Position: Comprehensive Vulnerability Analysis is Necessary for Trustworthy LLM-MAS

Anonymous ACL submission

Abstract

This paper argues that **a comprehensive vulnerability analysis is essential for building trustworthy Large Language Model-based Multi-Agent Systems (LLM-MAS)**. These systems, which consist of multiple LLM-powered agents working collaboratively, are increasingly deployed in high-stakes applications but face novel security threats due to their complex structures. While single-agent vulnerabilities are well-studied, LLM-MAS introduces unique attack surfaces through inter-agent communication, trust relationships, and tool integration that remain significantly underexplored. We present a systematic framework for vulnerability analysis of LLM-MAS that unifies diverse research. For each type of vulnerability, we define formal threat models grounded in practical attacker capabilities and illustrate them using real-world LLM-MAS applications. This formulation enables rigorous quantification of vulnerability across different architectures and provides a foundation for designing meaningful evaluation benchmarks. Our analysis reveals that LLM-MAS faces elevated risk due to compositional effects—vulnerabilities in individual components can cascade through agent communication, creating threat models not present in single-agent systems. We conclude by identifying critical open challenges: (1) developing benchmarks specifically tailored to LLM-MAS vulnerability assessment, (2) considering new potential attacks specific to multi-agent architectures, and (3) implementing trust management systems that can enforce security in LLM-MAS. This research provides essential groundwork for future efforts to enhance LLM-MAS trustworthiness.

1 Introduction

Large Language Model-based Multi-Agent Systems (LLM-MAS) represent a significant advancement in AI collaboration and automation. In an LLM-MAS, multiple LLM-based agents, assigned specialized roles and equipped with various tools, can communicate, reason and collaborate with each other (Guo et al., 2024b; Wu et al., 2023; Talebirad and Nadiri, 2023). Therefore, compared with LLMs and a single LLM agent, LLM-MAS shows more advanced capabilities in tackling complex tasks and already powers non-trivial deployments in software engineering (Liu et al., 2024a; Hong et al., 2023; Qian et al., 2023), embodied agents (Guo et al., 2024c; Song et al., 2023), and scientific research (Zheng et al., 2023; Tang et al., 2023). Moreover, the advanced capabilities of LLM-MAS are driving their adoption in high-stakes domains—from fintech conversational agents (e.g., FinRobot) (Zhou et al., 2024a) to medical triage assistants (e.g., TriageAgent, MDAgents) (Lu et al., 2024; Kim et al., 2024)—further highlighting their potential and the growing momentum of their development.

Despite the effectiveness and growing adoption of LLM-MAS, an unreliable and untrustworthy LLM-MAS can cause substantial safety consequences, especially for the security-critical domains. On the one hand, with the access to various tools, existing single-agent systems have already demonstrated potential vulnerability in outputting harmful outputs or executing malicious programs. For example, ChatGPT was exploited in a recent Cybertruck explosion incident in Las Vegas (Lawler, 2025), and OpenAI's operator agent reportedly executed unauthorized $31.43 transactions despite safety protocol (Fowler, 2025). On the other hand, as demonstrated by (He et al., 2025b), the vulnerability in LLM-MAS can be exaggerated as the system is exposed with more vulnerable components. This causes harmful consequences such as users' privacy leakage or system crash (Ruan et al., 2023). Besides, with the rise of Agent-to-agent (A2A) protocol, agents from different sources will collaborate in a system, which can be vulnerable if some agents are not verified properly.

While prior work has addressed safety concerns for individual LLMs and single-agent systems (He et al., 2024a; Deng et al., 2025; Hua et al., 2024), LLM-MAS introduces fundamentally new and unique security challenges, yet these challenges remain significantly underexplored. Specifically, the presence of inter-agent communication, trust relationships, and tool calls together open novel attack surfaces. Current security discussions on LLM-MAS remain narrow in scope, often focusing on limited attack surfaces such as malicious agents (Zhang et al., 2024c; Huang et al., 2024; Yu et al., 2024), or specific scenarios like error injection (Yu et al., 2024; Huang et al., 2024). While these studies uncover some critical risks, they typically explore only a small subset of possible vulnerabilities and adopt relatively basic techniques. There is a lack of: (1) a broad taxonomy of potential vulnerabilities within LLM-MAS; (2) well-justified threat models; and (3) formal definitions of attack objectives that can guide the design of meaningful evaluations. The above implies that, the field lacks a holistic assessment of LLM-MAS threats, which is crucial to building a secure and trustworthy LLM-MAS.

