Skillsets on the Chain: A Blockchain-based Trustworthy Agentic AI Networking Framework

Yayu Gao*, Yong Xiao*†‡, Bo Liu*, Xubo Li*, Aoyu Hu*, Wenhao Li*, Yingyu Li‡, Guangming Shi†‡§, and Ping Zhang¶
*School of Elect. Inform. & Commun., Huazhong Univ. of Science & Technology, Wuhan, China
† Peng Cheng Laboratory, Shenzhen, China
‡ Pazhou Laboratory (Huangpu), Guangzhou, China
§ School of Artificial Intelligence, Xidian University, Xi'an, China
¶ State Key Lab. of Networking & Switching Tech., Beijing Univ. of Posts & Telecom., Beijing, China

***Abstract*—Agentic AI networking (AgentNet) has attracted significant interest due to its promising potential to move traditional AI-based networking solutions beyond closed-loop and passive learning to proactive interaction and goal-driven action, offering a path to self-learning and generally intelligent networking systems. Despite its promise, ensuring the security and trustworthiness of such systems presents significant challenges, particularly concerning identity management, agent capability verification, and data integrity during collaborative learning. To address these issues, this paper proposes *TrustAgentNet*, a novel consortium blockchain-based framework for unified and trusted agent identification, traceable skillset and tag descriptions, and secure on-chain collaborative learning in AgentNet. In *TrustAgentNet*, a chain of skillset (CoS) is introduced, consisting of a skillset chain to distributedly store all the verified skillsets and associated tags, and a dedicated training chain for each distinct skillset can be jointly constructed and maintained by the authorized agents using a collaborative learning-based approach. Theoretical analysis suggests that there exists a three-way trade-off among the security level, skillset performance, and resource cost. This tradeoff is also empirically validated by the experimental results obtained from a hardware prototype implemented based on a Hyperledger Fabric-based consortium blockchain. To verify the practical performance of *TrustAgentNet*, we consider a real-world scenario of multi-agent collaborative learning under malicious attack. Experimental results suggest that TrustAgentNet can effectively guarantee the security of skillset training and enable rapid response and recovery from potential attacks within seconds.**

***Index Terms*—Agentic AI networking, security, authentication.**

I. Introduction

Agentic AI networking (AgentNet) is a novel AI-native networking ecosystem in which autonomous AI agents can collaborate, reason, and plan to solve complex, multi-step problems with minimal human intervention. It has the potential to overcome the limitations of existing AI-based solutions, positioning it as one of the possible architectures for next-generation networking systems, especially 6G and beyond [1].

Despite its promising potential, this fundamental shift in networking paradigm introduces profound challenges, particularly concerning security and trustworthiness issues. More specifically, the inherent autonomy, tool-use capabilities, and multi-agent interaction and communication interfaces within networking systems create novel attack surfaces susceptible to threats such as prompt injection, memory poisoning, and identity spoofing. If these vulnerabilities cannot be comprehensively addressed, they could lead to systemic failures, ultimately jeopardizing the reliability and integrity of the overall AgentNet systems. Despite these critical risks, the majority of existing research in the literature remains heavily focused on enhancing the individual or collaborative performance of agents, often overlooking the fundamental security and trustworthiness issues that are essential for their safe and robust deployment [2]–[4].

In particular, the increasing adaptation of various AI agents and agentic AI services and applications poses the following novel challenges to communication networking systems:

(1) Identity management and validation of diverse agents: A central challenge in the deployment of AgentNet systems is the management and validation of heterogeneous agent identities. As mentioned earlier, AgentNet systems are comprised of a multitude of diverse agentic entities, spanning embodied agents such as robots and drones, and virtual agents including foundation model-based agents like large language models (LLMs) and large vision models (LVMs)-based agents. The intrinsic disparity among these agents—in their physical form, operational protocols, and underlying architectures—necessitates the development of a unified identity management and validation framework. Such a framework would provide a cohesive mechanism to provision, manage, and correlate the identities of different agents with different skillsets, thereby facilitating secure and seamless interaction and collaboration within the networking system.

