. The audit mechanism itself introduces additional computation and communication overhead, and the analysis of gradient information requires a balance between privacy protection and auditability.

2.4. Trusted execution environment

TEE builds a trusted execution area on a general-purpose computing platform through hardware level isolation mechanisms (as shown in Figure 2), isolating sensitive data processing and critical computing logic from execution within the trusted area, ensuring that even if the operating system or hypervisor is breached, data and code within the trusted area are still protected from theft or tampering.

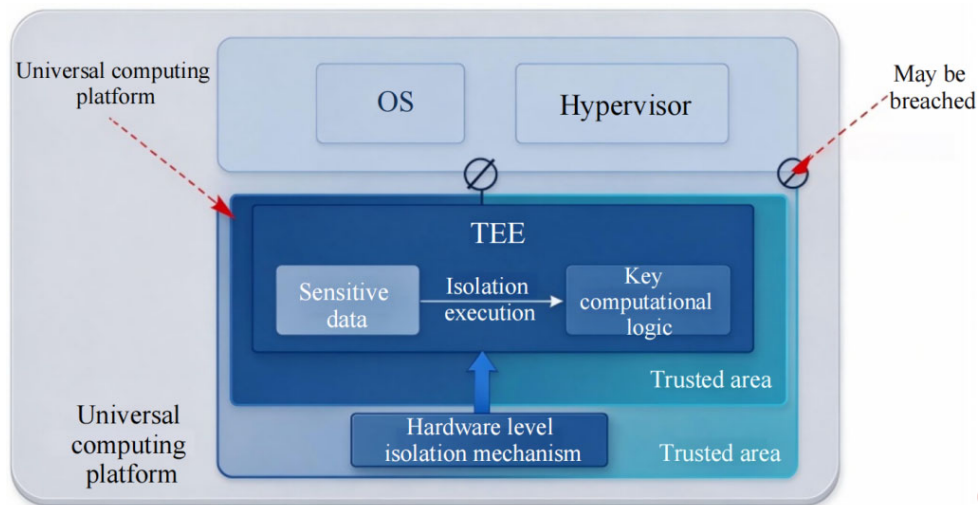


Figure 2. Principles of TEE technology

In the flow of data elements, TEE provides infrastructure level guarantees for the trusted processing of sensitive data, supporting scenarios such as trusted computing in data transactions and privacy protection in model training [7]. The security of this technology is built on the hardware trust root,

with high security strength, but it also faces security challenges such as side channel attacks, and the computing resources within the trusted area are limited. Mechanisms such as context switching and remote proof will bring certain performance overhead.

2.5. Technical comparison

Overall, the four technologies mentioned above each have their own emphasis in the AI trustworthy traceability system (as shown in Table 1). Blockchain forensics excels in record solidification under multi-party consensus, digital

watermarking focuses on source identification at the content level, FL auditing ensures the trustworthiness of collaborative training processes, and TEE provides infrastructure level isolation for sensitive computing. A single technology is difficult to cover the trustworthy requirements of the entire lifecycle of data from production, circulation to application.

Table 1. Comparison of key technologies

Technology	Core mechanism	Applicable scenarios	Safety	Scalability	Main limitations
Blockchain evidence storage	Distributed consensus+timestamp	Multi party consensus recording and transaction verification	Strong immutability, relying on consensus mechanism	Limited throughput and obvious scalability bottleneck	The contradiction between privacy and transparency
Digital Watermarking	Carrier embedding identification	Content verification and model asset tracking	Dependent on algorithm design, there is a risk of removal	Low cost, easy to deploy on a large scale	Balance between robustness and invisibility
FL Audit	Process recording+behavior verification	Multi party collaborative training audit	Detectable malicious behavior	Audit expenses increase with the participation of parties	Balance between privacy and auditability
TEE	Hardware level isolated execution	Sensitive data processing, trusted computing	Hardware trust root, high security strength	Resource constraints and high performance overhead	Side channel attack risk

