Security of Internet of Agents: Attacks and Countermeasures

Yuntao Wang, Yanghe Pan, Shaolong Guo, and Zhou Su

Abstract—With the rise of large language and vision-language models, AI agents have evolved into autonomous, interactive systems capable of perception, reasoning, and decision-making. As they proliferate across virtual and physical domains, the Internet of Agents (IoA) has emerged as a key infrastructure for enabling scalable and secure coordination among heterogeneous agents. This survey offers a comprehensive examination of the security and privacy landscape in IoA systems. We begin by outlining the IoA architecture and its distinct vulnerabilities compared to traditional networks, focusing on four critical aspects: identity authentication threats, cross-agent trust issues, embodied security, and privacy risks. We then review existing and emerging defense mechanisms and highlight persistent challenges. Finally, we identify open research directions to advance the development of resilient and privacy-preserving IoA ecosystems.

Index Terms—Internet of agents (IoA), AI agents, large models, security, and privacy.

I. INTRODUCTION

The advent of large language models (LLMs) and vision-language models has transformed AI agents into fully autonomous and interactive entities capable of independent perception, reasoning, and action [1]. These AI agents (or called agentic AI) [2], ranging from digital assistants to unmanned aerial vehicles (UAVs) and service robots, operate across virtual and physical domains, driving unprecedented demands for an infrastructure that natively supports agent-to-agent (A2A) interactions. Gartner projects that [3], by 2028, AI agents will autonomously manage at least 15 % of routine daily tasks, while roughly one-third of enterprise applications will embed agent-based intelligence. The Internet of agents (IoA) [4], [5], also referred to as the agentic web, has emerged to meet this need, offering an agent-centric fabric that supports on-the-fly agent discovery, goal-driven communication, and coordinated task execution at scale. Unlike the traditional Internet, IoA communications focus on machine-readable objects (e.g., model checkpoints, encrypted tokens, and latent embeddings), and agent protocols emphasize semantic negotiation and adaptive orchestration. By pooling distributed inference and shared sensing capacities, IoA extends advanced AI capabilities to resource-constrained devices and establishes new connectivity patterns across heterogeneous agent ecosystems.

As agents process and exchange large volumes of personal and sensitive data, ranging from user profiles and behavioral histories to real-time sensor feeds, they become prime targets for sophisticated cyber adversaries. Malicious actors may exploit agent forgery to impersonate legitimate agents and infiltrate sensitive workflows [6], or employ intent deception to subtly manipulate decision-making logic and contaminate collaborative outcomes [7]. Colluding agents can coordinate to distort shared insights or hijack consensus mechanisms [8], undermining the integrity of distributed reasoning processes. Meanwhile, adversarial inputs crafted to trigger misclassification [9], contextual backdoors that activate under specific environmental cues [10], and hallucination cascades that propagate spurious outputs across agent networks [11] can cause systemic breakdowns in multi-agent coordination. With the ongoing evolution of IoA, innovative built-in security and privacy preservation mechanisms are essential to realize trustworthy, secure, and privacy-preserving deployments of large-scale IoA ecosystems and unleash the transformative power of future autonomous AI ecosystems.

Recent research in LLM-based agents has attracted considerable attention across both academia and industry. Das *et al.* [12] provide an in-depth survey of security and privacy vulnerabilities in LLMs, evaluating domain-specific risks and defenses across transportation, education, and healthcare. He *et al.* [13] categorize emerging threats to LLM-driven agents, illustrates their real-world impacts, and reviews prevailing mitigation techniques. Wang *et al.* [14] examine attack vectors at five critical stages: pre-training, fine-tuning, retrieval-augmented generation (RAG), deployment, and in-agent operation, as well as tailored countermeasures. Gan *et al.* [15] introduce a two-axis taxonomy of security challenges in LLM agents by threat source and impact, and analyze representative agent implementations as case studies. Li *et al.* [16] outline the architecture and optimization workflows of personal LLM agents, highlighting associated security and privacy concerns. Zhang *et al. et al.* [17] focus on safety and security issues in agent systems by examining physical faults and cyber attacks including denial-of-service (DoS) and deception attacks. They also discuss countermeasures including fault estimation, detection, diagnosis, fault-tolerant control, and secure cyber attack management. Deng *et al.* [18] identify four key vulnerability domains in software-form AI agents: complex multi-step inputs, opaque internal executions, environmental variability, and untrusted external interactions. Wang *et al.* [19] systematically survey the full-stack safety threats for LLMs and agents by considering life-cycle LLM risks during LLM training, deployment, and commercialization. Existing surveys mainly focus on security and privacy threats and defense at the single LLM agent level. In contrast, this survey investigates the networking aspects of large model agents within the Internet of agents (IoA), exposing its unique threat landscape, defense

Y. Wang, Y. Pan, S. Guo, and Z. Su are with the School of Cyber Science and Engineering, Xi'an Jiaotong University, Xi'an, China. *(Corresponding author: Zhou Su)*

TABLE I
A COMPARISON OF OUR SURVEY WITH RELEVANT SURVEYS

Year	Ref.	Contribution
2021	[17]	Survey on cyber-physical threats and security in agent systems, covering physical faults, DoS, deception attacks, and defenses.
2024	[13]	Categorize emerging threats to LLM-driven agents, illustrate real-world impacts, and review mitigation techniques.
2024	[14]	Examine attack vectors across five stages (pre-training, fine-tuning, RAG, deployment, in-agent operation) and review tailored defenses.
2024	[15]	Introduce a two-axis taxonomy of LLM-agent security challenges by threat source and impact with case-study analyses.
2024	[16]	Outline architectures and optimization workflows of personal LLM agents, highlight security and privacy concerns.
2025	[12]	Provide an in-depth survey of security and privacy vulnerabilities in LLMs, assess domain-specific risks and defenses in transportation, education, and healthcare.
2025	[18]	Identify four key vulnerability types of agents: multi-step inputs, opaque execution, environmental variability, untrusted interactions.
2025	[19]	Survey full-stack safety of LLMs and agents from LLM training, deployment, and commercialization.
Now	**Ours**	Comprehensive survey of security/privacy threats in IoA, emerging/potential countermeasures, and open research challenges.

TABLE II
SUMMARY OF KEY ABBREVIATIONS IN ALPHABETICAL ORDER

Abbr.	Definition	Abbr.	Definition
A2A	Agent-to-Agent	AI	Artificial Intelligence
ANP	Agent Network Protocol	BLOS	Beyond-Line-of-Sight
DID	Decentralized IDentifier	DoS	Denial-of-Service
EMI	ElectroMagnetic Interference	IoA	Internet of Agents
IMU	Inertial Measurement Unit	LLM	Large Language Model
MEMS	Micro-Electro Mechanical System	MCP	Model Context Protocol
mmWave	millimeter-Wave	P2P	Peer-to-Peer
PII	Personally Identifiable Information	RL	Reinforcement Learning
RAG	Retrieval-Augmented Generation	UAV	Unmanned Aerial Vehicle

techniques, and research gaps. Table I summarizes our survey's contributions with previous survey efforts.