(2) Trustworthiness of agents' claimed skillsets: To ensure efficient multi-agent planning and task allocation, each agent's claimed skillsets must match its real capabilities. More specifically, for the agent controller or a collaborative set of agents to make a joint decision for a certain task, it must be able to verify that a prospective agent possesses the required skillsets. As agents operate within decentralized and distributed environments, it becomes imperative to establish a robust system for the verifiable and traceable documentation of their real capabilities when performing different tasks in various practical environments. The implementation of such a framework is essential for fostering trust and ensuring the reliable operation of multi-agent interaction and cooperation

within the networks.

(3) Data privacy guarantee in life-long collaborative learning: The challenge of supporting the continuous evolution of agents' skillsets while ensuring data privacy is a central concern. One of the key differences of AgentNet, compared to the traditional static model-based learning framework, lies in its capacity for supporting continuous, life-long learning, adaptation, and collaboration among diverse agents, each can dynamically refine its skillsets by exchanging and sharing new knowledge and information with others. Therefore, it is of critical importance to develop a novel collaborative learning paradigm that can facilitate the collective evolution of skillsets of different agents while strictly preserving the privacy of each agent's local datasets.

To address the above challenges, in this paper, we introduce a novel AgentNet architecture, called TrustAgentNet, that supports agent identity verification and provides skillset trustworthiness and data privacy guarantee. Our main contributions are summarized as follows:

(1) Novel architecture with identity verification, trustworthiness and data privacy guarantee: We propose *TrustAgentNet*, a novel consortium-blockchain-based management and networking framework that can address the aforementioned trustworthiness and security challenges of AgentNet. More specifically, TrustAgentNet is built based on a chain of skillset (CoS) framework that can support a unified interface for agent identity authentication and management for various types of agents. Second, the real performance and capability of agents can be recorded and tracked by the CoS, which will be compared with the skillsets reported by the agent tags and the agents' real performance in practical scenarios will be updated, and with its tag information to facilitate their future task planning and collaboration with each other. Finally, the continuous adaptation and improvement of agents' skillsets via distributed learning on the CoS among certificated agents can guarantee the data privacy of agents.

(2) Theoretical analysis of three-way tradeoff among security level, agent performance, and resource cost: We provide a theoretical analysis that captures the three-way tradeoff among the security level, the agents' model performance, as well as the communication and computational resource costs of the proposed TrustAgentNet.

(3) Hardware prototype and extensive experiments: We develop a hardware prototype to evaluate the performance of TrustAgentNet. Experimental results validate that the level of security can be improved at the cost of traffic volume, computational demands and running time for skillset training under a given performance target. We also demonstrate that the proposed TrustAgentNet can effectively address potential security risks by recognizing users' intent and selecting the appropriate learning mode accordingly.

II. System Model and Problem Formulation

A. System Model

We consider a general AgentNet system consisting of a set $\mathcal{K}$ of agents, deployed across various environments $\mathcal{E}$. In each environment e, K_e agents are deployed and can collaborate to solve a finite set of tasks $\mathcal{T}_e$. We assume that each task can be further divided into multiple sub-tasks, each can be solved by training and maintaining a specific AI model that can generate the intended output based on the input. We use the term "skillset" to refer to the function and input-to-output relation of a specific AI model. Let $\mathcal{S}$ be the set of all the skillsets that are accessible to the agents. We further assume that each agent k in environment e can access to a subset of skillsets $\mathcal{S}_{k,e}=\{s_1,\ldots,s_i\}$, $s_i\in\mathcal{S}$. Each agent k can also develop and update these local skillsets using an exclusive local dataset $\mathcal{D}_{k,e}$, sampled from an unknown data distribution $P_{k,e}$. Each agent will not expose its local data samples to others due to privacy concerns. Notably, agents deployed in different environments may be tailored for the same type of subtask and therefore possess one or multiple identical skillsets. As the local datasets and distributions are generally distinct among the agents, in this paper, we consider a collaborative learning framework in which multiple agents sharing the same skillsets across multiple environments can jointly construct and maintain these skillsets.