3. Multi layer fusion trustworthy traceability framework for data element circulation scenarios

In response to the core pain points of difficulty in confirming the source, tracking the flow, and defining responsibilities in the process of data element circulation, this article proposes a multi-layer fusion trustworthy traceability framework (as shown in Figure 3). This framework follows the design principles of traceable identification, auditable process, controllable behavior, and accountable responsibility. It embeds traceability capabilities into various stages of the data lifecycle, and constructs a four layer collaborative architecture covering the data layer, model layer, circulation layer, and application layer, achieving full chain trusted traceability from raw data production to the final application of AI models.

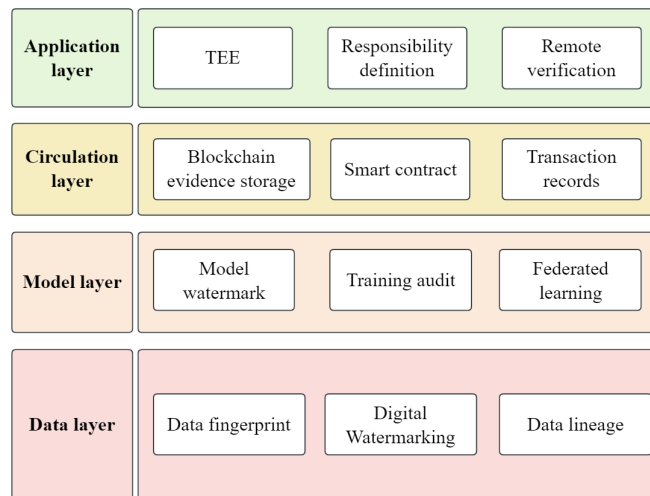


Figure 3. Multi layer fusion trustworthy traceability framework

The operation of the framework follows a closed-loop process of registration, authorization, circulation, and auditing. The data provider registers the data fingerprint, watermark information, and basic metadata on the blockchain to generate a digital identity credential; The data demander

applies for data usage authorization through a smart contract, and the authorization terms are automatically uploaded to the chain; The data flows within the authorized scope, generating new transaction records each time and storing them on the chain. During the model training process, a model watermark is embedded; Regulatory agencies or data providers can initiate audit requests at any time to fully restore the flow path and usage of data through on chain records, kinship maps, and watermark verification results.

3.1. Data layer

The data layer is the foundation of the entire traceability system, mainly responsible for assigning a unique and tamper proof digital identity to each piece of data entering the circulation process. Adopting a dual identification strategy of "data fingerprint+digital watermark". The data fingerprint generates a unique digest of the original data through hash algorithm, and any slight modification will cause the fingerprint to change, ensuring the integrity and verifiability of the data. Digital watermarking embeds identification information in an imperceptible way into the data content itself, and even if the data is copied, cropped, or formatted, the watermark can still be extracted for traceability. Establish a structured data lineage graph to record information such as the source system, collection time, processing history, and version evolution of the data. The blood relationship graph is stored in a directed acyclic graph (DAG) structure, which clearly depicts the derivation and dependency relationships between data, providing complete contextual information for upper level traceability.

3.2. Model layer

The model layer focuses on the traceability of AI models during data training and usage, addressing key questions such as what data the model uses and whether the model output is trustworthy. During the model training phase, specific watermark information is embedded into the model parameters. Model watermarking has robustness, and even after fine-tuning, pruning, or distillation, it can still be detected and verified, thereby achieving traceability of the model's source and training data compliance. Introducing FL

audit mechanism, in distributed training scenarios, all participants only exchange model parameters or gradient updates without sharing raw data. The audit module records the identities of participants, parameter contributions, and aggregation processes for each round of training, ensuring that the training process is verifiable and auditable, and preventing malicious participants from injecting poisoned data.