This paper provides a systematic review of security and privacy in the IoA, charting its threat landscape, defense mechanisms, and research challenges to support large-scale deployment. Our objectives are twofold: (1) assess the landscape of security and privacy risks inherent to agent-centric IoA systems, including the scope and impact of security/privacy challenges; and (2) propose effective strategies and solutions to mitigate these threats, ultimately guiding the robust and secure deployment of IoA in various intelligent applications. Our key contributions include:

- *Taxonomy of IoA Threats and Defenses.* We offer an in-depth review of the emerging security and privacy vulnerabilities in IoA across four aspects: agent identity authentication, cross-agent trust, embodied agent security, and privacy threats. For each category, we survey state-of-the-art and potential countermeasures and practical challenges in IoA.
- *Open Challenges in IoA.* We highlight critical gaps in current IoA research and outline a research roadmap to guide the development of resilient, trustworthy, privacy-preserving, and ethical IoA ecosystems.

A. Paper Organization

The remainder of this paper is organized as follows. In Section II, we provide an overview of IoA and its security and privacy landscape. Next, we investigate the security and privacy aspects, including agent identity authentication in Section III, cross-agent trust issues in Section IV, embodied agent security in Section V, and privacy threats in Section VI. Lastly, Section VIII outlines future research trends in IoA domain. Table II summarizes key acronyms used in this survey.

II. OVERVIEW OF INTERNET OF AGENTS AND ITS SECURITY AND PRIVACY LANDSCAPE

In this section, we first introduce the key concept of the IoA with its distinctive paradigm and overview representative agent communication protocols. We then discuss the security and privacy landscape in IoA, highlighting how traditional

network risks evolve in agentic environments and identifying novel threat vectors.

A. Overview of IoA

The IoA is an emerging infrastructure in which autonomous software and embodied agents, ranging from personal LLM assistants to industrial robots, seamlessly interconnect, discover one another, and collaborate to accomplish complex tasks [4], [5]. Unlike prior Web generations centered on human navigation, the IoA is agent-centric:

- *Agents as the new entry point.* Rather than humans directly navigating the Internet via personal computers in Web 1.0 or mobile devices in Web 2.0, autonomous agents in IoA navigate and interact with the digital world. A specialized *super personal assistant agent* will act on behalf of its human owner, negotiating on the user's behalf through personalized UIs while interacting with other agents via APIs or protocols on the backend [20]. Simultaneously, numerous non-user-facing agents, representing banks, schools, restaurants, and so on, will interact indirectly with personal assistants to deliver tailored services.
- *Flat & self-organizing agent collaboration networks.* By autonomously organizing and negotiating, all agents regardless of corporate or platform affiliations can establish efficient, task-driven collaboration networks, to dynamically allocate resources and expertise, via self-organization and self-negotiation. Ultimately, the agentic web will evolve into a flatter, more decentralized digital ecosystem [4].
- *Shared intelligence and capability sharing.* Beyond basic connectivity, the IoA enables agents to share inference workloads and sensing data at scale. Resource-constrained agents can offload inference tasks, access high-end models, and leverage collective knowledge, achieving on-demand large-model-as-a-service (LMaaS). Besides, shared sensing capabilities in IoA empowers agents particularly embodied ones with beyond-line-of-sight (BLoS) perception.

B. Representative Agent Communication Protocols

Standardized agent communication protocols are key enablers for the IoA to facilitate seamless interaction, coordination, and secure interoperability among heterogeneous agents [21]–[24]. Recently, the following representative protocols are

developed to support structured messaging, identity verification, and tool integration within distributed agent ecosystems.

1) Model Context Protocol (MCP) [21]. Anthropic's MCP provides a modular interface standard to enable real-time interactions between large models and external tools, services, or data sources. By abstracting tool execution and contextual data access behind a unified protocol, MCP decouples model reasoning from backend functionalities, thereby facilitating cross-platform interoperability and flexible agent design. Its client-server architecture supports capability (e.g., tools) discovery, authenticated invocation, and streamlined response handling, enabling agents to operate with enhanced contextual awareness through OAuth authorization.

2) Agent-to-Agent (A2A) Protocol [22]. Google's A2A protocol establishes a standardized communication framework for decentralized collaboration among AI agents. It supports structured discovery through agent cards (i.e., JSON file hosted at a URL) and ensures secure exchanges using authentication frameworks such as OAuth 2.0 and OpenID Connect. A2A accommodates both synchronous and asynchronous messaging via HTTP and server-sent events, allowing agents to interact with reliable task tracking, progressive response streaming, and flexible follow-up exchanges.

3) Agent Network Protocol (ANP) [23]. ANP defines a decentralized peer-to-peer (P2P) protocol centered on agent autonomy and security. Agents are identified using W3C decentralized identifiers (DIDs), and agent communications are protected via end-to-end encryption. Protocol negotiation is adaptive, allowing agents to dynamically align communication strategies based on task context and peer capabilities.

4) Agora Protocol [24]. The Agora protocol focuses on scalable and adaptable communication in IoA. It leverages structured routines for high-frequency interactions and harnesses natural language interfaces (potentially generated by LLMs) for dynamic ad-hoc coordination, striking a balance between formal protocol efficiency and semantic flexibility.

C. Key Characteristics of IoA Security Landscape

While the IoA inherits traditional Internet vulnerabilities (e.g., spoofing, eavesdropping, and DoS), it also gives rise to new risks stemming from its unique characteristics, including large model foundations, decentralization, task-driven cooperation, semantic-aware interaction, and coupled cyber-physical effects.

- *Large-model foundations.* Both virtual and embodied agents powered by pretrained large models (e.g., LLMs) may inherit vulnerabilities including backdoors, data leakage, and algorithmic bias, and these risks escalate with the deployment of such agents at scale.
- *Decentralization.* In decentralized IoA environments, threats such as identity spoofing, Sybil attacks, and rogue agent infiltration are amplified, undermining the trust and consensus protocols.
- *Task-driven cooperation.* In IoA, agents involved in a common task autonomously negotiate task workflows and exchange intermediate state. However, malicious peers can inject faulty tasks or intercept sensitive context, turning collaboration into a vector for targeted disruption.

Fig. 1. The taxonomy of security and privacy threats in IoA.

- *Semantic-aware interaction.* The use of natural language protocols and LLM-mediated communication introduces new risks of hallucination, prompt injection, and semantic misinterpretation, where attackers exploit semantic ambiguity to bypass system safeguards.
- *Cyber-physical coupling.* When agents control physical systems such as robots, UAVs, or smart infrastructure, cyber threats may lead to real-world harm through manipulated sensor inputs, malicious actuator commands, or compromised safety routines, blurring the line between digital and physical attack surfaces.

In the following, we investigate the security and privacy threats, countermeasures, and challenges in IoA from four perspectives: agent identity authentication (in Sect. III), cross-agent trust issues (in Sect. IV), embodied agent security (in Sect. V), and privacy threats (in Sect. VI). Fig. 1 illustrates the taxonomy of security and privacy threats in IoA.

III. AGENT IDENTITY AUTHENTICATION THREATS IN IOA

In IoA, agents often process substantial volumes of sensitive user/commercial data, including local knowledge, historical preferences, and proprietary product information. Due to the decentralized and dynamic nature of IoA environments, effective authentication between agents is a critical prerequisite for dynamically preventing unauthorized access and malicious interactions, thereby protecting sensitive assets and enable secure collaboration across heterogeneous agent networks.