In this paper, we consider a blockchain-based on-chain model training and updating solution. We define the chain of skillset (CoS) as a blockchain system consisting of a consortium blockchain that stores all the global skillsets $\mathcal{S}$ as well as the corresponding AI models and a number of parallel training trains for collaborative training of each agent s_i. Each agent, if authorized to access the CoS, can jointly train and update the AI models on the CoS using its local dataset.

B. Problem Formulation

While security is a paramount performance metric for AgentNet, its enhancement often introduces a trade-off with other critical system properties, such as model performance and resource consumption. For instance, the implementation of more rigorous security criteria to verify an agent's qualifications–both for network participation and for collaborative learning–may inadvertently exclude agents with unique or unconventional characteristics. This exclusion can lead to unintended consequences, including model bias, slower convergence of distributed learning processes, and increased computational overhead.

Suppose the model of each skillset s_i is parameterized by w_i that can output the specific act $y\in\mathcal{Y}$ based on observed states $x\in\mathcal{X}$. Specifically, for a specific skillset $s_i\in\mathcal{S}:=\bigcup_{k\in\mathcal{K}}\mathcal{S}_k$, we assume a subset of agents $\mathcal{R}_i\subseteq\mathcal{K}$ possess this skillset and are authenticated to join the system.

In this paper, we mainly focus on the following key performance metrics:

Security level: The security level of a consensus-based blockchain system can be paradoxically evaluated by the total

number of agents that have been authorized to participate in the model training and updating on a training chain in the CoS. While the conventional assumption is that a larger number of nodes enhances a network's resilience, in permissioned or authorized systems, this metric can indicate increased vulnerability because the more authorized agents, the higher the risk for malicious agents or actors have been granted entry. Consequently, the larger the set of authorized agents, the higher the risk of a successful collusion attack where these malicious agents can agree to a malicious consensus. Therefore, the total number of authorized agents serves as an important measure of the overall network's security level. In this paper, we use $\mathcal{R}_i$ to denote the universal set of agents that have skillset s_i, and $\mathcal{M}_i$ to denote the subset of agents that can pass the authorization process of CoS to maintain and update skillset s_i. Therefore, the number of agents M_i in the set $\mathcal{R}_i$ is proportional to the level of security of the CoS, i.e., we define the security level of the CoS as $L(M_i)$ where L can be any security measurement function that is proportional to the total number of agents M_i being authorized to train and update CoS.

Skillset performance: The performance of each skillset is defined as the performance of its AI model that is jointly trained by the set of authenticated participating agents $\mathcal{M}_i$. This can be measured by the gap between the losses $F_i(w_{\mathcal{M}_i,T})$ of the model trained by M_i authorized agents and T coordination iterations and the optimal loss value $F^*_{\mathcal{R}_i}$ of the skillset s_i, which can be written as

$$\mathcal{E} = \mathbb{E}[F_i(w_{\mathcal{M}_i,T})] - F^*_{\mathcal{R}_i}. \tag{1}$$

Resource cost: While CoS offers distinct security benefits, it also introduces a higher total resource cost due to the additional communication and computational resources required for distributed ledger operations. The computational cost of CoS development is closely related to the total number of agents M_i participating in the training and updating of each specific skillset s_i and the communication cost depends mainly on the number of multi-agent coordination iterations T for skillset/model training.

In this paper, we introduce TrustAgentNet, a novel AgentNet architecture that can balance the three-way tradeoff among security level, skillset performance, and resource cost. The TrustAgentNet relies on the CoS, a consortium blockchain-based framework to achieve identity authorization and management of agents and also secure AI model and skillset development and updating among various agents in different environments.

III. TrustAgentNet Architecture

In this section, we propose a novel consortium-blockchain-based trustworthy AgentNet framework, called *TrustAgentNet*, to support trusted authentication, verifiable and traceable skillsets and tags acquisition and privacy-preserving skillset evolution of various types of agents across different environments.

