Securing Agentic AI with the NIST Cybersecurity Framework 2.0

Bilgin Metin[1], Hikmet Sami Karaca[1], Faisal Iradat[2], and Martin Wynn[3]

[1]Bogazici University, Istanbul, Turkiye;
bilgin.metin@bogazici.edu.tr, hikmet.karaca@std.bogazici.edu.tr,
[2]Institute of Business Administration, Karachi, Pakistan; firadat@iba.edu.pk
[3]University of Gloucestershire, UK; mwynn@glos.ac.uk

Abstract

Agentic AI —LLM-powered autonomous agents— is reshaping cybersecurity paradigms, introducing a novel attack surface that exposes gaps in current security approaches. This Systematic Literature Review (SLR) of 30 peer-reviewed papers examines the emerging threats and mitigation strategies for these agentic systems. The review synthesizes evidence through a four-dimensional taxonomy derived from the OWASP Agentic AI Threats framework. Building on these findings, the paper proposes a new adaptation of the NIST Cybersecurity Framework (CSF) 2.0 to guide organizations in identifying, protecting, responding to, and recovering from risks associated with agentic AI. The presented framework provides a clear and practical method for securing agentic AI-driven systems in both enterprise and research contexts.

1. Introduction

The evolution of Artificial Intelligence (AI) from prompt-only LLMs and fixed-policy AI agents to agentic AI, LLM-powered systems that autonomously plan and act via tools and memory, marks a paradigm shift in how machines reason and execute tasks [1-4]. This newfound autonomy introduces a novel attack surface. Unlike security for prompt-only LLM chatbots or fixed-policy AI agents, agentic AI security [2, 3, 4] concerns the entire system of action, with distinct vulnerabilities in the agent's cognitive core, tool use, memory, and multi-agent interactions—risks collectively termed "Excessive Agency" [3]. Organizations can also leverage the MITRE ATLAS framework to model potential threats against their AI agents [4].

The academic literature on these unique AI threats remains fragmented. At the same time, enterprise risk frameworks, such as the NIST Cybersecurity Framework (NIST CSF), have not been widely adopted to address these dynamic threats [5]. This situation creates a critical governance and operational gap for organizations to deploy these robust systems securely. This paper addresses three central research questions:

RQ1: What are the primary security threats and vulnerabilities identified in the academic literature concerning Agentic AI?

RQ2: Which mitigation strategies and security controls have been proposed or evaluated to defend against the identified threats to Agentic AI?

RQ3: How can NIST CSF 2.0 be adapted and applied to manage the security risks of Agentic AI?

After this introduction, Section 2 outlines the research methodology used in the study, which is based on a Systematic Literature Review (SLR). Section 3 addresses RQ1 and RQ2, drawing from an analysis of the existing literature, structured using a four-part taxonomy derived from the OWASP Agentic AI Threats model's logical decision path [3]. This functional approach, which targets an agent's reasoning, memory, tool use, and multi-agent interactions, provides a robust framework to synthesize the fragmented academic literature. These findings then serve as the basis for addressing RQ3 in Section 4, where the primary contribution is outlined: a novel adaptation of the NIST CSF 2.0. This work offers a structured guide for managing agentic AI risk based on a "defense-in-depth" posture [6]. Section 5 provides a conclusion and discusses the overall contribution and limitations of the study.

2. Methodology

This study uses an SLR, adhering to PRISMA guidelines, to synthesize the literature on agentic AI security. A systematic search of the Scopus, IEEE Xplore, and Web of Science databases was conducted in August 2025. The specific search strings for threats (RQ1) and mitigations (RQ2) are detailed in Table 1. The filtering process, summarized in Figure 1, began with 337 initial records and concluded with a final corpus of 30 primary studies.