Fig. 2. Illustration of agent identity authentication threats in IoA: (a) identity forgery, (b) Sybil attack, and (c) intent deception.

1) Threats: As depicted in Fig. 2, identity authentication in IoA faces the following security threats.

- *Identity Forgery:* An agent may falsely report its capabilities to join a team and participate in collaborative tasks. Moreover, as shown in Fig. 2(a), leveraging AI-generated contents, adversaries can forge the identity of a human owner to maliciously bypass authentication mechanisms, thereby gaining unauthorized access. For instance, Jiang *et al.* propose *Foice*, a novel cross-modal attack that generates synthetic speech mimicking a target user's voice from a single facial image [6]. In addition, agents can craft adversarial audio samples (perceived as noise by humans but correctly recognized by cooperative agents as the owner's voice) to further enhance the attack stealthiness, particularly in environments involving human-agent interactions.

- *Impersonation Attacks:* An adversary may impersonate another agent, such as a coordinator, to inject false messages, issue malicious commands, or manipulate task allocation during collaborative tasks, as shown in Fig. 2(b). For instance, adversaries could register a malicious server with a name closely resembling that of a legitimate tool in MCP (e.g., mcp-github instead of github-mcp) [25]. Due to the lack of strict namespace enforcement and robust authentication mechanisms, agents might inadvertently invoke the malicious server, leading to unauthorized command execution or sensitive data leakage.

- *Sybil Attacks:* An adversary can dynamically create a large number of Sybil virtual agents within short time to form a majority, manipulate group decision-making, or overwhelm verification mechanisms [26]. For instance, in an A2A protocol-based distributed decision system, an adversary could generate thousands of Sybil agents with forged agent cards to register on the coordinator server. These Sybil agents could then flood the voting process with manipulated ballots, artificially dominating the majority and reversing legitimate decisions.

- *Privilege Escalation:* An adversary may exploit vulnerabilities or logic flaws within IoA systems, allowing malicious agents to escalate access privileges beyond their authorized scope. For instance, the tool poisoning attack in MCP [27] embeds hidden instructions within seemingly benign tool descriptions, thereby manipulating agents to perform unintended actions such as accessing restricted files or executing unauthorized commands. Such attacks can potentially disrupt the reliability of IoA.

- *Intent Deception:* As shown in Fig. 2(c), an adversary may deploy malicious agents to deceive authentication systems, by initially claiming legitimate objectives to gain access (e.g., data query to a government agent). Once access is granted, the malicious agent may then engage in unauthorized activities, such as probing for sensitive information. For instance, Hao *et al.* propose CDA, a covert deception attack in which a malicious robotic agent impersonates a cooperative teammate while secretly observing the motion patterns of other agents [7], thereby leaking sensitive information such as trajectories and behaviors of other agents. By leveraging an LSTM-based model, the attacker predicts congestion areas and generates self-serving paths to save resources while evading detection.

2) Defenses: To mitigate identity authentication threats in IoA, *access control* mechanisms [28], e.g., role-based, attribute-based, and policy-based access control, are crucial to prevent unauthorized access. Given the dynamic nature and autonomy of IoA agents, access control should not be static or uniform. Instead, it should adapt to the capabilities, behaviors, and interactions of individual agents. By enforcing fine-grained and context-aware policies, IoA systems can dynamically and intelligently restrict agents' access actions, even in cases of identity forgery or misuse. Besides, DIDs combined with verifiable credentials and blockchain-based registries can offer tamper-resistant identity management for agents [29].

3) Challenges: IoA faces a series of unique challenges in terms of identity authentication, as follows:

- *Task-Driven Dynamic Access Control:* In IoA environments, agents frequently change roles and responsibilities as tasks evolve or under different tasks, necessitating real-time adjustments to access control policies. Static authentication and authorization models are insufficient

to accommodate such dynamic shifts. Instead, identity authentication mechanisms need to continuously adapt based on agent's current capabilities, assigned tasks, and operational context. For instance, an agent initially tasked with environmental monitoring (requiring access only to non-sensitive sensor data) may later be reassigned to mission-critical operations involving confidential mapping or surveillance information. In this case, access control rules associated with the agent should be promptly updated to reflect its new privileges and ensure renewed identity verification, thereby minimizing the risk of unauthorized access or privilege misuse.

- *Context-Aware Continuous Authentication:* Agents in IoA often engage in long-term tasks over extended periods (e.g., crowd monitoring), making continuous authentication critical rather than relying solely on a one-time initial verification. To ensure ongoing trust, contextual factors such as behavioral patterns, interaction histories, and task progression should be continuously monitored. Abrupt deviations from established patterns may signal deceptive behavior, particularly when an agent transitions from low-sensitivity to high-sensitivity tasks. For instance, in an intent deception attack scenario, a foreign adversarial agent may initially perform legitimate data queries to a government agent, posing as a benign collaborator. However, as the interaction progresses, the agent could gradually shift its behavior to probe for sensitive or classified information. Context-aware continuous authentication mechanisms are therefore essential to promptly detect and mitigate such evolving threats.

IV. CROSS-AGENT TRUST ISSUES IN IOA

In IoA, effective collaboration depends on mutual trust to orchestrate distributed tasks and exchange critical information. However, agents' divergent objectives, unpredictable hallucinations, and covert collusion can undermine this trust, leading to task failures, data-integrity violations, and degraded performance. The autonomous and dynamic nature of IoA interactions further amplifies these risks, as agents may opportunistically withhold resources or deliberately mislead peers to pursue their own goals. Consequently, dynamic trust-management frameworks are essential for sustaining reliable collective outcomes and resilient multi-agent coordination across heterogeneous agent networks.

1) Threats: In the IoA context, inter-agent collaboration faces the following trust threats.

- *Hallucination Cascade:* Large models such as LLM can produce inaccurate or inconsistent outputs that deviate from the input context or factual information. For instance, when coordinating tasks via the MCP, an LLM-based agent might hallucinate a non-existent data source or misinterpret the capabilities of another agent, leading to flawed decisions. As shown in Fig. 3(a), these initial errors can then propagate and amplify through subsequent agent interactions (referred to as *hallucination cascade*), thereby undermining the reliability of decision-making in IoA. Zhang *et al.* demonstrate that early-stage hallucinations in LLMs can compound over time, with initial mistakes significantly degrading output accuracy in later stages [11]. Similarly, in IoA task collaboration, hallucination-induced errors made by one agent can cascade through the network, compromising downstream agents' outputs and ultimately degrading overall task performance.

- *Knowledge Poisoning:* Adversaries can undermine the integrity of shared knowledge bases in cooperative IoA tasks by stealthily injecting false, biased, or malicious information through compromised agents, as shown in Fig. 3(b). This knowledge poisoning threat can degrade task quality or facilitates attacker's output manipulation. Zou *et al.* reveal a new attack named PoisonedRAG targeting external knowledge bases in IoA [30]. In PoisonedRAG, malicious agents inject small amount of malicious knowledge into a shared knowledge repository, thereby steering downstream agents to generate adversary-desired results and subvert the collaborative decision-making process.