A. Architectural Framework

The architectural framework TrustAgentNet is illustrated in Fig. 1, consisting of the following key components:

- **Infrastructure:** includes the hardware infrastructure, such as cloud and edge computing and storage resources and communication networks that connect agents and the CoS, and the software systems including high-quality datasets, accumulated skillsets, and comprehensive world models available for agents' utilization.
- **Skillset chain**: is a consortium blockchain in the CoS which contains the following subfunctional modules: 1) *ID Authentication*: provides authenticated digital ID for new agents k joining the network; 2) *Skillset Storage*: stores all the available and up-to-date skillsets s_i, $s_i \in \mathcal{S}$, each associated with a tag to describe its capability and life-cycle traceability. 3) *Tag and skillset authorization*: delivers trusted tags and skillsets to agents according to their capability requests.
- **Training chain of a specific skillset** s_i: facilitates decentralized learning among agents with identical skillsets s_i in the CoS. While often initiated by a single agent, other agents interested in collaboratively training this skillset s_i can subsequently join. Joining an existing chain necessitates endorsement, commonly requiring a consensus of given percentage of the current participating agents. Upon successful endorsement, each agent can submit its locally trained model parameters to the training chain for aggregation into a shared global model. CoS accommodates numerous parallel training chains, with each chain focusing on the collaborative learning and evolution of a specific AI skillset $s_i \in \mathcal{S}$.
- **Agents**: include various types of task-oriented agents. Each agent k that implemented in a specific environment e has a unique set of locally observed dataset $\mathcal{D}_{k,e}$, and has a set of skillsets $\mathcal{S}_{k,e}$ so as to accomplish certain tasks $t \in \mathcal{T}_e$.
- **Skillset/Agent tag**: is a trusted digital certificate, endorsed by the CoS, used to advertise the capabilities of an agent, which allows dynamic capability discovery so an agent controller can find the most suitable agent for a given subtask. Typically, an agent tag can be a JSON metadata file, which includes the tags of all of its skillsets.
- **Agent controller**: coordinates the collaboration and task cognition actions among agents as defined in [1]. It consists of the following subfunctional modules: 1) *Task cognition*: formulates the corresponding task t on behalf of users' semantics; 2) *Task separation*: separates the task t into multiple subtasks that need to be performed in order to fulfill the user's semantic goal; 3) *Agent adaptation*: assigns each subtask to a specific agent according to the agent tags.

B. Operational Procedure

If a new agent wants to join TrustAgentNet, it needs to follow the following procedures:

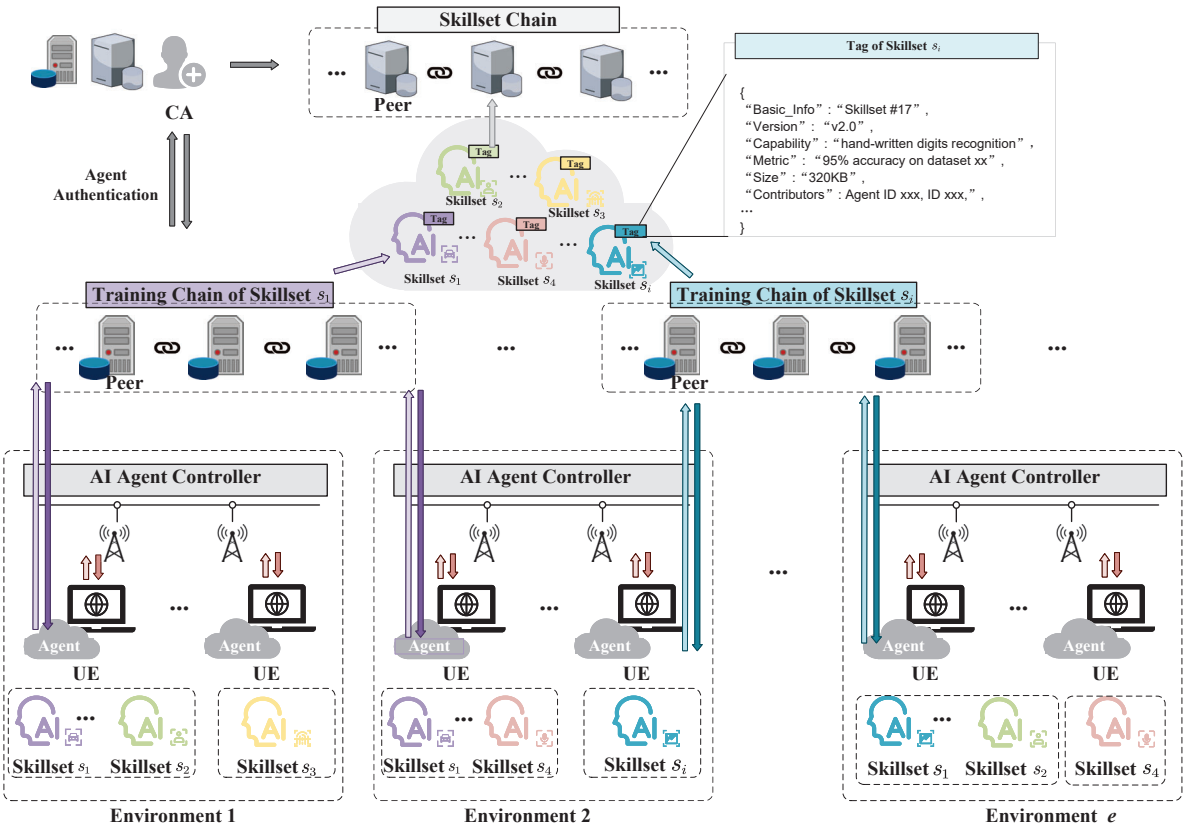


Fig. 1: An architectural framework of TrustAgentNet.

1) **Agent authentication:** To participate, an agent first applies for an identity from the consortium blockchain's Certificate Authority (CA). The CA then issues a digital certificate to the agent, encompassing a public key, a CA-signed certificate, and a corresponding private key, adhering to the standard Public Key Infrastructure (PKI) system.

2) **Task-skillset mapping:** In AgentNet systems, agents are tailored for specific tasks. Therefore, a task-skillset mapping agent is crucial for efficient operation, which serves as a central hub for associating specific tasks with the necessary skillsets required for their completion. Through continuous learning and accumulation of knowledge within its knowledge base, the mapping agent refines its understanding of the mapping functions over time. When a certificated agent has a task to perform, it submits its requirements to the mapping agent, which in turn replies with a list of the corresponding skillsets. This mechanism ensures that task objectives are effectively matched with agents possessing the appropriate capabilities within the network.

3) **Skillset acquiring and downloading from CoS:** The network's available skillsets, each tagged with descriptive capability and lifecycle traceability information, are stored on the peer nodes of the skillset chain's database in the CoS. Certificated agents can proactively acquire relevant skillsets and associated tags by submitting a proposal to the skillset chain. Notably, skillset acquisition is not limited to initialization but can occur regularly throughout an agent's operation. Once an agent obtains the necessary skillsets, its certificated tag – which can be stored as a JSON metadata file – is updated and then advertised to the agent controller for further task assignment.

4) **Skillset uploading to CoS:** Upon downloading a skillset s_i from the chain, the agent can perform local training on the base model using its private dataset. For each skillset s_i, a dedicated training chain i is established as illustrated in Fig. 1, where only allowed agents can join to conduct private and confidential model training. An agent possessing skillset s_i must be authenticated and authorized by a Membership Services Provider (MSP) to join training chain i in the CoS, after which it can submit its locally trained parameters as a transaction to this training chain.

5) **SkillSet evolving on CoS:** For each training chain i for skillset s_i in the CoS, after receiving updated model parameters from a number of authorized agents, a smart contract then aggregates these updates from multiple agents (e.g., using FedAvg) to compute refined global model parameters. The updated global model and tag of a skillset s_i is written into the public datasets of the CoS. Agents that joined the skillset chain can view and download the updated skillset data information in the public dataset, including model parameters and skillset tag. Consequently, skillsets on the chain can iteratively evolve through decentralized collaborative learning, leveraging different agents' unique datasets without compromising their data privacy.

