Proceedings of CONF-SEML 2026 Symposium: Computational Analysis and Modeling in Complex Intelligent Systems
DOI: 10.54254/2755-2721/2026.GU33556

A Comprehensive Review of Threat Models in Federated AI Agents

Xintong Ji

Department of Data Science and Artificial Intelligence, The Hong Kong Polytechnic University, Hong Kong, China
xintong.ji1207@gmail.com

Abstract. The rapid growth of Large Language Model (LLM) agents and their ability to work as federated systems creates both exciting opportunities and new complex security challenges. This survey provides systematic insights into the threat models and defensive paradigms for federated trustworthy AI agents from 2024 to 2026. The threats focus on the different phases of training and runtime execution, as well as collaborative protocols that focus on trust boundaries and inter-agent authorization dilemma resolutions. Some of the findings highlight the following problems: indirect prompt injections in RAG pipelines, vulnerabilities caused by tool use, and "infectious jailbreaks," that spread among different agents. The research looks for chances in gaps among border line standards such as Model Context Protocol (MCP) and the "Security Tax," which examines the balance between a robust defense and loss of collaborative effectiveness, and dynamic trust. New defensive paradigms explaining self-evolving immune mechanisms, identity-centric security, and zero-trust paradigms are briefly discussed. Also, the research presents plans that move from a filtering approach to a trust-consistent self-managed and self-healing approach toward a distributed agent system.

Keywords: Federated AI Agents, Threat Models, LLM Security, Prompt Injection, Model Context Protocol

1. Introduction

The rise of large language model (LLM) intelligence has given rise to autonomous AI systems that can perform complex tasks in multiple application scenarios. Unlike single LLM, these intelligent bodies run in a federal environment and interact with various intelligent bodies, tools and services across different boundaries of trust. This architecture evolution introduces security risks unique to multi-intelligent composition, including new vulnerabilities that are not available in single models, such as prompt word injection, tool manipulation, permission upgrade and protocol vulnerabilities [1-3].

Relying on the federal learning framework, the federal AI intelligence realizes cross-organizational collaboration to ensure data ownership. Although traditional federal learning research has made progress in resisting poisoning and backdoor attacks [4-7], the rapid deployment of intelligent AI in enterprises and open environments has made it face the threat of confrontational

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.GU33556

attacks such as prompt infection, malware transmission and hijacking between intelligent bodies [8,9]. At present, the literature is still relatively fragmented, focusing mostly on isolated research on single-intelligent, federal learning or multi-intelligent attacks. There is an urgent need for a global method to systematically review threats at the training, runtime and protocol levels.

This paper aims to fill this research gap and comprehensively sort out the threat model, attack methods, system vulnerabilities and cutting-edge defense schemes of the federal trusted AI intelligence. This paper focuses on cross-organizational trust mechanisms, dynamic authorization and identity-based protocols, and especially focuses on the federal multi-intelligent system. By integrating the latest research and industrial interoperability standards such as MCP and A2A, this paper provides a theoretical blueprint for building a safe, interoperable and reliable AI intelligent deployment system [10-12].

Table 1. Threat taxonomy for federated AI agents

Threat Category	Attack Target	Representative References
Training-time	Model parameters, client data	[4-7]
Runtime	Agent execution, prompts, tools	[1,3,13-15]
Protocol-level	Communication protocols, cross-agent trust	[10-12]

2. Threat taxonomy

Threats to federal AI intelligence can be classified according to their life cycle and trust boundaries. Unlike previous studies that only classified a single threat, this paper takes a more cautious perspective and delves into the impact of each type of threat and the vulnerability behind it. As shown in Table 1, this paper focuses on three core categories: training-stage threats, runtime threats and protocol-layer threats, which bring unique challenges in a distributed multi-intelligent environment.

Training-time threats mainly occur in the coordination of model training and federal learning. Poisoning and backdoor attacks will directly affect the model parameters [5,7]. In addition, Discrepancies in the federal aggregation process may also be used to weaken the effectiveness of the global model [4,6]. Although defense mechanisms such as robust aggregation and abnormal detection can be deployed, its effect often depends on the number of malicious intelligences and the diversity of local data. Therefore, in the large-scale federal learning deployment with semi-trusted or unknown participants, the threat of the training phase poses a serious challenge.