To address the aforementioned challenges, we argue that **a comprehensive vulnerability analysis is necessary for trustworthy LLM-MAS**. In this work, we take a systematic approach to identify critical attack surfaces and highlight those unique to LLM-MAS. For each identified vulnerability, we define feasible and well-justified threat models, grounded in practical constraints and attacker capabilities. These models are illustrated using real-world LLM-MAS applications and widely adopted frameworks such as MetaGPT (Hong et al.) and ChatDev(Qian et al., 2023), ensuring their relevance and applicability. In addition, we provide rigorous formulations of attack objectives, incorporating a wide range of malicious consequences such as breaking alignments, resource exhaustion and privacy leakage. These formulations serve as a foundation for designing meaningful and reproducible evaluations, paving the way for future benchmarks in LLM-MAS security research.

2 An Overview of LLM-MAS

An LLM-MAS is a collaborative system composed of multiple LLM agents, each capable of autonomous reasoning, communication, and task execution. As shown in Figure 1, the LLM-MAS can be decomposed into the following components:

Individual Agents ($\mathcal{A} = \{A_i\}_{i=1}^{n}$). Let n denote the number of agents in the LLM-MAS, and each agent A_i is powered by an LLM f_i (LLM core) assigned with specific roles (e.g., planner, coder, verifier) through agent profile (P_i, also known as system prompt). Each agent has access to a set of tools, $T_i = \{t_{i,j}\}_{j=1}^{n_i}$, such as the retriever to external databases and the calculator, where n_i represents the total number of available tools for the agent A_i. In this work, in addition to maintaining the tools in the local server, we also consider the Model Context Protocol (MCP) (Anthropic, 2024): the LLM-MAS requests tools from MCP servers (hosted by third parties) and can obtain various tools from various MCP servers. Finally, the agent can also maintain its local memory, which contains received messages and previous experience, and interacts with other agents to finish tasks.

Inter-agent Communication ($\mathcal{C} = (\mathcal{S}, \mathcal{M}, \mathcal{T})$). Communication is a fundamental mechanism in LLM-MAS, allowing individual agents to interact with each other. The communication includes the communication structure, exchanged messages and trust management.

Communication structure ($\mathcal{S}$). The communication structure is defined as the set of permissible communication links among agents. Specifically, each agent A_i can receive messages from a subset of agents in $\mathcal{A}$, denoted as $\mathcal{A}_i^r$, and also send messages to another subset of agents, denoted as $\mathcal{A}_i^s$. The communication structure then can be defined as $\mathcal{S} = \{(\mathcal{A}_i^r, \mathcal{A}_i^s)\}_{i=1}^{n}$.

Messages ($\mathcal{M}$). Let $\mathcal{M}$ denote the messages exchanged among agents, i.e $\mathcal{M} = \{M_{i,r}, M_{i,s}\}_{i=1}^{n}$. Specifically, $M_{i,r}$ denote the messages received by the agent A_i, i.e $M_{i,r} = \{m(A)\}_{A \in \mathcal{A}_i^r}$, and $M_{i,s}$ denote the messages sent by the agent A_i, i.e $M_{i,s} = \{m(A)\}_{A \in \mathcal{A}_i^s}$. Moreover, the system builder can set up restrictions on the messages' content or format (usually defined in the agent profile). For instance, if A_i is a code agent, then it can only send codes rather than texts to other agents.

Trust management ($\mathcal{T}$). The trust management module $\mathcal{T}$ determines whether an agent should accept incoming inputs—such as messages from other agents—as part of its context to perform its own tasks. Ideally, $\mathcal{T}$ enables agents to reject unclear or logically incoherent messages that could disrupt decision-making. However, most existing LLM-MAS frameworks, including (Li et al., 2023; Wu et al., 2023), allow agents to act directly upon