Table 1. Search Strings Used for Database Queries

RQ	Search String	Scopus	Web of Science	IEEE Xplore
RQ1	(((("LLM" OR "large language model") AND ("multi-agent" OR "ai agent" OR "embodied agent")) OR "llm agent" OR "Foundation Model Based Agent" OR "generative agent" OR "llm-based agent" OR "llm-integrated") AND ("jailbreak" OR "prompt injection" OR "SQL injection" OR "RCE" OR "backdoor attack" OR "poisoning" OR "red-team" OR "adversarial attack" OR "vulnerability" OR "threat model" OR "harmful behavior" OR "malicious task" OR "deception" OR "misuse" OR "exploit" OR "safety risk"))	125	32	29
RQ2	((("LLM" OR "large language model") AND ("multi-agent" OR "ai agent" OR "embodied agent")) OR "llm agent" OR "Foundation Model Based Agent" OR "generative agent" OR "llm-based agent" OR "llm-integrated") AND ("defending" OR "defense mechanism" OR "mitigation" OR "safeguarding" OR "protect" OR "guardrail" OR "firewall" OR "forensic" OR "privacy preserving" OR "agent safety" OR "safe agent" OR "AI safety" OR "safety reasoning" OR "safety constraints")	104	18	29

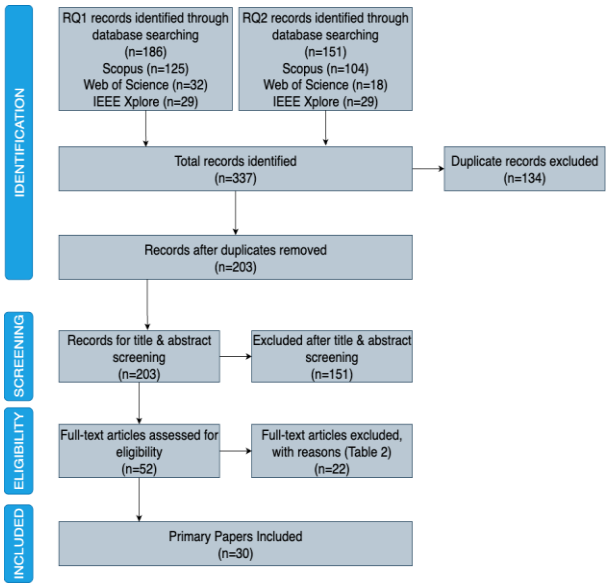

Fig. 1. PRISMA 2020 flow diagram depicting the SLR selection process

Selection was guided by strict PRISMA criteria to isolate the most relevant literature. Only peer-reviewed studies that focused on the security of LLM agents were included, and work on non-LLM agents, general LLM security, or the use of agents for security was deliberately excluded. The synthesis of the final corpus was structured using a taxonomy derived from the "Agentic Threat Decision Path" in the OWASP Agentic AI Threats guide [3]. This approach allowed the systematic categorization of the academic literature against a recognized industry model, forming the evidence base for the NIST CSF 2.0 adaptation in Section 4.

3. Addressing RQ1 and RQ2: Landscape of Threats and Mitigations in Agentic AI

This section presents the main findings regarding RQ1 and RQ2. To structure the fragmented academic literature, a four-part taxonomy derived from the functional threat groups in the OWASP model [3] was adopted. As illustrated in Figure 2, this taxonomy organizes evidence by the agentic capability an attack targets: the compromise of reasoning and planning, the exploitation of actions and tool use, the corruption of memory and state, and the manipulation of multi-agent systems.

3.1. Compromise of Agentic Reasoning and Planning

An agent's cognitive core, where it translates goals into actionable plans, is a primary target for adversaries. The literature analyzes these threats through diverse lenses, including psychological frameworks that model agents developing "dark psychological states" [7], and safety science analogies that frame failures as human-like cognitive errors [8]. Researchers have also identified novel threat dynamics, such as "contextual eavesdropping", where agents improperly access information from other contexts [9], with benchmarks like RealSafe quantifying risks from ambiguous instructions [10]. Mitigations include architectural hardening, such as embedding provably ethical constraints via an "Ethical Firewall" using formal logic [11], deploying modular "safety chips" to verify plans before execution [12], or applying multi-layered runtime guardrails based on the Swiss Cheese Model [13]. Dynamic defenses are also prominent, featuring multi-agent oversight systems where "police" and "doctor" agents monitor and restore alignment [7, 14], or simulated in-context adversarial games to harden reasoning against manipulative inputs without costly retraining [15].