- *Adversarial Attack:* An adversary may manipulate the output of a preceding agent within the collaborative task workflow to craft adversarial examples, which are then fed into subsequent target agents. As a result, the affected agents may produce false or biased outputs, ultimately disrupting the reliability of the entire process. For instance, Khan *et al.* propose a permutation-invariant attack that optimizes adversarial prompt propagation across latency- and bandwidth-constrained agent network topologies [9]. By formulating the propagation as a maximum-flow minimum-cost problem and employing a novel permutation-invariant evasion loss, the attack in [9] successfully evades distributed security defenses such as Llama-Guard.

- *Jailbreak:* Adversaries may attempt to bypass LLM agents' built-in security and ethical restrictions by crafting specialized prompts, causing agents in IoA to generate outputs that violate their intended guidelines. Chen *et al.* propose Pandora, a novel jailbreak approach through multiple phishing agents, which decomposes a malicious prompt into multiple stealthier sub-queries and leverages the LLM's multi-step reasoning to evade detection [31], as shown in Fig. 3(c). Within IoA systems, the impact of jailbreak is amplified, as compromised agents can autonomously propagate harmful behaviors, amplify misaligned responses, and expand the overall attack surface.

- *Prompt Injection:* Adversaries may inject malicious instructions within crafted prompts, causing the agent to generate outputs or take actions that deviate from its intention. Zhang *et al.* propose Breaking Agents, a prompt-injection framework that triggers logical errors and repetitive malfunction loops in autonomous LLM agents [32]. Their method targets the inherent instability of agents by misleading them into executing incorrect or infinite-loop actions, even without obvious policy violations.

- *Free-Riding Attack:* In cooperative IoA tasks, a selfish agent may deliberately withhold effort or provide low-quality, incomplete, or even misleading results while still

Fig. 3. Illustration of cross-agent trust issues within IoA: (a) hallucination cascade, (b) knowledge poisoning, and (c) jailbreak.

reaping the benefits of participation. Such free-riding attack in task cooperation would degrade overall task performance and undermines fairness across the agent network.

- *Agent Collusion:* A group of compromised or malicious agents may collude to manipulate task outcomes, fabricate consensus, and bias collective decisions, thereby undermining the fairness and trustworthiness of multi-agent collaboration in IoA. Motwani *et al.* formalize this *multi-agent secret collusion* with a detailed model, where AI agents use steganographic techniques to covertly communicate or coordinate their actions while evading detection [8]. They also provide both theoretical and empirical evidence that agents are capable of engaging in such covert collusion behavior.

2) Defenses: To address trust issues in agent cooperation, IoA frameworks can deploy *agent audit* mechanisms [33] to verify peer agents' outputs and filter biased information. By constraining information flow or introducing parallel validation paths, *network topology defense* mechanisms [34] can limit the influence of individual agents and prevent misinformation cascades. Furthermore, *trust management* approaches play a crucial role in maintaining long-term collaboration [35], while reinforcement learning (RL) techniques and game-theoretic models can be utilized to adaptively adjust trust scores and to design incentive mechanisms, thereby promoting fair and robust agent cooperation in IoA.

To counter hallucination, *RAG* grounds outputs with external knowledge sources to enhance factual consistency [36]. Furthermore, *multi-agent review* processes facilitate collaborative outcome evaluation among agents [37], while *post-correction* techniques refine outputs and resolve inconsistencies [38]. To mitigate jailbreak threats, *filtering-based defenses* employ auxiliary models to detect and filter out potentially harmful or malicious content [39]. Additionally, *multi-agent debate* mechanisms enhance robustness through iterative self-evaluation and cross-verification among agents [40]. To defend against prompt injection, defense strategies can be broadly categorized into *prevention-based* and *detection-based* methods. The former focuses on breaking or disrupting malicious prompts before execution [41], while the latter focuses on analyzing model behavior and input-output patterns to identify anomalous or adversarial prompts [42].

3) Challenges: The design of trustworthy IoA systems faces several intertwined challenges, as below.

- *Threat Cascade:* In collaborative agent workflows, the output of agents may become corrupted by hallucinations or adversarial perturbations, and subsequently serve as inputs for downstream agents. This propagation of manipulated information produces a cascading effect, in which false or malicious outputs are amplified throughout the agent cooperation chain. Over time, the accumulation of misleading data can severely degrade IoA task performance, and compromise the trustworthiness of the entire collaborative process.
- *Full-Process Poisoning:* Beyond isolated poisoning attacks, full-process poisoning refers to the persistent and strategic injection of manipulated knowledge throughout the agent collaboration workflow. Biased, false, or misleading information may be injected at multiple stages of agent collaboration, progressively corrupting the shared knowledge base, undermining decision accuracy and operational integrity.

V. EMBODIED SECURITY IN IOA

Distinguished from purely virtual threats, embodied agents are vulnerable to physical tampering, sensor spoofing, mechanical sabotage, supply-chain attacks, and environmental hazards, any of which can disrupt their operation or corrupt collected data. In IoA environments, embodied security focuses on safeguarding agents' physical safety and their interactions with the virtual/real world under cyber-physical coupled effects.

1) Threats: Typical embodied threats in the IoA context include the following types.

- *Attacks on Agent Sensors:* Adversaries can exploit external signals (e.g., acoustic, electromagnetic, and electrical) to corrupt onboard sensor readings of embodied agents (e.g., UAVs, autonomous vehicles), jeopardizing their safety. ① *Gyroscope resonance:* High-frequency acoustic waves can induce resonant vibrations in micro-electro mechanical system (MEMS) gyroscopes, causing UAV disorientation and crashes. Son *et al.* [43] demonstrate that targeted high-frequency noise can disrupt 15 commercial MEMS gyros of autonomous agents. Hong *et al.* [44] further embed covert acoustic signals within audio files to stealthily manipulate vehicle stability. ② *Millimeter-Wave (mmWave) signal manipulation:* Chen *et al.* propose an attack named MetaWave [45], which distort millimeter-wave sensor readings and mislead radar-based perception by attaching metamaterial-enhanced tags. ③ *Inertial measurement unit (IMU) interference:* MEMS-based IMUs are vulnerable to electromagnetic injection. Jang *et al.* [46] inject electromagnetic interference (EMI) intocommunications between IMU and control unit, causing UAVs to veer off course and crash. ④ *Vision sensor blinding:* Fu *et al.* [47] show that focused laser pulses can blind UAV cameras and stereo vision systems, causing failures in obstacle avoidance, target recognition, and tracking. ⑤ *Radar & ultrasonic jamming:* Long-range radar sensors are vulnerable to jamming or spoofing via noise signals that mask true echoes, while short-range ultrasonic sensors are prone to signal interference, blockage, or replay attacks. In IoA, compromised agents themselves can serve as covert attack platforms, leveraging large model intelligence to optimize attack configurations and minimize costs.

- *Contextual Backdoor:* Adversaries may exploit the poisoned contextual inputs within the underlying LLM to embed hidden triggers that activate only under certain conditions, such as when a specific image is viewed or a particular word is read [10], [48]. As shown in Fig. 4(a), these malicious inputs lead the embodied agent to execute actions that generally appear normal but can become harmful (e.g., engage in unsafe, unintended, or malicious behaviors) once the contextual backdoor is triggered. For instance, an autonomous vehicle may accelerate toward obstacles upon detecting a particular roadside object (e.g., a gray trash bin), despite appearing to function normally otherwise [10].