Runtime threats occur in the process of task execution, interaction between intelligent bodies, and interaction with tools. Prompt Injections and tool call vulnerabilities may manipulate the behavior of the intelligent body or lead to the leakage of confidential information; at the same time, attacks on the retrieval enhanced generation (RAG) pipeline may cause damage to the model output [1,3,13,15] Due to the dynamics of the intelligent body and the variety of tools used, it is extremely difficult to deal with runtime threats. In order to resist such threats, the intelligent body needs to have real-time monitoring and abnormal detection capabilities, but this often increases the computing burden and prolongs the response time.

Protocol-level threats originate from attacks on communication and coordination protocols, which are the basis of cross-organization and cross-intelligence collaboration. Defects in the model context protocol (MCP) or intelligent inter-body (A2A) trust mechanism may lead to permission upgrade, illegal authorization or message tampering [10-12]. Cross-protocol vulnerabilities have also opened up a new attack path, and its defense may require protocol consistency, verifiable

DOI: 10.54254/2755-2721/2026.GU33556

credentials and Zero-trust schemes. It is worth noting that although protocol-level measures can mitigate systemic threats, incorrect configuration or lack of unified adoption standards may still lead to security vulnerabilities in a diversified intelligent environment.

3. Key challenges

Federal AI intelligence operates across different organizational trust domains and dynamic access strategies, which leads to loopholes that go beyond abstract "trust uncertainty" and have practical usability.

Trust-authorization mismatch makes power abuse and "Confused-deputy" attacks in cross-organizational workflows possible. The intelligent body runs under delegated permissions (including API keys and OAuth tokens), but lacks a runtime trust decomposition mechanism: the low-privilege intelligent body may induce the high-privileship equipe to perform unauthorized operations or inherit excessive permissions through loose trust relationships between the intelligent bodies. This violates the principle of minimum privilege and leads to the accidental excessive disclosure of sensitive data in the process of benign collaboration. The problem is formalized as "trust-authorization mismatch", that is, static access control cannot track the dynamic changes of trust at runtime [8,9].

In addition, the security-intelligence paradox has been confirmed in specific prompt injection and poisoning attacks that spread across intelligent boundaries. In addition to the basic prompt hijacking, Prompt infection realizes self-replicating LLM-to-LLM injection: malicious prompts embedded in documents or messages spread from one intelligence to another, forcing them to steal data or tools. Misuse [3]. For example, AgentPoison pollutes long-term memory or RAG knowledge base through hidden backdoors; when triggers appear, the intelligent experience retrieves malicious examples and performs confrontational behaviors (such as unauthorized API calls, violations of policies, etc.), and there are no obvious abnormalities [15]. Although tightening information sharing can alleviate these risks, it will weaken the responsiveness and interoperability of intelligent bodies; while relaxing sharing will amplify the offensive side [1,2].

Finally, Disparities in protocol interoperability has created exploitable vulnerabilities in cross-smart communication and tool arrangement. The use of heterogeneous frameworks (such as MCP, A2A and Agora) often leads to inconsistency between data formats and authentication mechanisms, thus contributing to insecure message delivery, identity deception and unauthorized tool execution [11]. Attackers may use these verification vulnerabilities to launch protocol-level attacks, including manipulating Context frames and forging intelligent inter-body requests to disintegrate the arrangement layer [1,12]. Therefore, the establishment of verifiable identity and unified security protocols is crucial to alleviate the inherent vulnerabilities of cross-ecosystem interaction [10].

4. Defensive paradigms

To mitigate security threats in federal AI intelligence, it is necessary to realize a shift from "model-centered" static defense to adaptive, identity perception and zero-trust architecture. Based on the recent research results of multi-intelligent security and federal trusted AI, this paper summarizes the main defense paradigms as follows.

Dynamic runtime monitoring realizes continuous audit of intelligent body interaction and tool calls, and can detect abnormal behaviors such as prompt word injection, hidden collusion or external API overruntime call in real time [9,13]. Compared with static defense, runtime monitoring allows the intelligent body to mitigate sudden threats during operation while maintaining its collaborative

DOI: 10.54254/2755-2721/2026.GU33556

ability. Its core technologies include Trajectory-level policy enforcement, abnormal detection of intelligent inter-body communication, and input and output Sanitization, etc.