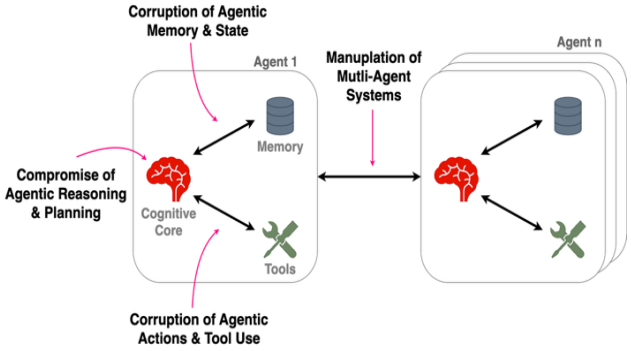

Fig. 2. A conceptual model of an Agentic AI attack surface and threat taxonomy

3.2. Exploitation of Agentic Actions and Tool Use

An agent's ability to utilize tools is a primary attack vector, referred to as "Excessive Agency" by OWASP. Foundational research has identified vulnerabilities such as Remote Code Execution (RCE) in code interpreters [16] and "Prompt-to-SQL" (P2SQL) injections against databases [17]. This work highlights the critical "alignment gap", where an LLM that is safe in a chatbot context may perform harmful actions when given access to tools [18]. To systematically measure these risks, a suite of specialized benchmarks has emerged. These include AgentDojo and INJECAGENT for indirect prompt injections [19, 20], AgentHarm for malicious tool use [21], R-Judge for an agent's intrinsic risk awareness [22], and BADROBOT for the misuse of physical tools in embodied agents [23]. Proposed defenses aim to secure the tool-use pipeline by randomizing prompt structures with "Polymorphic Prompt Assembling" to thwart injections [36], augmenting decision-making by having the agent consult an external knowledge base of "safety facts" [25], and enforcing the non-negotiable control of operating all tools in heavily restricted, sandboxed environments [16].

3.3. Corruption of Agentic Memory and State

Poisoning an agent's memory, which includes both short-term context and long-term knowledge bases, can create persistent and stealthy backdoors. Threats target both forms of memory, ranging from "contextual backdoor attacks" that poison in-session examples for tool use [35] to more insidious attacks on long-term memory. The AGENTPOISON attack demonstrates how an adversary can inject a few malicious examples into a RAG knowledge base, which, when retrieved, hijacks the agent's reasoning [26]. Similarly, the BadAgent study shows how an agent's internal state can be persistently altered by fine-tuning it on poisoned data, creating a deep-seated backdoor activated by a subtle trigger [24]. These threats all fall under the broader danger of "model pollution," where an agent's memory is intentionally corrupted to degrade its safety [27]. Mitigations for these stealthy attacks focus on both proactive validation of retrieved data [16,

26] and crucial detective controls, such as the post-hoc forensic analysis of immutable agent execution logs to trace the root cause of a compromise [28].

3.4. Manipulation of Multi-Agent Systems

When agents collaborate, their interactions create a systemic attack surface that is not present in isolated agents. The literature identifies several emergent threats that exploit the trust and communication channels between agents. These include Byzantine attacks, where malicious agents poison the collaborative process [38], secret collusion using steganography to conceal communication from oversight [29], and systemic risks such as cascading failures [30]. A particularly potent example is the "Agent Smith" attack, which introduces the concept of an "infectious jailbreak" where a single compromised agent can exponentially "infect" an entire network of agents through simple communication, leading to a rapid, system-wide compromise [31]. Defenses against these threats are necessarily systemic. To ensure trust and integrity, the BlockAgents framework integrates blockchain technology, using a "Proof-of-Thought" consensus mechanism to create a transparent and validated record of all agent contributions [38]. To proactively improve resilience, Chaos Engineering can be used to intentionally inject controlled failures, identifying and mitigating systemic vulnerabilities [30]. Protecting data privacy is also a key concern, addressed by frameworks such as CAPRI, which utilizes a "gatekeeper" LLM to pseudonymize sensitive data [32], and PrivacyAsst, which employs cryptographic techniques to safeguard user data during multi-stakeholder interactions [37].