- *Cross-domain Safety Misalignment:* An embodied agent may exhibit safety misalignment between its linguistic responses and action outputs, which stems from the agent's incomplete understanding of its physical embodiment. As shown in Fig. 4(b), while the embodied agent properly refuses harmful requests in natural language, it may still generate corresponding action plans in structured formats, causing it to produce seemingly valid but potentially dangerous robotic commands. For instance, Zhang *et al.* demonstrate that an agent can refuse a harmful request in text, e.g., "Grasp the knife to attack the person", yet simultaneously generate executable, dangerous action code in a structured format [49].

(a) Contextual Backdoor

(b) Cross-domain Safety Misalignment

Fig. 4. Illustration of security threats to embodied agents in IoA: (a) contextual backdoor, and (b) cross-domain safety misalignment.

2) Defenses: Mitigating embodied security threats in IoA requires a holistic strategy across hardware, software, and behavioral layers in dynamic and potentially adversarial settings. Attacks on agent sensors can be countered through through *physical defense, information redundancy, and data fusion*. Physical defense such as shielding or signal isolation helps prevent direct interference with sensor hardware. Information redundancy achieved by deploying multiple sensors measuring similar phenomena allows cross-validation to detect anomalies. Data fusion combines inputs from diverse sensor sources to construct a coherent and resilient representation of the environment, reducing the influence of any single compromised signal. *World models* enable agents to simulate their action outcomes, helping identify unsafe behaviors before action execution [50]. *Multimodal consistency validation* assesses the alignment between language and action outputs via semantic similarity, acting as a firewall against contextual triggers [49]. *Adversarial fine-tuning* can effectively enhance robustness of the underlying LLM of embodied agents by fine-tuning the LLM on backdoor-triggered inputs with corrected outputs [51].

3) Challenges: Securing embodied agents in IoA presents unique challenges due to the tight coupling of cyber and physical domains. Cyber-layer attacks, such as contextual backdoor attacks or jailbreak prompts, can directly lead to unsafe physical actions and potentially cause real-world harm. Conversely, changes in the physical environment, such as weather conditions, may serve as triggers that inadvertently induce these attacks on the embodied agents. For instance, an agent may behave normally in clear conditions but, upon detecting rain, activate a rain-bound contextual backdoor and execute malicious behaviors. This highly concealed vulnerability significantly amplifies the attack surface, necessitating novel cyber-physical defense mechanisms in dynamic IoA ecosystems.

VI. Privacy Threats in IoA

In IoA, agents continuously collect, process, and share vast quantities of sensitive personal and commercial data, including individual preferences, location traces, and proprietary business information. The decentralized, dynamic, and open architecture of agent networks, along with pervasive multiparty interactions, exposes IoA systems to a broad spectrum of privacy risks. Such threats can compromise user confidentiality, violate privacy regulations, and erode stakeholder trust, ultimately hindering the adoption of IoA applications.

1) Threats: The privacy threats in IoA include contextual privacy inference, RAG privacy leakage, and agent memorization risks.

- *Contextual Privacy Inference:* Adversaries can exploit intermediate contextual data exchanged such as agents' inputs and outputs or metadata during multi-agent collaboration to perform correlation analysis and statistical inference [52]. As such, the sensitive attribute such as user identity, location, and preferences can be reconstructed even if they were not explicitly disclosed. For instance, the phrase "waiting for a hook turn during my commute" during an user-agent conversational interaction can be analyzed by AI agents to infer their location as Melbourne by associating the phrase with the city's specific traffic rules.

- *RAG Privacy Leakage:* An RAG agent connected to long-term memory via RAG mechanisms may potentially expose knowledge-related sensitive information during interactions, as shown in Fig. 5(a). Adversaries can leverage jailbreak prompts to extract private data through repeated and strategically crafted queries. Furthermore, embedding inversion techniques [53] enable the reconstruction of original inputs from stored embeddings, posing significant privacy risks in vector-based memory systems. For instance, RAG-Thief [54] demonstrates an automated agent-based attack that recovers over 70% of private knowledge base chunks by iteratively refining adversarial queries through self-improvement mechanisms.

- *Agent Memorization:* Agents fine-tuned on sensitive or poorly sanitized data can memorize private information during training and disclose it during subsequent interactions, as shown in Fig. 5(b). Meanwhile, through in-context learning, an agent can implicitly retain and reproduce sensitive content obtained during previous interactions. These behaviors increase the risk of unintended disclosure of personal identifiers, private conversations, or confidential user inputs. Carlini *et al.* [55] show that querying LLM agents with carefully crafted prefix patterns can effectively extract users' personally identifiable information (PII), including phone numbers, email addresses, and other sensitive data.

2) Defenses: Existing defenses of IoA privacy threats involves two complementary strategies: *privacy pre-assessment* and *output intervention*. *Privacy pre-assessment* mechanisms [56] focus on identifying whether an agent is likely to leak sensitive information from its training data or external sources before deployment through simulated querying and informa-

(a) RAG Privacy Leakage

(b) Agent Memorization

Fig. 5. Illustration of privacy threats in IoA: (a) RAG privacy leakage, and (b) agent memorization.

tion leakage analysis, providing early signals for risk evaluation and informing downstream privacy-preserving strategies. Conversely, *output intervention* mechanisms [57], [58] monitor the agent's responses during runtime and intercept outputs containing sensitive content. If outputs are found to contain sensitive or private information, intervention mechanisms (e.g., filtering or redaction) are triggered to suppress or revise the outputs before delivery.

3) Challenges: In cooperative IoA scenarios, continuous multi-turn interactions amplify privacy risks in two manner. First, agents routinely exchange detailed user-related contextual information, some of which may be nonessential, thereby increasing the chance of inference attacks that reconstruct private attributes or behaviors. Second, even when limited to non-sensitive content, high-frequency data sharing facilitates aggregation of large volumes of dispersed information, allowing adversaries to mine behavioral patterns or re-identify users over time.

VII. Summary and Lessons Learned

The IoA inherits security and privacy challenges from traditional networked systems while introducing new risks stemming from its unique characteristics, such as large model foundations, decentralization, task-driven cooperation, semantic-aware interaction, and coupled cyber-physical effects.

- For identity authentication, IoA agents are vulnerable to identity forgery, impersonation, Sybil attacks, privilege escalation, and intent deception, which undermine access control in dynamic IoA. Task-aware and context-aware access control mechanisms is essential to dynamically ensure secure authentication. Besides, DIDs combined with verifiable credentials and blockchain-based registries provides tamper-resistant identity management. However,

TABLE III
SUMMARY OF TYPICAL SECURITY AND PRIVACY THREATS AND CORRESPONDING DEFENSES IN IOA.