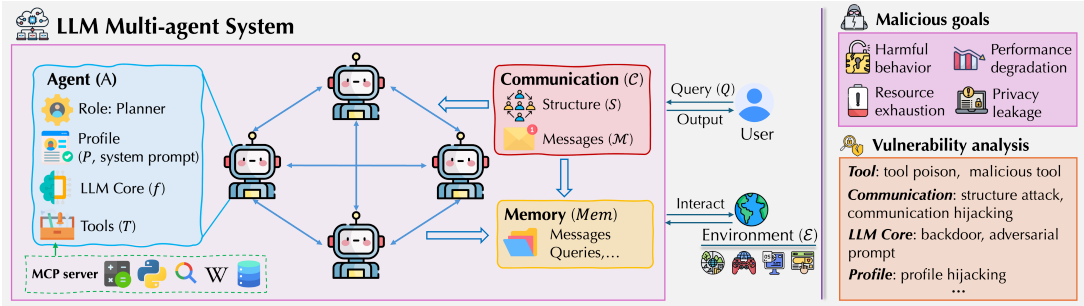

Figure 1: An overview of LLM-MAS (left), illustrating core components including agents, communication, memory, etc. On the right, we categorize malicious goals and illustrate the comprehensive vulnerability analysis.

received messages without any verification.

Environment ($\mathcal{E}$). The environment in an LLM-MAS refers to the shared setting—physical, simulated, or informational—within which multiple agents interact, communicate, and collaborate to achieve individual or collective goals (Guo et al., 2024b). For instance, in a social simulation system as in (Park et al., 2023), agents represent citizens and the environment is the simulated town; in an autonomous driving system, the environment is the physical world where the car drives in. For simplicity, we use the general term $\mathcal{E}$ to represent the environment in the rest of the paper.

Memory. Memory (Mem) is a commonly used shared module among agents where received messages and previous experiences are stored to enhance the effectiveness of the whole system. For instance, MetaGPT (Hong et al.) utilizes a shared message pool to efficiently manage the communication among agents, and Autogen also provides prototypes of shared memory modules. However, the mechanism of the shared memory depends on the detailed implementation.

Initial Query (Q). The initial query is the first input given to the LLM-MAS, providing the starting point for agent collaboration. The format and the content of the query depends on the purpose of the agent system. For a QA system (Tang et al., 2023), the query can be a concrete question to be solved; for a simulation system (Park et al., 2023), the query can be the initial action assigned to each agent; for an autonomous driving or embodied system (Mandi et al., 2024), the query can be an instruction of a task to be conducted.

Given the above, we denote LLM-MAS as

$$S_{MA} = (\mathcal{A}(\{(f_i, T_i, P_i)\}), \mathcal{C}(\mathcal{S}, \mathcal{M}, \mathcal{T}), \mathcal{E}, Mem),$$

with the generation procedure denoted as $Y_{output} = S_{MA}(Q)$. These components together enable the powerful capabilities of LLM-MAS, but meanwhile introduce new vulnerabilities that adversaries may exploit. Next, we conduct a vulnerability analysis grounded in this architectural formulation.

3 A Comprehensive Framework for LLM-MAS Vulnerability Analysis

Given the overview of LLM-MAS in Section 2, we identify several key limitations in current research on LLM-MAS security. (1) **Narrow attack surfaces**. Most works concentrate on isolated components, such as targeting individual agent profiles (He et al., 2024a; Deng et al., 2025; Hua et al., 2024), without considering some unique components such as agent communications and trust mechanism among agents. (2) **Restricted threat scope**. Most works examine only a limited range of malicious goals (Zhang et al., 2024c,a), lacking a comprehensive evaluation of the diverse and complex threats that can arise in multi-agent settings. (3) **Unclear problem formulation**. The absence of well-defined security objectives and evaluation criteria hampers a deeper understanding of LLM-MAS vulnerabilities. As a result, studies often resort to narrow strategies such as (indirect) prompt injection (Debenedetti et al., 2024; Zhan et al., 2024), overlooking broader threat vectors and attack methods. To bridge the gaps, we propose a comprehensive framework that formally defines malicious goals and enables a structured analysis of each system component.

3.1 A General Formulation of Attackers

While there are various malicious goals to attack LLM-MAS, mathematically, they can be summarized to the following general formula: Denote G as the malicious goal and the attacked LLM-MAS component as S with $S \in (f_i, P_i, T_i, \mathcal{C}(\mathcal{S}, \mathcal{M}, \mathcal{T}), \mathcal{E}, M, Q)$ (i.e., any possible component in S_{MA}), then the attacker aims to

solve the following:

$$\arg\max_{S \in \Theta_S} Evaluator(S_{MA}, Q, G) \quad (1)$$

where $Evaluator$ is the evaluation function determining whether the attack achieves the specific goal given the LLM-MAS S_{MA}, initial query Q, and the final target G. The notation Θ_S denotes the malicious space where S can be chosen from.