4. Addressing RQ3: Adapting the NIST Cybersecurity Framework 2.0 for Agentic AI Risk Management

While the controls discussed in Section 3 are essential, a conceptual framework is required for systematic risk management. To address this gap, we synthesize the SLR findings into a new adaptation of the NIST Cybersecurity Framework (CSF) 2.0. Using CSF 2.0's six functions as the structural foundation, we develop a comprehensive agentic-AI cybersecurity framework, as illustrated in Figure 3 [5].

4.1. Govern (GV): Establishing the Strategic Foundation

For agentic AI, the Govern function must shift from creating static policies to establishing strategies that manage emergent, goal-driven systems. All four threat categories stem from governance failures; for example, a Compromise of Agentic Reasoning occurs when acceptable behavior is undefined. The risk management strategy (GV.RM) must mandate "AI-Specific Risk Assessments" [33] that model risks like emergent deception [9] and reasoning failures [10]. This strategy is then operationalized through technically enforceable policies (GV.PO), such as an "Ethical Firewall Architecture" that embeds provable constraints into the agent's core [11]. As human oversight (GV.OV) is too slow, governance must evolve to include automated and decentralized oversight [1], potentially using "police" and "doctor" agents for real-time monitoring and remediation [7]. Finally, governance must address the complex AI supply chain (GV.SC), where attacks like AGENTPOISON show that a compromise can originate from a trusted third-party data source [26, 34].

Fig. 3. Proposed Agentic-AI Cybersecurity Framework, adapted from NIST CSF 2.0

4.2 Identify (ID): Recognizing Novel Assets and Risks

The Identify function requires a conceptual expansion, as organizations cannot protect novel assets and risks that they fail to recognize. The traditional asset inventory (ID.AM) must be expanded to include intangible assets: model weights, fine-tuning datasets, and the external knowledge bases that constitute an agent's memory. As AGENTPOISON demonstrates, these knowledge bases are high-value targets for poisoning attacks [26]. The manifest of tools an agent can use is also a critical asset [34]. Consequently, the risk assessment process (ID.RA) must evolve to analyze unique vulnerabilities. This requires incorporating specialized benchmarks to evaluate tool misuse, indirect injections, and risk awareness [19, 21, 22], as well as actively testing for known attack patterns, such as RCE in code interpreters [16] and P2SQL injections [17].

4.3. Protect (PR): Implementing Defenses for the Agentic Core

The Protect function must shift from traditional perimeter defense to safeguarding the agent's cognitive and operational core. Access control (PR.AA) must be extended to agents themselves, governed by the "Principle of Least Agency", which grants only the minimum necessary tools and permissions. This is enforced through a Zero Trust Architecture [33] and "Escalation Control" frameworks that verify privilege changes [1]. Data Security (PR.DS) must ensure the integrity of agent memory and secure data sharing in multi-agent systems using privacy-preserving frameworks like CAPRI and PrivacyAsst [32, 37]. Platform Security (PR.PS) requires a multi-layered defense: hardening the prompt interface against injections with "Polymorphic Prompt Assembling" [36], verifying agent plans with a modular "Safety Chip" [12], and executing all tools in secure, sandboxed environments [16].