3.3. Circulation layer

The circulation layer is the core link of market-oriented allocation of data elements, ensuring transparency and controllability of the circulation process through blockchain and smart contract technology. Upload key information such as data fingerprints, transaction hashes, and authorization credentials to the blockchain for authentication. The immutability and distributed consensus mechanism of blockchain ensure that once a record is recorded, it cannot be unilaterally tampered with, providing trusted third-party evidence support for data transactions. Encode data usage rules into smart contracts to automatically execute and validate data usage behavior. When data is called, the smart contract automatically checks the caller's authorization credentials, and unauthorized access will be denied and recorded, achieving automatic compliance control of "code as law".

3.4. Application layer

The application layer is oriented towards the final data usage scenario, ensuring that the output of AI models in the inference and application stages is trustworthy and responsibilities can be defined. During the model inference phase, use TEE to construct an isolated and secure execution environment. The data is decrypted and processed within TEE, and cannot be accessed or tampered with externally, ensuring the credibility of the inference process. Meanwhile, TEE can generate remote proofs to prove to data providers that the model is indeed running in a protected environment and has not been tampered with. When data abuse or model output errors occur, the problem can be quickly identified through on chain evidence records, model watermark information, and watermark extraction results, clarifying the responsibility boundaries of data providers, model developers, platform operators, and other parties.

4. Application prospects

The multi-layer fusion trustworthy traceability framework proposed in this article is based on the practical needs of traceable sources, destinations, and responsibilities in the circulation of data elements. It embeds mechanisms such as data identification, model watermarking, on chain authentication, and smart contracts into the entire lifecycle of data. With the accelerated implementation of data infrastructure systems and the large-scale deployment of AI applications, this framework has broad application prospects in areas such as data trading, model services, industry data sharing, trusted data space construction, regulatory compliance, and AIGC governance.

Data transactions have long been plagued by difficulties in property rights confirmation, traceability, and evidence collection, making it difficult for both parties to form stable expectations regarding the source and scope of use of the subject matter. This framework establishes a digital identity based on "data fingerprint+digital watermark" as the

transaction subject, records processing and derivative relationships through a data blood relationship graph, and implements authorization, performance, and call tracing throughout the entire process through on chain authentication and smart contracts. Once unauthorized use or data abuse occurs, the flow path can be restored based on on chain records and watermark verification results, forming an evidence chain, providing verifiable objective basis for transaction matching, pricing evaluation, and even dispute arbitration. Against the backdrop of the gradual rise of model as a service and model trading, model assets face risks such as theft, unauthorized secondary distribution, and infringement fine-tuning. This framework embeds robust model watermarks to verify ownership and authorization status even after fine-tuning, pruning, or distillation of the model, providing technical means for model ownership confirmation, authorization management, and infringement evidence collection; At the same time, with the help of the model layer audit mechanism, the question of which data the model used can be answered, alleviating the compliance risk caused by unclear sources of training data for large models.

This type of industry has high data sensitivity and strict regulatory requirements, making it difficult to directly aggregate data. The TEE and FL audit mechanisms in the framework can support cross institutional joint risk control and joint modeling without the original data being out of domain or visible, and ensure the trustworthiness of participant identities, verifiability of processes, and accountability of behaviors through remote proof and audit logs. In recent years, the engineering practice of trustworthy data space in industries such as medical insurance and energy, which revolves around cross institutional data sharing, has preliminarily verified the feasibility of such technological paths. The trusted data space emphasizes the ability to access authentication, contract performance certification, and full traceability of data resource development and utilization. This framework can serve as a trusted control technology component, providing support for identity mutual recognition, authorization circulation, and responsibility definition among multiple entities in the space. Regulatory agencies or data providers can initiate audit requests at any time based on the framework, quickly locate problematic links through on chain records, kinship maps, and watermark verification results, clarify the responsibility boundaries of data providers, model developers, and platform operators, shift post event accountability to full process auditability, and provide instrumental support for data element security governance and law enforcement evidence collection. In response to the identification and copyright issues brought about by the large-scale dissemination of generative content, the watermark identification and on chain authentication mechanism in the framework can be extended to source tagging, propagation path tracking, and training data compliance review of generated content, providing technical basis for content identification management and copyright dispute resolution on content platforms.