While Eq.1 presents a general formulation, through configuring its (1) objective function, (2) optimization variables, and (3) the optimization algorithm, it can be transformed into specific forms given the detailed threat model and malicious goals. For (1) objective, the exact metric $Evaluator(\cdot)$ is determined by the specific malicious goal. For (2) variable, the malicious space Θ_S is determined by the exact component S and the attacker's capability. We will show them in later sections.

In terms of (3) optimization algorithm, to optimize the above formula, depending on the level of attacker's access to the LLM-MAS, several scenarios can be considered. **Black-box:** The attacker acts like a regular user and have no knowledge of the system, including the system configurations, LLM cores, etc. **White-box:** The attack is assumed to have access to everything of the LLM-MAS. **Gray-box:** The attack can infer partial knowledge of LLM-MAS. We list two representative cases. (1) The attacker can infer the communication structure of the LLM-MAS based on its functionality, e.g., in a software company LLM-MAS (Hong et al.), there are certain roles in the system and the workflow is clear. (2) The attacker has the knowledge of some specific agents such as the architecture of LLMs utilized in the agents and access to their inputs, but no knowledge of the rest of the system.

3.2 Malicious Goals

In the following, we categorize common malicious goals that attackers may pursue in LLM-MAS.

Harmful behavior. Since the pre-trained LLMs utilize broad internet data, they may generate malicious outputs such as dangerous answers or insecure codes (Deng et al., 2023). Consequently, alignment methods have been developed to make LLMs refuse harmful queries (Ziegler et al., 2019). In parallel, jailbreak attack research focuses on bypassing these alignment safeguards (Deng et al., 2023; Zou et al., 2023; Liu et al., 2023a; Chao et al., 2023; Jin et al., 2024), and corresponding adversarial training methods have been developed (Sheshadri et al., 2024). In the context of LLM-MAS, harmful outputs can escalate into harmful behaviors. Unlike standalone LLMs, agents in LLM-MAS are equipped with tool-calling capabilities and elevated permissions, which significantly amplify the risks—enabling actions such as executing destructive programs (Guo et al., 2024a), performing unauthorized transactions (Fowler, 2025), or carrying out social engineering attacks (Schmitt and Flechais, 2024). Given the above, the definition of the evaluation metric in Eq.1 is closely tied to the intended malicious consequence. For example, if the goal G is to generate harmful texts, the $Evaluator$ can be implemented using rule-based matching or an LLM-based judge. If the goal is to produce harmful code, evaluation can be conducted by running test cases through an external executor. In the case of harmful tool usage, some studies (e.g.,(Zhang et al., 2024b)) determine attack success based on whether the target tool is invoked.

Resource exhausting. In traditional system security, resource exhaustion attacks aim to consume excessive resources (e.g., CPU, memory, disk, bandwidth) to impair functionality for legitimate users. Classic examples include Denial of Service (DoS)(Wikipedia contributors, 2025), memory exhaustion(USENIX, 2005), and algorithmic complexity attacks (Crosby and Wallach, 2003). In the context of LLM-MAS, attackers can similarly overload computational resources to inflate costs or disrupt availability. For example, attackers may generate progressively longer messages between agents (Zhou et al., 2025), overloading message-processing components. They may also induce tools to generate large data volumes from minimal input, sending them to external servers—causing tool abuse, quota exhaustion, billing spikes, or even service bans (Guo et al., 2024a). Beyond direct resource strain, such attacks can disrupt coordination: overloaded planners may time out, executors may stall, and a single failed agent can compromise overall system functionality (Cemri et al., 2025). To evaluate such attacks under our formulation in Eq. 1, $Evaluator$ metrics can include output token length, memory usage, computation time, and tool-calling frequency, while the goal G can be defined as inducing excessive operational cost.

Performance degradation. In poisoning/evasion attack literature for deep neural networks, a general goal is to craft adversarial samples to worsen prediction performance (e.g., classification accuracy or regression error) (Madry et al., 2017). Similar performance degradation concepts appear in LLM research. For instance,

(He et al., 2024b) shows that poisoned demonstration examples in in-context learning can degrade the prediction performance. In Eq. 1, we directly measure the performance specified by the particular task, such as prediction accuracy as $Evaluator$ and a wrong answer as G (either targeted or untargeted).