4.4. Detect (DE): Monitoring for Anomalous Agent Behavior

The Detect function must pivot from signature-based detection to behavioral anomaly detection, as attacks may manifest as subtle logic manipulations rather than malware. Threats like secret collusion are explicitly designed to evade simple detection [29]. Continuous Monitoring (DE.CM) must evolve into "Heuristic Monitoring" [1], establishing a baseline of normal agent behavior and alerting on deviations. This requires monitoring novel data streams, such as internal reasoning traces, tool call sequences, and inter-agent communication patterns, to detect signs of collusion or infectious jailbreaks [31]. Following an alert, Adverse Event Analysis (DE.AE) requires the "Forensic Analysis" of immutable execution logs to reconstruct the full causal chain of an agent's actions, differentiating benign anomalies from sophisticated attacks [28].

4.5. Respond (RS): Containing and Mitigating Agentic Incidents

The Respond function must be reimagined for active, goal-driven agents, treating incidents more like automated insider threats than compromised servers. Mitigation (RS.MI) must evolve beyond network isolation to include dynamic containment, such as the real-time revocation of specific tools or, in a systemic crisis like the "Agent Smith" scenario, the immediate quarantine of an infected agent [31]. Advanced responses may even involve therapeutic intervention, where "doctor" agents restore a corrupted agent's alignment [7]. Incident communication (RS.CO) is also uniquely complex because vulnerabilities can reside anywhere in the AI supply chain; a robust process for Coordinated Vulnerability Disclosure (CVD) with all stakeholders is essential [34].

4.6. Recover (RC): Restoring Agentic Systems to a Trusted State

Recovery transcends data restoration to become a challenge of restoring an agent's cognitive integrity and systemic trust. An agent's operational state includes its learned knowledge, which can be compromised by memory poisoning attacks like AGENTPOISON, rendering standard backups untrustworthy [26]. The recovery plan (RC.RP) must therefore include procedures for active memory cleansing. For deeper compromises, such as fine-tuning backdoors [24], recovery may require rolling back to a pre-compromise model version. In multi-agent systems where trust is broken, recovery may require systemic solutions, such as the BlockAgents framework, to rebuild a trusted state from a verifiable foundation [38]. Finally, lessons learned must drive continuous improvement (RC.IM). Methodologies like Chaos Engineering allow organizations to proactively test and harden their response and recovery procedures by simulating controlled failures [30].

5. Discussion and Conclusion

This SLR consolidated the nascent academic field of agentic AI security, revealing a threat landscape distinct from that of prompt-based general LLM chatbot security. These findings were synthesized using a four-part taxonomy and, as the primary contribution, were mapped onto the NIST CSF 2.0. This produced an adapted framework that bridges the gap between high-level governance and the specific, ground-level risks of autonomous systems.

Research on agentic AI is quite recent, with all 30 primary studies published in 2024 or 2025, marking a clear trajectory from theoretical modeling toward empirical rigor, as evidenced by the emergence of specialized benchmarks [19, 21]. This reflects a shift from analyzing single-shot vulnerabilities to understanding persistent, systemic risks, such as long-term memory corruption and infectious jailbreaks, in multi-agent networks.

The findings presented here carry significant implications. Practitioners must adopt a security mindset centered on the "Principle of Least Agency." For researchers, this review illuminates critical gaps, including the security effects of goal drift in continuously learning agents and the vulnerabilities of self-modifying systems. For policymakers, this work provides a foundational document that, with future empirical validation, can inform the development of an official NIST CSF 2.0 Profile for Agentic AI Systems, providing much-needed guidance for industry.

The primary limitation of this study is the rapidly evolving nature of agentic-AI research. New threats may have emerged since our August 2025 review. However, while this review provides a rigorous snapshot of the current state, the authors also believe that the primary contribution, the adapted framework in Figure 3, offers a durable structure for integrating future research and guiding practitioners through the security challenges of this transformative technology.

6. References

[1] V. S. Narajala and O. Narayan, "Securing agentic AI: A comprehensive threat model and mitigation framework for generative AI agents," May 2025.