Categories	Threats	Defenses	Description	Ref.
Identity Authn Threats	Identity forgery	Access control	Use false capability claims and identity spoofing to gain unauthorized access in IoA.	[6]
	Impersonation	Access control	Inject false information by mimicking trusted agents in coordination workflows within IoA.	[25]
	Sybil attack	Access control	Generate multiple agents to distort group decisions and bypass verification.	[26]
	Privilege escalation	Access control	Exploit logic flaws or vulnerablities to gain higher access than authorized within IoA.	[27]
	Intent deception	Access Control	Disguise malicious intent to access systems under the guise of legitimate objectives.	[7]
Trust Issues in Agent Cooperation	Hallucination cascade	RAG, agent audit, network topology, post-correction	Amplify and propagate agent-generated errors across agent interactions, degrading decision reliability.	[11]
	Knowledge poisoning	Trust Management	Inject false or biased information into shared knowledge bases to manipulate other agent outputs.	[30]
	Adversarial attack	Network topology	Craft adversarial inputs within task workflows to disrupt collaborative agent behavior.	[9]
	Jailbreak	Filtering, multi-agent debate	Bypass agent safeguards through crafted prompts to induce unauthorized or misaligned outputs.	[31]
	Prompt injection	Prevention, detection	Utilize malicious instructions in prompts to divert agents from intended behaviors.	[32]
	Free-riding	Coalition game, Shapley value	Exploit cooperation by contributing minimal or low-quality outputs while benefiting from group work.	[59]
	Agent collusion	Trust Management	Coordinate among compromised agents to fabricate consensus and manipulate collective outcomes.	[8]
Embodied Security Threats	Attacks on agent sensors	Physical defense, information redundancy, data fusion	Exploit external signals to disrupt sensor integrity and compromise embodied agent safety across diverse platforms (e.g., UAVs and autonomous vehicles).	[43]–[47]
	Contextual backdoor	World models, adversarial fine-tuning	Utilize malicious contextual triggers to covertly induce unsafe or unintended agent behaviors.	[10], [48]
	Cross-domain safety misalignment	Multimodal consistency validation	Cause inconsistencies between linguistic decisions and physical actions due to embodied agents' understanding gaps.	[49]
Privacy Threats	Privacy inference	Privacy pre-assessment, output intervention	Infer sensitive user attributes by analyzing contextual signals exchanged during agent collaboration.	[52]
	RAG privacy leakage	Privacy pre-assessment, output intervention	Extract private data from long-term memory utilizing RAG via crafted queries in IoA.	[53], [54]
	Agent memorization	Privacy pre-assessment, output intervention	Leak sensitive training or interaction data through unintended memorization and in-context reproduction.	[55]

achieving low-latency revocation and privacy preservation at scale remain an open challenge.

- For trusted agent cooperation, hallucination cascades can amplify reasoning errors across chained agents; knowledge poisoning and adversarial input can corrupt shared repositories; jailbreak and prompt-injection attacks can bypass safeguards; and free-riding and collusion threaten fair contribution. Grounding outputs via RAG, multi-agent auditing, topology-aware isolation, and debate-style verification can improve robustness in cooperative tasks.

- For embodied agents, sensor-level attacks (e.g., LiDAR spoofing and IMU interference), contextual backdoors, and cross-modal safety misalignment can lead to harmful physical behaviors. Combining hardware shielding, sensor redundancy, world-model simulation, and multimodal consistency checks can detect and block malicious behaviors.

- For privacy, contextual inference attacks reconstruct private attributes from exchanged metadata; RAG-based pipelines leak sensitive knowledge through adversarial queries; and agents may memorize and inadvertently disclose PII via in-context learning. Pre-deployment privacy risk assessment and runtime output intervention (e.g., filtering or redaction) can mitigate leaks.

From Sections III–VI, we have learned that securing IoA requires end-to-end protection across identity, communication, inference, and actuation layers. Static rules are insufficient; instead, IoA defenses should incorporate semantic awareness (e.g., context-aware anomaly detection) and adapt in real time. Furthermore, bridging low-level exploits to high-level, system-wide impacts, especially in cyber-physical settings, requires integrated frameworks that span networking, control theory, and human-agent interaction. Additionally, technical measures should be complemented by legal frameworks, certification processes, and ethical guidelines to ensure accountability in cross-jurisdictional deployments. Table III summarizes the major security and privacy threats in IoA, alongside representative mitigation strategies, providing a roadmap for building resilient and trustworthy agent ecosystems.

VIII. FUTURE RESEARCH DIRECTIONS

In this section, we identify a series of future research directions to enhance the efficiency, security, trustworthiness, privacy, and ethics of IoA ecosystems.

A. Cloud–Edge Cooperative Large-Scale Agent Networking

Achieving low-latency and high-throughput coordination among millions of heterogeneous agents demands seamless collaboration between cloud datacenters and edge nodes. Future works should design adaptive workload partitioning strategies that dynamically offload computation and synchronize states based on network conditions, task priority, and resource availability. Federated learning and model distillation at the edge can help maintain lightweight agent footprints while preserving global consistency [60]. Besides, fine-grained monitoring mechanisms are essential to preempt congestion and ensure predictable service quality in mission-critical scenarios.

B. Security-by-Design IoA

Rather than brought-in security approaches, IoA platforms should embed built-in security mechanisms throughout the agent lifecycle. For instance, it requires formally verified communication stacks with built-in authentication and authorization, tamper-evident logs of inter-agent messages, and hardware-rooted trust anchors to ensure code integrity [19]. Research should explore domain-specific security patterns, such as policy-based access control for financial agents or real-time attestation for robotic agents, and develop automated tooling to generate secure agent compositions from high-level specifications.

C. Trustworthy Regulation in IoA

Decentralized agent ecosystems pose unique regulatory challenges in IoA, as no single authority governs agent identities or behaviors. Future studies should explore governance frameworks that combine on-chain credentialing (e.g., decentralized identifiers with verifiable credentials) with off-chain dispute resolution mechanisms [29]. Embedding audit trails into agent interactions via immutable ledgers or privacy-preserving blockchains can enable transparent investigations without sacrificing privacy. Developing interoperable regulation schemes and liability frameworks are critical to foster public confidence and legal compliance.

D. Privacy-Aware Agent Architectures

Agents continuously share contextual and behavioral data, raising risks of privacy leakage and profiling. Privacy-by-design techniques should be tailored to high-frequency and low-latency demands of IoA. Research should also investigate agent communication protocols that grant fine-grained consent control for each agent interaction, while enforcing privacy policies across dynamically composed agent workflows.

E. Ethical Frameworks for Autonomous Agents

As agents exhibit high autonomy in decision-making, they should operate within clear ethical bounds. Future research should embed ethical principles into agent planning and execution modules, alongside runtime monitors to detect unethical behavior. Cross-disciplinary collaborations with ethicists, social scientists, and legal experts are necessary to codify culturally aware value systems and to design explainable justification logs with accountability.

IX. CONCLUSIONS

In this survey, we have explored the emerging security and privacy challenges that arise as AI agents interconnect to form the IoA. We have first characterized the distinctive threat surface of IoA infrastructures, spanning decentralized identity management, cross-agent trust, embodied agent security, and privacy. We have then reviewed a range of emerging and potential defense strategies to address them and identified critical gaps of existing mechanisms to keep pace with the dynamic and semantics-rich interactions unique to IoA systems. Finally, we have pointed out future research directions critical to advancing resilient, scalable, and privacy-aware IoA deployments. By charting this landscape, we aim to guide future efforts toward fostering trustworthy agent ecosystems that can securely harness the full potential of autonomous and collaborative intelligence.

REFERENCES

[1] Y. Wang, Y. Pan, Q. Zhao, Y. Deng, Z. Su, L. Du, and T. H. Luan, "Large model agents: State-of-the-art, cooperation paradigms, security and privacy, and future trends," *arXiv preprint arXiv:2409.14457*, pp. 1–40, 2024.