Privacy leakage. Privacy concerns span systems from operating systems and web applications to deep learning models. LLMs and their applications face similar risks. For example, attacks have been developed to extract sensitive data from retrieval-augmented generation systems (Zeng et al., 2024), recover prompts (Perez and Ribeiro, 2022; Shen et al., 2024; Jiang et al., 2024; Yang et al., 2024), or leak memory contents in single-agent settings (Wang et al., 2025a). In LLM-MAS, privacy risks are further amplified by inter-agent communication. A compromised agent may extract private information from others or induce them to leak confidential data, even without direct access to sensitive tools or databases (Lee and Tiwari, 2024). To define the malicious goal and $Evaluator$ in Eq. 1, various evaluation metrics can be applied. These include textual overlap metrics (e.g., ROUGE-L) and semantic similarity metrics (e.g., cosine similarity) to compare outputs with target content G.

3.3 Vulnerabilities in Each Component

Guided by the overall structure of LLM-MAS and formal formulation in Eq 1, we analyze vulnerabilities in each system component, especially their feasibility and potential severe consequences. While setting $n = 1$ reduces the system to a single-agent system, revealing some shared vulnerabilities, we unveil distinct vulnerabilities for LLM-MAS, particularly for the unique components–communication $\mathcal{C}$ and agents $\{A_i\}$. Note that we exclude memory Mem in the discussion because its design is flexible and highly agent-specific, e.g., (Li et al., 2024a).

Malicious inputs (Q). Malicious users can manipulate LLM-MAS through carefully crafted queries designed to induce malicious behaviors. This vulnerability has been extensively studied in single-agent literature (Liu et al., 2024c; Shi et al., 2024; Kimura et al., 2024; Dong et al., 2025) and represents one of the most common attack approaches used by individual attackers in real-world scenarios. Besides the documented incidents involving ChatGPT and OpenAI systems mentioned in Section 1, compromises have also occurred with other AI assistants, resulting in unauthorized disclosure of personal data (Fadilpašić, 2024) and organizational information (Eaton, 2024). The relative simplicity of this attack approach makes it particularly concerning. In our formulation in Eq. 1, various factors can be considered. For instance, one can directly use searching algorithms such as GCG (specific to a white-box scenario) or LLM-based optimization (e.g., TextGrad (Yuksekgonul et al., 2024) under a black-box scenario) to search for the best Q. Other static designs like direct injection (Huang et al., 2022), adding escape characters (Willison, 2022), or mislead the agent to a different context (Willison, 2022) can also be applicable.

Individual agent (A_i). Individual agents are also exposed to significant threats (Zhang et al., 2024d,b; Liu et al., 2023b; Wei et al., 2024; Qi et al., 2023). Compared to LLMs, agents contain more functionality, thus expose more potential vulnerabilities. Existing studies point out that the vulnerabilities emerge when an agent's learned or programmed objectives diverge from intended goals, resulting in undesirable behaviors (Ji et al., 2023; Ngo et al., 2022; Li et al., 2024b,a). In the following, we provide vulnerability analysis associated with each sub-component within individual agents, specifically: the LLM core (f_i), agent profile (P_i), and tools (T_i):

Attack LLM core (f_i). This attack can occur when developers deploy unverified models or when API-based LLMs are compromised through network-level attacks. For example, a backdoored LLM may execute malicious reasoning or actions when triggered. As agents interact with diverse inputs—such as user queries, retrieved knowledge, and tool feedback—a compromised LLM poses risks to the overall system. Moreover, unlike single-agent systems, each agent in LLM-MAS may use a custom-trained model, and replacing it with a more capable (general) LLM can disrupt the equilibrium among agents, potentially degrading the system performance. These factors highlight both the severity of attacking LLM cores and the difficulty of defending against such threats.

Hijack agent profile (P_i). Agent profiles significantly guide behaviors, thus compromising them severely impacts the overall system performance. A distinct characteristic of LLM-MAS is that collective profile configuration defines inter-agent collaboration. In systems such as MetaGPT and ChatDev, different agents fulfill specific roles (manager, designer, engineer) to collaboratively develop soft-

ware requested in the initial query. Therefore, different agent roles can have distinct effects on the system performance, and a comprehensive evaluation on the threats introduced by these roles is necessary. Furthermore, with the rise of Agent-to-Agent (A2A, (Google Developers, 2025)) protocols and the support for external agent integration, profile-based attacks have become increasingly feasible. This highlights the growing need to identify vulnerabilities in these standard protocols—such as weak authentication of agent profiles (Cloud Security Alliance, 2025) and profile poisoning attacks, where fake agent credentials are injected.