IV. Theoretical Analysis on Three-way Tradeoff Among Security Level, Agent Performance, and Resource Consumption

This section presents the theoretical analysis of three-way tradeoff between security, agent skillset performance, and resource consumption in TrustAgentNet. As described in Section II-B, for a specific skillset $s_i \in \mathcal{S} := \bigcup_{k\in\mathcal{K}} \mathcal{S}_k$, a subset of agents $\mathcal{M}_i \subseteq \mathcal{K}$ that possess this skillset and obtain the consensus of the current participating agents will spontaneously engage in the co-training, which can be formulated as:

$$\min_{w_i\in\mathbb{R}^d} F_i(w_i) := \sum_{k\in\mathcal{M}_i} p_{i,k} F_{i,k}(w_i), \tag{2}$$

where $F_{i,k}(w_i) := \frac{1}{|\mathcal{D}_k|}\sum_{(x_{k,j},y_{k,j})\in\mathcal{D}_k} f(w_i; x_{k,j}, y_{k,j})$ represents the local objective function for agent k, $k \in \mathcal{M}_i$, and $p_{i,k}$ denotes the weight of agent k for skillset s_i, where $0 \le p_{i,k} \le 1$ and $\sum_{k\in\mathcal{M}_i} p_{i,k} = 1$. We define $f : \mathbb{R}^d \to \mathbb{R}^+$ as the non-negative loss function reflecting the error of the model w_i evaluated on sample $(x_{k,j}, y_{k,j})$.

Suppose the optimal model for skillset s_i is given by $w^*_{\mathcal{R}_i} \in \arg\min_{w_i} \sum_{k\in\mathcal{M}_i} p_{i,k} F_{i,k}(w_i)$ where $\mathcal{R}_i$ denotes the universal set of agents that have skillset s_i without considering security endorsement. We then have the following definition on the impact of the introduction of security authentication.

Definition 1: The impact of removing a subset $\mathcal{Q}_i = \mathcal{R}_i \backslash \mathcal{M}_i$ of agents on skillset s_i due to their failure to meet security certifications can be defined as difference between the global optimal performance with all R_i agents and that with only the M_i certified agents. This discrepancy is given by:

$$C^*_{\mathcal{M}_i}(F_i) = \sum_{k\in\mathcal{R}_i} p_{i,k}(F_{i,k}(w^*_{\mathcal{M}_i}) - F_{i,k}(w^*_{\mathcal{R}_i})). \tag{3}$$

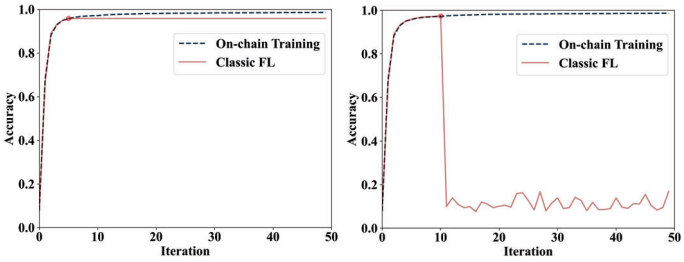
We can then derive the following theoretical bound of the skillset performance with security endorsement on blockchain.

Theorem 1: Suppose the following assumptions hold: *(i)* The objective function is L-smooth and μ-convex, i.e., $\frac{\mu}{2}\|w - w'\|^2 \le F_{i,k}(w) - F_{i,k}(w') - (w - w')^\top \nabla F_{i,k}(w') \le \frac{L}{2}\|w - w'\|$; *(ii)* The stochastic gradient of the loss function satisfies $\mathbb{E}\|\nabla F_{i,k}(w)\| \le G$ and $\mathbb{E}\|\nabla F_{i,k}(w) - \mathbb{E}[\nabla F_{i,k}(w)]\| \le \sigma_k$. Then, with $\kappa = \frac{L}{\mu}, \gamma = \max\{\frac{8L}{\mu}, E\}$, and the learning rate $\beta_t = \frac{2}{\mu(\gamma+t)}$, we have

$$\mathbb{E}[F_i(w_{\mathcal{M}_i,T})] - F^*_{\mathcal{R}_i} \le \frac{4\kappa}{\gamma + ET}\Big(\sum_{k\in\mathcal{R}_i}\frac{p^2_{i,k}\sigma^2_k}{D_k} + 8E^2G^2 + A_{\mathcal{M}_i} + 6LC^*_{\mathcal{M}_i}(F_i)\Big) + LB_{\mathcal{M}_i},$$

where $D_k = |\mathcal{D}_k|$ is the cardinality of the set $\mathcal{D}_k$, $A_{\mathcal{M}_i} = \frac{\mu^2(\gamma+1)}{4}\|w_0 - w^*_{\mathcal{M}_i}\|^2$, and $B_{\mathcal{M}_i} = \|w^*_{\mathcal{M}_i} - w^*_{\mathcal{R}_i}\|^2$.