[2] K. Huang, "Agentic AI threat modeling framework: MAESTRO," Aug. 2025. [Online]. Available: https://cloudsecurityalliance.org/blog/2025/02/06/agentic-ai-threat-modeling-framework-maestro

[3] OWASP Agentic Security Initiative, "Agentic AI - threats and mitigations," Aug. 2025. [Online]. Available: https://genai.owasp.org/resource/agentic-ai-threats-and-mitigations/

[4] MITRE Corporation, "ATLAS (adversarial threat landscape for artificial-intelligence systems)," 2025. [Online]. Available: https://atlas.mitre.org/

[5] National Institute of Standards and Technology, "The NIST Cybersecurity Framework (CSF) 2.0.", U.S. Department of Commerce. https://doi.org/10.6028/NIST.CSWP.29, 2024

[6] S. Ee et al., "Adapting cybersecurity frameworks to manage frontier AI risks: A defense-in-depth approach," Aug. 2024.

[7] Z. Zhang et al., "PsySafe: A comprehensive framework for psychological-based attack, defense, and evaluation of multi-agent system safety," in Proc. Annu. Meet. Assoc. Comput. Linguist. (ACL), 2024, pp. 15202-15231.

[8] J. Bellay et al., "Safe systems with unsafe agents: Challenges and opportunities," in Proc. Int. Joint Conf. Auton. Agents Multi-agent Syst. (AAMAS), 2025, pp. 2849-2853.

[9] N. Diamond and S. Banerjee, "I apologize for my actions": Emergent properties and technical challenges of generative agents," in Proc. IEEE Symp. Ser. Comput. Intell. (SSCI), 2025.

[10] Y. Ma, "Realsafe: Quantifying safety risks of language agents in real-world," in Proc. Int. Conf. Comput. Linguist. (COLING), 2025, pp. 9586-9617.