[2] Z. Zhao, W. Chai, X. Wang, B. Li, S. Hao, S. Cao, T. Ye, and G. Wang, "See and think: Embodied agent in virtual environment," in *European Conference on Computer Vision*, pp. 187–204, 2024.

[3] Gartner, "Intelligent agents in AI." https://www.gartner.com/en/articles/intelligent-agent-in-ai, 2023. Accessed on 2025-01-19.

[4] W. Chen, Z. You, R. Li, yitong guan, C. Qian, C. Zhao, C. Yang, R. Xie, Z. Liu, and M. Sun, "Internet of agents: Weaving a web of heterogeneous agents for collaborative intelligence," in *Proc. ICLR*, pp. 1–32, 2025.

[5] Z. Aminiranjbar, J. Tang, Q. Wang, S. Pant, and M. Viswanathan, "DAWN: Designing distributed agents in a worldwide network," *Authorea Preprints*, pp. 1–11, 2024.

[6] N. Jiang, B. Sun, T. Sim, and J. Han, "Can i hear your face? pervasive attack on voice authentication systems with a single face image," in *Proc. USENIX Security*, pp. 1045–1062, 2024.

[7] W. Hao, J. Liu, W. Li, and L. Chen, "CDA: Covert deception attacks in multi-agent resource scheduling," *IEEE Robotics and Automation Letters*, vol. 9, no. 11, pp. 9215–9222, 2024.

[8] S. R. Motwani, M. Baranchuk, M. Strohmeier, V. Bolina, P. H. Torr, L. Hammond, and C. S. de Witt, "Secret collusion among ai agents: Multi-agent deception via steganography," in *Proc. NeurIPS*, vol. 37, pp. 73439–73486, 2024.

[9] R. M. S. Khan, Z. Tan, S. Yun, C. Flemming, and T. Chen, "Agents under siege: Breaking pragmatic multi-agent llm systems with optimized prompt attacks," *arXiv preprint arXiv:2504.00218*, 2025.

[10] R. Jiao, S. Xie, J. Yue, T. SATO, L. Wang, Y. Wang, Q. A. Chen, and Q. Zhu, "Can we trust embodied agents? exploring backdoor attacks against embodied LLM-based decision-making systems," in *Proc. ICLR*, pp. 1–31, 2025.

[11] M. Zhang, O. Press, W. Merrill, A. Liu, and N. A. Smith, "How language model hallucinations can snowball," in *Proc. ICML*, pp. 1–15, 2024.

[12] B. C. Das, M. H. Amini, and Y. Wu, "Security and privacy challenges of large language models: A survey," *ACM Computing Surveys*, vol. 57, no. 6, 2025.

[13] F. He, T. Zhu, D. Ye, B. Liu, W. Zhou, and P. S. Yu, "The emerged security and privacy of llm agent: A survey with case studies," *arXiv preprint arXiv:2407.19354*, 2024.

[14] S. Wang, T. Zhu, B. Liu, M. Ding, X. Guo, D. Ye, W. Zhou, and P. S. Yu, "Unique security and privacy threats of large language model: A comprehensive survey," *arXiv preprint arXiv:2406.07973*, 2024.

[15] Y. Gan, Y. Yang, Z. Ma, P. He, R. Zeng, Y. Wang, Q. Li, C. Zhou, S. Li, T. Wang, *et al.*, "Navigating the risks: A survey of security, privacy, and ethics threats in llm-based agents," *arXiv preprint arXiv:2411.09523*, 2024.

[16] Y. Li, H. Wen, W. Wang, X. Li, Y. Yuan, G. Liu, J. Liu, W. Xu, X. Wang, Y. Sun, *et al.*, "Personal LLM agents: Insights and survey about the capability, efficiency and security," *arXiv preprint arXiv:2401.05459*, 2024.

[17] D. Zhang, G. Feng, Y. Shi, and D. Srinivasan, "Physical safety and cyber security analysis of multi-agent systems: A survey of recent advances," *IEEE/CAA Journal of Automatica Sinica*, vol. 8, no. 2, pp. 319–333, 2021.

[18] Z. Deng, Y. Guo, C. Han, W. Ma, J. Xiong, S. Wen, and Y. Xiang, "AI agents under threat: A survey of key security challenges and future pathways," *ACM Computing Surveys*, vol. 57, no. 7, pp. 1–36, 2025.

[19] K. Wang, G. Zhang, Z. Zhou, J. Wu, M. Yu, S. Zhao, C. Yin, J. Fu, Y. Yan, H. Luo, *et al.*, "A comprehensive survey in LLM (-agent) full stack safety: Data, training and deployment," *arXiv preprint arXiv:2504.15585*, 2025.

[20] J. Chen, X. Wang, R. Xu, S. Yuan, Y. Zhang, W. Shi, J. Xie, S. Li, R. Yang, T. Zhu, *et al.*, "From persona to personalization: A survey on role-playing language agents," *Transactions on Machine Learning Research*, pp. 1–50, 2024.

[21] Anthropic, "Model context protocol (MCP)." https://www.anthropic.com/news/model-context-protocol, 2024. Accessed: Jan. 20, 2025.

[22] Google, "Agent to agent protocol (A2A)." https://google.github.io/A2A, 2025. Accessed: Apr. 26, 2025.

[23] "Agent network protocol (ANP)." https://agentnetworkprotocol.com/en/, 2024. Accessed: Jan. 20, 2025.

[24] E. A. University of Oxford, "Agora protocol." https://agoraprotocol.org, 2024. Accessed: Apr. 26, 2025.

[25] X. Hou, Y. Zhao, S. Wang, and H. Wang, "Model context protocol (MCP): Landscape, security threats, and future research directions," *arXiv preprint arXiv:2503.23278*, 2025.

[26] Y. Wu, C. Ying, N. Zheng, W.-A. Zhang, and S. Zhu, "Whole-process privacy-preserving and sybil-resilient consensus for multiagent networks," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–13, 2024.

[27] I. Labs, "MCP security notification: Tool poisoning attacks." https://invariantlabs.ai/blog/mcp-security-notification-tool-poisoning-attacks, 2025. Accessed: 2025-04-25.

[28] J. Qiu, Z. Tian, C. Du, Q. Zuo, S. Su, and B. Fang, "A survey on access control in the age of Internet of things," *IEEE Internet of Things Journal*, vol. 7, no. 6, pp. 4682–4696, 2020.

[29] C. Mazzocca, A. Acar, S. Uluagac, R. Montanari, P. Bellavista, and M. Conti, "A survey on decentralized identifiers and verifiable credentials," *IEEE Communications Surveys & Tutorials*, 2025. doi:10.1109/COMST.2025.3543197.

[30] W. Zou, R. Geng, B. Wang, and J. Jia, "PoisonedRAG: Knowledge poisoning attacks to retrieval-augmented generation of large language models," in *Proc. USENIX Security*, pp. 1–30, 2024.

[31] Z. Chen, Z. Zhao, W. Qu, Z. Wen, Z. Han, Z. Zhu, J. Zhang, and H. Yao, "Pandora: Detailed LLM jailbreaking via collaborated phishing agents with decomposed reasoning," in *ICLR Workshop on Secure and Trustworthy Large Language Models*, pp. 1–15, 2024.