Remark 1: Theorem 1 indicates that a higher level of authentication results in a fewer number of collaborative model constructing agents, which results in larger $B_{\mathcal{M}_i}$ and $C^*_{\mathcal{M}_i}$, that is, a larger performance gap from the optimal model, especially when the agents that fail the security verification have unique characteristics. A clear tradeoff between security and skillset performance emerges, controlled by the threshold for security authentication. Indeed, increasing the threshold leads to greater security but poorer performance, and vice versa. To guarantee the performance of the skillset and security, it requires more global collaboration T and local training E, which undoubtedly brings burdensome communication and computation overhead. In other words, if we aim to achieve security and stable performance at the same time, it requires more local computation by each agent and more knowledge sharing among agents. This shows another tradeoff between performance, security, and resource consumption.



(a) (b)

Fig. 2: Training accuracy versus the number of iterations with the on-chain training in TrustAgentNet and classic federated learning. (a) With single point of failure at the coordinator. (b) With poisoning attack at 30% agents.

V. Prototype and Experimental Results

In this section, we introduce our developed TrustAgentNet prototype and present experimental results under various scenarios.

A. Prototype and Experimental Setup

For our prototype implementation, we utilize a desktop server equipped with an Intel i5-10500T CPU and 64GB of memory to host both the blockchain nodes and the agents. We establish a network with 20 users, each implemented with an agent and a unique local dataset. A specific skillset is considered for illustration: a hand-written digits recognition model with LeNet-5 for MNIST. FedAvg is adopted as the model aggregation method.

We adopt the open source Hyperledger Fabric [5] as the consortium blockchain platform for both skillset chain and training chain. Etcdraft and Raft are adopted as the consensus protocol and algorithm, respectively. Specifically, the skillset chain is configured with three orderer nodes and involves seven distinct organizations. Each individual training chain is set up with three orderer nodes and comprises five organizations. Each organization has one peer node.

B. Security Enhancements for Skillset Training under Malicious Attacks

To evaluate the security enhancements of TrustAgentNet for distributed learning comparing to classic federated learning, we examine its resilience against two prevalent attack types: (a) attacks targeting the coordinator server, leading to a single point of failure, and (b) poisoning attacks where 30% agents

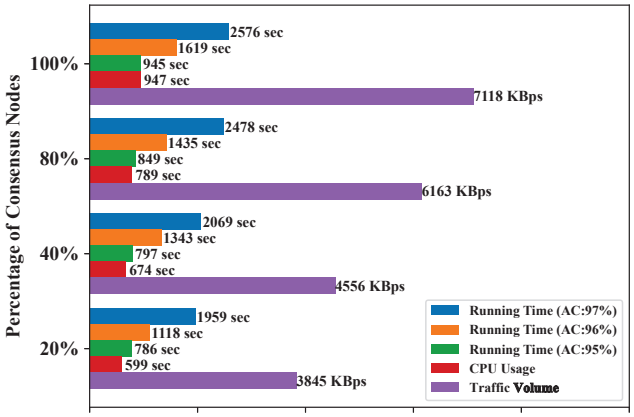


Fig. 3: Running time, real-measured CPU occupancy time, and traffic volume under different percentage of nodes for agreeing to the endorsement in the TrustAgentNet prototype.

maliciously manipulate their uploaded local model parameters. As illustrated in Fig. 2(a), when the coordinator server is shut down at iteration $\#5$, classic federated learning ceases to function; while the TrustAgentNet with on-chain training continues training uninterrupted thanks to the distributed ledgers on all the servers. Fig. 2(b) further demonstrates that when 30% of the participating agents are compromised by poisoning attacks, classic FL fails to converge whereas TrustAgentNet maintains convergence with the consensus and endorsement mechanisms. Experimental results underscore the improvement in security offered by TrustAgentNet.