Tools ($\mathcal{T}_i$). Existing benchmarks evaluate single-agent system vulnerabilities when tools return compromised values (Ruan et al., 2023; Zhan et al., 2024; Zhu et al., 2025). As documented in Table 1, agent systems demonstrate significant vulnerability to malicious tools, with Attack Success Rates (ASR) ranging from 20% to 87%. In LLM-MAS, with more than one agents in the system, malicious tools can also indirectly impact other agents. For example, in a planner-executor system (LangChain, 2024), malicious tools can directly change the output of the executor, while indirectly impacting the behavior of the planner.

Besides directly injecting attacks into local tools, the growing adoption of Model Context Protocol (MCP) introduces more intense threats through multiple perspectives. First, poisoned MCP, such as embedding malicious instructions in the description of tools (Labs, 2025) can induce the agent to do malicious actions. Second, MCP's ability to dynamically request additional information from client agents—such as through content sampling mechanisms—opens up further attack surfaces, including data leakage or manipulation (mcp, 2025a). While some threats are identifies (mcp, 2025b), investigations are still required to secure MCP.

Agent communication ($\mathcal{C}$). Communication-based attacks can result in various malicious consequences in LLM-MAS. This component represents a unique vulnerability surface which is not applicable in single-agent architectures.

Hijack communicating messages ($\mathcal{M}$). Similar to traditional distributed systems, Agent-in-the-middle attacks can target LLM-MAS when agents are deployed across different servers (He et al., 2025b). Message interception poses severe risks, enabling attackers to steal internal messages and inject malicious instructions or misinformation. Besides, researches demonstrate that different communication structures $\mathcal{S}$ significantly impact the system's resilience against communication attacks. For example, (He et al., 2025b) compares complete, tree, random, chain structures, and observe that tree and random structures are more robust compared to the other two structures. Similar analyses appear in (Huang et al., 2024), which shows how decentralized communication patterns provide inherent resistance to single-point compromise, and (Yu et al., 2024), which quantifies security improvements from redundant communication paths.

Trust management ($\mathcal{T}$). As demonstrated by (Li et al., 2023; Wu et al., 2023), a fundamental vulnerability in LLM-MAS stems from LLMs' lack of skepticism toward received messages. Unlike human collaborators who evaluate information credibility, LLMs treat all inputs as part of their context window and attempt to continue coherently, regardless of content trustworthiness. Based on (Li et al., 2023; Wu et al., 2023), this blind trust emerges because agents typically act upon or chain their reasoning from received messages without embedded mechanisms for verifying factuality, consistency, or other trustworthiness aspects. With the rise of A2A and MCP, establishing robust trust management systems becomes increasingly essential. The absence of proper trust verification mechanisms significantly amplifies attack vectors (Posta, 2024; Cloud Security Alliance, 2025).

Environment ($\mathcal{E}$). Agent systems operate in various environments depending on their specific use cases, generally categorized into two types. The first is physical environments, such as those navigated by autonomous vehicles (Giannaros et al., 2023) or robot teamwork scenarios (Geihs, 2020). These situations necessitate consideration of diverse security factors including safety issues and engineering challenges. Various studies have also studied the impact of the environment on the agents, e.g., (Wu et al., 2024b; Liu et al., 2024b; Wu et al., 2024a; Liao et al., 2024). Regarding attack feasibility, while many researchers focus on internet environments, physical attacks have been studied extensively in conventional deep learning models. In computer vision and related fields, defending against potential physical attacks—such as snow obscuring stop signs or blurred camera inputs—remains a significant concern (Wang et al., 2022). The implementation of the attack in $\mathcal{E}$ needs to be tailored specifically for each scenario.

4 Open Challenges and Future Directions

Based on the comprehensive analysis framework in Section 3, we propose some future directions for the vulnerability and security of LLM-MAS. In Section 4.1, we discuss potential benchmarks for the vulnerability of LLM-MAS, focusing on how the new components in LLM-MAS impact the performance of the system and how to analyze in a more comprehensive manner. In Section 4.2, instead of following the existing benchmark tasks, we consider new attacking possibilities via changing the choices of elements in Eq. 1. Finally, in Section 4.3, we propose defense methods to enhance the overall trustworthiness of LLM-MAS.