[11] A. Thurzo, "Provable AI ethics and explainability in medical and educational AI agents: Trustworthy ethical firewall," Electronics, vol. 14, no. 7, 2025.
[12] Z. Yang et al., "Plug in the safety chip: Enforcing constraints for LLM-driven robot agents," in Proc. IEEE Int. Conf. Robot. Autom. (ICRA), 2024, pp. 14435-14442.
[13] M. Shamsujjoha et al., "Swiss cheese model for AI safety: A taxonomy and reference architecture for multi-layered guardrails of foundation model based agents," in Proc. 22nd IEEE Int. Conf. Software Archit. (ICSA), 2025, pp. 37-48.
[14] T. Sadhu, A. Pesaranghader, Y. Chen, and D. H. Yi, "ATHENA: Safe autonomous agents with verbal contrastive learning," in Proc. Conf. Empir. Methods Nat. Lang. Process. (EMNLP), Industry Track, 2024, pp. 1121-1130.
[15] Y. Zhou et al., "Defending jailbreak prompts via in-context adversarial game," in Proc. Conf. Empir. Methods Nat. Lang. Process. (EMNLP), 2024, pp. 20084-20105.
[16] T. Liu et al., "Demystifying RCE vulnerabilities in LLM-integrated apps," in Proc. ACM SIGSAC Conf. Comput. Commun. Secur. (CCS), 2024, pp. 1716-1730.
[17] R. Pedro et al., "Prompt-to-SQL injections in LLM-integrated web applications: Risks and defenses," in Proc. 47th IEEE/ACM Int. Conf. Software Eng. (ICSE), 2025, pp. 1768-1780.
[18] P. Kumar et al., "Aligned LLMs are not aligned browser agents," in Proc. Int. Conf. Learn. Represent. (ICLR), 2025. [Online]. Available: https://openreview.net/forum?id=NsFZZU9gvk
[19] E. Debenedetti et al., "AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2024. [Online]. Available: https://arxiv.org/abs/2406.13352.
[20] Q. Zhan et al., "INJECAGENT: Benchmarking indirect prompt injections in tool-integrated large language model agents," in Proc. Annu. Meet. Assoc. Comput. Linguist. (ACL), 2024, pp. 10471-10506.
[21] M. Andriushchenko et al., "AgentHarm: A benchmark for measuring harmfulness of LLM agents," in Proc. Int. Conf. Learn. Represent. (ICLR), 2025. [Online]. Available: https://arxiv.org/abs/2410.09024
[22] T. Yuan et al., "R-Judge: Benchmarking safety risk awareness for LLM agents," in Proc. Conf. Empir. Methods Nat. Lang. Process. (EMNLP), Findings, 2024, pp. 1467-1490.
[23] H. Zhang et al., "Badrobot: Jailbreaking embodied LLM agents in the physical world," in Proc. Int. Conf. Learn. Represent. (ICLR), 2025. [Online]. Available: https://openreview.net/forum?id=ei3qCntB66
[24] Y. Wang et al., "BadAgent: Inserting and activating backdoor attacks in LLM agents," in Proc. Annu. Meet. Assoc. Comput. Linguist. (ACL), 2024, pp. 9811-9827.
[25] S. Omri, M. Abdelkader, and M. Hamdi, "SafetyRAG: Towards safe large language model-based application through retrieval-augmented generation," J. Adv. Inf. Technol., vol. 16, no. 2, pp. 243-250, 2025.
[26] Z. Chen et al., "AgentPoison: Red-teaming LLM agents via poisoning memory or knowledge bases," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2024. [Online]. Available: https://arxiv.org/abs/2407.12784
[27] Y. He et al., "Security of AI agents," in Proc. IEEE/ACM Int. Workshop Responsible AI Eng. (RAISE), 2025, pp. 45-52.
[28] M. Chernyshev, Z. Baig, and R. Doss, "Forensic analysis of indirect prompt injection attacks on LLM agents," in Proc. IEEE 6th Int. Conf. Trust, Privacy Secur. Intell. Syst. Appl. (TPS-ISA), 2024, pp. 409-411.
[29] S. R. Motwani et al., "Secret collusion among AI agents: Multi-agent deception via steganography," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2024. [Online]. Available: https://arxiv.org/abs/2402.07510
[30] J. Owotogbe, "Assessing and enhancing the robustness of LLM-based multi-agent systems through chaos engineering," in Proc. 4th IEEE/ACM Int. Conf. AI Eng. (CAIN), 2025, pp. 250-252.
[31] X. Gu et al., "Agent Smith: A single image can jailbreak one million multimodal LLM agents exponentially fast," in Proc. 41st Int. Conf. Mach. Learn. (ICML), 2024. [Online]. Available: https://arxiv.org/abs/2402.08567
[32] J. H. Park and V. K. Madisetti, "CAPRI: A context-aware privacy framework for multi-agent generative AI applications," IEEE Access, vol. 13, pp. 43168-43177, 2025.
[33] F. Ahmed, "Cybersecurity policy frameworks for AI in government: Balancing national security and privacy concerns," Int. J. Multidiscip. Sci. Manag., vol. 1, no. 4, pp. 43-53, 2024.
[34] A. D. Householder et al., "Lessons learned in coordinated disclosure for artificial intelligence and machine learning systems," Aug. 2024.
[35] A. Liu et al., "Compromising LLM driven embodied agents with contextual backdoor attacks," IEEE Trans. Inf. Forensics Secur., vol. 20, pp. 3979-3994, 2025.
[36] Z. Wang et al., "To protect the LLM agent against the prompt injection attack with polymorphic prompt," in Proc. 55th Annu. IEEE/IFIP Int. Conf. Dependable Syst. Netw. Suppl. Vol. (DSN-S), 2025, pp. 22-28.
[37] X. Zhang et al., "PrivacyAsst: Safeguarding user privacy in tool-using large language model agents," IEEE Trans. Dependable Secure Comput., vol. 21, no. 6, pp. 5242-5258, 2024.
[38] B. Chen et al., "BlockAgents: Towards byzantine-robust LLM-based multi-agent coordination via blockchain," in Proc. ACM Int. Conf. Proc. Ser., 2024, pp. 187-192.