C. Three-way Tradeoff among Security Level, Agent's Training Performance and Resource Consumption

To empirically validate the inherent trade-off between security level, skillset training, and resource consumption within the proposed TrustAgentNet, as analytically established in Section IV, we conduct an experiment on our prototype. This experiment involved 20 agents, each equipped with an identical skillset for hand-written digit recognition but possessing a unique non-i.i.d. dataset. For the training chain associated with this skillset, we preconfigured five peer nodes to endorse the participation of agents in the training process. The endorsement policy stipulated that for an agent to be granted access, $x\%$ of these peer nodes must agree. We further assume a linear correlation between the increase in $x\%$ and the number of agents failing to meet the security certification requirements. Each agent that successfully joined the training process performs local model training for 5 epochs before submitting its updated parameters to the training chain.

As depicted in Fig. 3, we conducted a comparative analysis of the real-measured running time, CPU occupancy time, and traffic volume under varying percentages of nodes required to agree on an endorsement. Specifically, running time is defined as the cumulative time duration for the agents to collaboratively train the model until a target accuracy is achieved. CPU occupancy time represents the aggregate time during which the CPUs of the blockchain servers are utilized. Lastly, traffic volume refers to the total volume of data transmitted across the blockchain servers. Fig. 3 clearly demonstrates that enhancing system security by demanding agreement from a higher percentage of peer nodes for endorsement leads to increased computational demands (CPU usage) on the blockchain infrastructure. This is a direct consequence of a greater number of nodes needing to validate transactions according to the endorsement policy. Furthermore, the communication overhead escalates substantially as all endorsing peers are required to broadcast their endorsement decisions and digital signatures across the blockchain network. Notably, the running time for collaboratively training the skillset to achieve target accuracies of 95%, 96%, and 97% also increases with a greater number of required endorsing peer nodes, indicating that improvements in security come at the expense of the distributed training performance of the skillset.

VI. Conclusion

This paper proposes TrustAgentNet, a novel consortium-blockchain-based framework to foster trustworthy AgentNet. The framework provides distributed-ledger-based authentication to support secure and unified digital identify management for agents, trusted acquisition and updating of agents' skillsets on the chain, and life-long evolution of skillsets via on-chain collaborative learning. We have detailed the general architectural framework and introduced its primary operational procedures. Our theoretical analysis reveals an intrinsic trade-off between the achieved level of security and the performance of skillset training, alongside the consumption of computational and communication resources, which is further validated by experimental results. Finally, by considering a use case of collaborative learning mode selection under malicious attacks, we have demonstrated the efficacy of the proposed TrustAgentNet in significantly enhancing the security of AI agents during distributed collaboration and interaction.

Acknowledgment

This work is supported in part by the National Natural Science Foundation of China (NSFC) under grants 62571208, 62525109, 62301516, 62293483, 62293480, and 62293481, and the Mobile Information Network National Science and Technology Key Project under grant 2024ZD1300700.

References

[1] Y. Xiao, G. Shi, and P. Zhang, "Towards agentic AI networking in 6G: A generative foundation model-as-agent approach," *IEEE Communications Magazine*, vol. 63, no. 9, Sep. 2025.

[2] Y. Xiao, G. Shi, Y. Li, W. Saad, and H. V. Poor, "Toward Self-Learning Edge Intelligence in 6G," *IEEE Communications Magazine*, vol. 58, no. 12, pp. 34–40, Dec. 2020.

[3] K. Dev *et al.*, "Advanced architectures integrated with agentic ai for next-generation wireless networks," *arXiv preprint arXiv:2502.01089*, 2025.

[4] X. Li *et al.*, "The agentic-ai core: An ai-empowered, mission-oriented core network for next-generation mobile telecommunications," *Engineering*, 2025.

[5] E. Androulaki *et al.*, "Hyperledger fabric: a distributed operating system for permissioned blockchains," in *ACM EuroSys*, Porto, Portugal, Apr. 2018, pp. 1–15.