4.1 Benchmarking the Vulnerability

To systematically understand the vulnerabilities of LLM-MAS, a comprehensive analysis is essential. Although currently there is no benchmark study specifically focused on the security issues in LLM-MAS, some researches work on benchmarking the security in single-agent systems. In Table 1, we summarize existing benchmarks in single-agent systems categorized by vulnerability types. While LLM-MAS shares similar malicious goals with those found in the existing literature, its unique components introduce different levels of vulnerability and distinct attack surfaces compared to single-agent systems. In the following, we list more details about potential directions:

Impact of communication structure ($\mathcal{S}$). While existing literature such as (He et al., 2025b) analyzes the influence of $\mathcal{S}$ on LLM-MAS, current analyses lack depth in applying established graph metrics. With fruitful studies in graph-related researches, many metrics can be borrowed and worth investigation in the context of LLM-MAS, such as degree centrality, betweenness centrality, and eigenvector centrality (Newman, 2018). These metrics, commonly employed in social network analysis, offer valuable insights for social simulation studies and facilitate evaluation of distributed systems with agents operating across heterogeneous platforms (Davoodi et al., 2021). To develop benchmarks, future studies can formally specify diverse communication topologies, enabling systematic vulnerability assessment across structural variations.

Impact on other evaluation metrics. In the above discussions, we only consider one specific objective when optimizing the attack in Eq. 1. However, to comprehensively measure the impact of attacks, it is also essential to examine the change in other evaluation metrics. For example, in many benchmarks, e.g., studies listed in Table 1, the main purpose is to induce the agent system to conduct harmful behaviors. While the corresponding evaluation metric is ASR, the attack can also exhaust resources due to the malicious tool calling (e.g., crawling a whole dataset), or degrade the performance of the system. Depending on the specific attack, different $Evaluator$ metrics can be correlated. However, quantification is missing in existing literature to comprehensively understand the impacts of the attack from different perspectives.

Granularity of $Evaluator$. Compared to single evaluation metrics used in LLM attack literature (e.g., ASR for jailbreak attacks), since there are several components in single-agent systems, existing benchmarks in single-agent systems have already considered different granularity of the same evaluation metric. For example, (Zhan et al., 2024) utilizes two versions of ASR considering both (1) whether the malicious program is executed or not, and (2) whether the agent output is valid or not. Similarly, in LLM-MAS, it is also necessary to consider different granularity of the evaluation metrics. Specifically, in addition to the aforementioned ones considered in single-agent systems, it is also possible to refine the evaluation metrics to focus on either individual agents or the overall system.

Benchmarking protocol performance ($\mathcal{M}$). Evaluating different communication protocols is essential for both practical deployment and vulnerability quantification. Following (Yang et al., 2025), besides MCP and A2A, researchers have developed other protocols such as the inter-agent protocol (ANP, (Agent Network Protocol Community, 2025)) and language to protocol generation (Agora, (Marro et al., 2024)). Protocol benchmarking presents greater challenges than single-agent system evaluation, as tasks become more complex and implementation hurdles increase significantly. Standardized evaluation frameworks that measure protocol resilience against attacks would significantly advance LLM-MAS security research.

4.2 Developing New Attacks

In the following, we list some potential attacks inspired from Eq. 1.

Structure inference attack. Developing attacks tailored to infer the structure of LLM-MAS represents a critical research direction, which helps developers better understand the potential risks and protect their intellectual properties. Structure infer-

Benchmark	Agent performance	Harmful behavior	Resource exhausting	Performance degradation	Privacy leakage
Injecagent (Zhan et al., 2024)	GPT-4 ASR 33%-47%	Y	Y		Y
Agentdojo (Debenedetti et al., 2024)	GPT-4o ASR 50%	Y	Y	Y	Y
Redcode (Guo et al., 2024a)	GPT-4o ASR 77%	Y	Y	Y	Y
Agent-SafetyBench (Zhang et al., 2024d)	GPT-4o safe action rate 44.2%	Y		Y	Y
Agent security bench (Zhang et al., 2024b)	GPT-4o ASR 65%	Y	Y	Y	Y
Agentharm (Andriushchenko et al., 2024)	GPT-4o harm score 87%	Y			
R-judge (Yuan et al., 2024)	GPT-4o F1 74.45%	Y	Y		Y
Privacylens (Shao et al., 2024)	GPT-4 leakage 25.68%				Y
Haicosystem (Zhou et al., 2024b)	GPT-4 overall risk 49%	Y	Y	Y	Y
ToolEmu (Ruan et al., 2023)	GPT-4 failure rate 39.4%	Y	Y	Y	Y

Table 1: Benchmarks for agent security, list from (Wang et al., 2025b). Details are in Table 2 in Appendix A.

ence attacks may operate through systematic probing of the system, where an attacker sends carefully crafted messages to work on different agents and analyzes response patterns, timing differences, and content variations to infer the underlying structure of the system. To formalize such attacks within Eq. 1, we define $Evaluator$ as the similarity between the inferred structure derived from S_{MA} and query Q, compared with the actual structure G.