It should be pointed out that the realization of the above application prospects is still subject to several practical conditions. The performance and storage pressure brought by massive fine-grained data on chain, interoperability across technology stacks and platforms, balance between privacy protection and traceability, lack of relevant standard specifications, and the ongoing game between watermark robustness and adversarial attacks in AIGC scenarios. These

are also the key directions that need to be broken through in future research in this field. Overall, the value of this framework lies not in replacing a single technology, but in organizing dispersed trusted capabilities into a closed loop covering the entire data production, circulation, and application chain. It has the potential to move from a theoretical framework to engineering implementation and gradually expand application boundaries with the development of the data element market.

5. Conclusion

Blockchain is good at multi-party consensus recording but is constrained by throughput bottlenecks. Digital watermarking focuses on source identification but has robustness and security risks. FL auditing ensures the trustworthiness of collaborative training but faces the challenge of balancing privacy and auditability. TEE provides infrastructure level isolation but has limited resources and the risk of side channel attacks. On this basis, this article proposes a multi-layer fusion trustworthy traceability framework for data element circulation scenarios, following the design principles of traceable identification, auditable process, controllable behavior, and accountable responsibility. It constructs a four layer collaborative architecture covering the data layer, model layer, circulation layer, and application layer. Through the dual identification of "data fingerprint+digital watermark", model watermark and FL audit, on chain certificate and smart contract, TEE remote proof and other mechanisms, it realizes the full chain trustworthy traceability from raw data production to the final application of AI models. This framework has broad prospects in data trading, model as a service, trusted data space construction, regulatory compliance, and AIGC governance. In the future, continuous efforts will be needed to address performance overhead, cross platform interoperability, privacy and traceability balance, and AIGC watermark robustness.

References

- [1] Shah, A. M., & Ganesh, B. U. (2026). Delays to distributed ledgers: Examining the rationale for leveraging blockchain technology in construction sector. *Global Knowledge, Memory and Communication*, 75(7-8), 3834-3853. <https://doi.org/10.1108/GKMC-03-2024-0177>
- [2] Gull, S., Hussan, M., & Parah, A. S. (2026). Automatic segmentation based watermarking technique for medical images in e-healthcare. *Health and Technology*, 16(2), 1-14. <https://doi.org/10.1007/S12553-025-01045-8>
- [3] Wang, J., Liu, Y., & Yang, W. (2026). Secure multi-party computation-based distributed state estimation against hybrid attacks. *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*, 240(6), 1087-1099. <https://doi.org/10.1177/09596518261422591>
- [4] Ramachandran, A., Ramadevi, P., Alkhayyat, A., & Kumar, S. (2023). Blockchain and data integrity authentication technique for secure cloud environment. *Intelligent Automation & Soft Computing*, 36(2), 2055-2070. <https://doi.org/10.32604/IASC.2023.032942>
- [5] Naima, S., Boukhamla, E. Z. A., Narima, Z., & Ahmed, M. (2024). Secure and imperceptible frequency-based watermarking for medical images. *Circuits, Systems, and Signal Processing*, 44(1), 1-22. <https://doi.org/10.1007/S00034-024-02814-Y>
- [6] Zeng, H., Yang, A., Weng, J., & Li, M. (2026). Publicly auditable federated learning with privacy and Byzantine robustness. *IEEE Transactions on Dependable and Secure Computing*, 23(1), 577-589. <https://doi.org/10.1109/TDSC.2025.3609353>
- [7] Chao, Z., Weidong, Z., An, X. W., & Hong, L. (2026). An improved cloud data auditing scheme with enhanced security based on trusted execution environment-based multi-replica data audit. *Journal of High Speed Networks*, 32(2), 147-162. <https://doi.org/10.1177/09266801261419605>