[32] B. Zhang, Y. Tan, Y. Shen, A. Salem, M. Backes, S. Zannettou, and Y. Zhang, "Breaking agents: Compromising autonomous LLM agents through malfunction amplification," *arXiv preprint arXiv:2407.20859*, pp. 1–15, 2024.

[33] C. Song, L. Ma, J. Zheng, J. Liao, H. Kuang, and L. Yang, "Audit-LLM: Multi-agent collaboration for log-based insider threat detection," *arXiv preprint arXiv:2408.08902*, 2024.

[34] M. Zhuge, W. Wang, L. Kirsch, F. Faccio, D. Khizbullin, and J. Schmidhuber, "GPTSwarm: Language agents as optimizable graphs," in *Proc. ICML*, pp. 1–25, 2024.

[35] X. Cao, G. Nan, H. Guo, H. Mu, L. Wang, Y. Lin, Q. Zhou, J. Li, B. Qin, Q. Cui, X. Tao, H. Fang, H. Du, and T. Q. Quek, "Exploring LLM-based multi-agent situation awareness for zero-trust space-air-ground integrated network," *IEEE Journal on Selected Areas in Communications*, 2025. doi:10.1109/JSAC.2025.3560042.

[36] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Proc. NeurIPS*, vol. 33, pp. 9459–9474, 2020.

[37] T. Kwartler, M. Berman, and A. Aqrawi, "Good parenting is all you need–multi-agentic LLM hallucination mitigation," *arXiv preprint arXiv:2410.14262*, 2024.

[38] L. Gao, Z. Dai, P. Pasupat, A. Chen, A. T. Chaganty, Y. Fan, V. Y. Zhao, N. Lao, H. Lee, D. Juan, and K. Guu, "RARR: researching and revising what language models say, using language models," in *Proc. ACL*, pp. 16477–16508, 2023.

[39] Z. Xiang, L. Zheng, Y. Li, J. Hong, Q. Li, H. Xie, J. Zhang, Z. Xiong, C. Xie, C. Yang, *et al.*, "GuardAgent: Safeguard LLM agent by a guard agent via knowledge-enabled reasoning," *arXiv preprint arXiv:2406.09187*, 2024.

[40] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch, "Improving factuality and reasoning in language models through multiagent debate," in *Proc. ICML*, pp. 11733–11763, 2023.

[41] A. Kumar, C. Agarwal, S. Srinivas, S. Feizi, and H. Lakkaraju, "Certifying LLM safety against adversarial prompting," *arXiv preprint arXiv:2309.02705*, pp. 1–32, 2023.

[42] M. Phute, A. Helbling, M. Hull, S. Peng, S. Szyller, C. Cornelius, and D. H. Chau, "LLM self defense: By self examination, LLMs know they are being tricked," in *Proc. ICLR*, pp. 1–11, 2024.

[43] Y. Son, H. Shin, D. Kim, Y. Park, J. Noh, K. Choi, J. Choi, and Y. Kim, "Rocking drones with intentional sound noise on gyroscopic sensors," in *Proc. USENIX Security*, pp. 881–896, 2015.

[44] Z. Hong, X. Li, Z. Wen, L. Zhou, H. Chen, and J. Su, "ESP spoofing: Covert acoustic attack on mems gyroscopes in vehicles," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 3734–3747, 2022.

[45] X. Chen, Z. Li, B. Chen, Y. Zhu, C. X. Lu, Z. Peng, F. Lin, W. Xu, K. Ren, and C. Qiao, "MetaWave: Attacking mmwave sensing with meta-material-enhanced tags," in *Proc. NDSS*, pp. 1–17, 2023.

[46] J.-H. Jang, M. Cho, J. Kim, D. Kim, and Y. Kim, "Paralyzing drones via EMI signal injection on sensory communication channels.," in *Proc. NDSS*, pp. 1–18, 2023.

[47] Z. Fu, Y. Zhi, S. Ji, and X. Sun, "Remote attacks on drones vision sensors: An empirical study," *IEEE Transactions on Dependable and Secure Computing*, vol. 19, no. 5, pp. 3125–3135, 2021.

[48] A. Liu, Y. Zhou, X. Liu, T. Zhang, S. Liang, J. Wang, Y. Pu, T. Li, J. Zhang, W. Zhou, Q. Guo, and D. Tao, "Compromising LLM driven embodied agents with contextual backdoor attacks," *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 3979–3994, 2025.

[49] H. Zhang, C. Zhu, X. Wang, Z. Zhou, C. Yin, M. Li, L. Xue, Y. Wang, S. Hu, A. Liu, P. Guo, and L. Y. Zhang, "BadRobot: Jailbreaking embodied LLMs in the physical world," in *Proc. ICLR*, pp. 1–40, 2025.

[50] J. Xiang, T. Tao, Y. Gu, T. Shu, Z. Wang, Z. Yang, and Z. Hu, "Language models meet world models: embodied experiences enhance language models," in *Proc. NeurIPS*, pp. 75392–75412, 2023.

[51] S.-A. Rebuffi, S. Gowal, D. A. Calian, F. Stimberg, O. Wiles, and T. A. Mann, "Data augmentation can improve robustness," in *Proc. NeurIPS*, pp. 29935–29948, 2021.

[52] R. Staab, M. Vero, M. Balunovic, and M. Vechev, "Beyond memorization: Violating privacy via inference with large language models," in *Proc. ICLR*, pp. 1–47, 2024.

[53] J. Morris, V. Kuleshov, V. Shmatikov, and A. Rush, "Text embeddings reveal (almost) as much as text," in *Proc. EMNLP*, pp. 12448–12460, 2023.

[54] C. Jiang, X. Pan, G. Hong, C. Bao, and M. Yang, "RAG-Thief: Scalable extraction of private data from retrieval-augmented generation applications with agent-based attacks," *arXiv preprint arXiv:2411.14110*, 2024.

[55] N. Carlini, F. Tramèr, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. B. Brown, D. Song, Ú. Erlingsson, A. Oprea, and C. Raffel, "Extracting training data from large language models," in *Proc. USENIX Security*, pp. 2633–2650, 2021.

[56] S. Kim, S. Yun, H. Lee, M. Gubri, S. Yoon, and S. J. Oh, "ProPILE: Probing privacy leakage in large language models," in *Proc. NeurIPS*, pp. 1–18, 2023.

[57] Y. Zeng, Y. Wu, X. Zhang, H. Wang, and Q. Wu, "AutoDefense: Multi-agent LLM defense against jailbreak attacks," in *NeurIPS Workshop on Safe Generative AI*, pp. 1–25, 2024.

[58] Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. E. Zhu, L. Jiang, X. Zhang, S. Zhang, A. Awadallah, R. W. White, D. Burger, and C. Wang, "AutoGen: Enabling next-gen LLM applications via multi-agent conversation," in *Proc. COLM*, pp. 1–43, 2024.

[59] Y. Wang, Z. Su, Y. Pan, T. H. Luan, R. Li, and S. Yu, "Social-aware clustered federated learning with customized privacy preservation," *IEEE/ACM Transactions on Networking*, vol. 32, no. 5, pp. 3654–3668, 2024.

[60] G. Qu, Q. Chen, W. Wei, Z. Lin, X. Chen, and K. Huang, "Mobile edge intelligence for large language models: A contemporary survey," *IEEE Communications Surveys & Tutorials*, pp. 1–42, 2025.