Identity-based security uses intelligent identity authentication, authentication and Verifiable Credentials to manage trust boundaries. By establishing a robust identity and access management (IAM) mechanism, the system can perform fine-grained permission control, detect impersonation, and accurately trace malicious operations [14]. Encryption technology, verifiable logs and tamper-proof credentials help to strengthen the Accountability mechanism and selective information disclosure across intelligence bodies, thus reducing the risk of misuse of trust in the federal environment.

Zero-trust architectures and self-healing mechanisms assume that any intelligence or communication channel is not inherently reliable. Each operation must be authenticated, authorized and continuously verified [1,8]. The self-healing security mechanism enables the intelligent body to automatically detect damaged states, malicious prompts or abnormal behaviors, and initiate corrective measures, such as isolating affected instructions or resetting the damaged memory state.

5. Conclusion

In summary, this paper builds a comprehensive threat and vulnerability framework, describing the main obstacles and defense strategies faced by federal AI intelligence. The security picture is divided into three distinct levels: vulnerabilities in the training stage, risks in the reasoning stage (execution period), and the systemic challenges caused by the decentralized multi-intelligent environment with dynamic cross-organization trust boundaries and authority management. The study identified three core challenges: the ambiguity of the trust boundary, the trade-off between security and information gain, and the lack of interoperability between heterogeneous components. These problems can be solved by evaluating existing protective measures, implementing people-oriented real-time dynamic monitoring, and adopting a "zero trust" architecture equipped with a self-repair protocol. Together, these measures form a comprehensive security framework in an open federal environment.

Looking to the future, the realization of reliable federal AI systems will depend on the "Secure-by-design" standardization framework that integrates scalable self-recovery functions and adaptive protection mechanisms. This evolutionary trend means that there will be a deeper integration between academic research and business implementation. With the rapid popularity of large language model platforms in various industries, building a robust, commutable and secure multi-intelligent system will depend on rigorous threat modeling, standardized protocol verification and continuous trust management.

References

[1] Ferrag, M., et al. (2025). From prompt injections to protocol exploits: Threats in LLM-powered AI agents workflows. arXiv preprint arXiv: 2503.00657.
[2] He, F., et al. (2024). The emerged security and privacy of LLM agent: A survey with case studies. ACM Computing Surveys, 1–36.
[3] Lee, D., & Tiwari, M. (2024). Prompt infection: LLM-to-LLM prompt injection within multi-agent systems. arXiv preprint arXiv: 2406.12244.
[4] Abad, G., Picek, S., & Urbieta, A. (2021). SoK: On the security & privacy in federated learning. arXiv preprint arXiv: 2102.11290.
[5] Bagdasarian, E., et al. (2018). How to backdoor federated learning. arXiv preprint arXiv: 1807.00456.
[6] Hallaji, E., et al. (2024). Decentralized federated learning: A survey on security and privacy. IEEE Transactions on Big Data, 194–213.

DOI: 10.54254/2755-2721/2026.GU33556

[7] Lyu, L., et al. (2020). Privacy and robustness in federated learning: Attacks and defenses. IEEE Transactions on Neural Networks and Learning Systems, 8726–8746.
[8] Shi, G., et al. (2025). SoK: Trust-authorization mismatch in LLM agent interactions. arXiv preprint arXiv: 2503.00788.
[9] He, P., et al. (2025). Comprehensive vulnerability analysis for trustworthy LLM multi-agent systems. arXiv preprint arXiv: 2501.04720.
[10] Atri, S. (2026). Standardizing AI agent protocols: Definitions, layered architectures, and future research directions. International Journal of Artificial Intelligence, Data Science, and Machine Learning.
[11] Anbiaee, Z., et al. (2026). Security threat modeling for emerging AI-agent protocols: A comparative analysis of MCP, A2A, Agora, and ANP. Unknown Journal.
[12] Maloyan, N., & Namiot, D. (2026). Breaking the protocol: Security analysis of the model context protocol specification. arXiv preprint arXiv: 2601.04567.
[13] Arzanipour, A., et al. (2025). RAG security and privacy: Formalizing the threat model and attack surface. arXiv preprint arXiv: 2501.03810.
[14] Topol, A., et al. (2025). Secure authentication design for AI agents. AHFE International.
[15] Chen, Z., et al. (2024). AgentPoison: Red-teaming LLM agents via poisoning memory or knowledge bases. arXiv preprint arXiv: 2406.12781.