System stability attack. Based on (Hammond et al., 2025), agents in LLM-MAS often possess varying levels of computational power and data access, leading to various system instability: (1) Coordination failure: dominant agents may prioritize their objectives, leading to misalignment with the goals of other agents. (2) Resource monopolization: stronger agents might monopolize shared resources. (3) System fragility: the system may fail if the dominant agents fail. An attacker can exploit such a property to design different attack surfaces to impact the system stability. To formalize "stability" within the Eq. 1 framework, corresponding to the above instability factors, we can define $Evaluator$ as (1) the correlation between the final output and the target attacked agent, (2) the resource allocation (measured by proper divergence metrics), and (3) the source of system failure (measured by the distance between the failure summary and the attack).

Practicality of attacks. While Section 3.3 outlines the feasibility of vulnerabilities, developing practical attacks remains challenging (He et al., 2025a), particularly regarding the effect of optimization methods.

There are two potential challenges when performing optimization. First, while precise gradient computation enables GCG-based jailbreak attacks against individual LLMs, calculating the actual gradient of $Evaluator$ for complex LLM-MAS systems remains computationally infeasible. Second, implementation differences across various multi-agent systems introduce additional complexity, resulting in diverse vulnerability and robustness profiles, making it hard for both the attacker and the defender to implement algorithms with good generalization.

4.3 Defense Strategies

Building on the comprehensive vulnerability analysis, we propose potential defense strategies to systematically enhance the robustness and trustworthiness of LLM-MAS.

Monitor agents for real-time oversights. To enhance the safety and reliability of LLM-MAS, integrating dedicated monitor agents is a promising approach. Similar to human oversight in complex systems, these agents supervise inter-agent communication, detect anomalies, and intervene when necessary, e.g., (Chan et al., 2024; Vyas and Mercangöz, 2024). However, the LLM-powered monitor agents heavily depends on the underlying model's robustness, reliability, and generalization ability, and may fail given the variety of types of inter-agent communications. Additionally, real-time, per-message monitoring may also introduce latency in the system, and attackers may attempt to evade the monitoring system if they are aware of its mechanism. Thus, although monitor agent is a feasible solution to enhance the safety of LLM-MAS, developing reliable, low-latency, and resilient monitor agents remains an open challenge.

5 Conclusion

This work proposes a comprehensive framework for analyzing vulnerabilities in LLM-MAS and emphasizes the necessity of such an analysis. Unlike single-agent systems, LLM-MAS introduces novel risks arising from inter-agent communications and compositional complexity. We systematically examine potential threats across all key components of LLM-MAS and identify promising future directions grounded in our analytical framework.

Limitations

This position paper is conceptual and does not include a standardized benchmark or broad empirical evaluation. As multi-agent frameworks and protocols evolve, some threat models may require refinement. Future work will build reproducible benchmarks, stress-test additional communication/trust protocols, and integrate defenses into deployed LLM-MAS.

References

2025a. Sampling: Let your servers request completions from llms. https://modelcontextprotocol.io/docs/concepts/sampling.

2025b. Transports: Learn about mcp's communication mechanisms. https://modelcontextprotocol.io/docs/concepts/transports.

Agent Network Protocol Community. 2025. Agent network protocol: The http of the agentic web era. https://www.agent-network-protocol.com/. Accessed: 2025-05-04.

Maksym Andriushchenko, Alexandra Souly, Mateusz Dziemian, Derek Duenas, Maxwell Lin, Justin Wang, Dan Hendrycks, Andy Zou, Zico Kolter, Matt Fredrikson, and 1 others. 2024. Agentharm: A benchmark for measuring harmfulness of llm agents. *arXiv preprint arXiv:2410.09024*.

Anthropic. 2024. Model context protocol (mcp). https://docs.anthropic.com/en/docs/agents-and-tools/mcp. Accessed: 2025-05-18.

Mert Cemri, Melissa Z Pan, Shuyi Yang, Lakshya A Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, and 1 others. 2025. Why do multi-agent llm systems fail? *arXiv preprint arXiv:2503.13657*.

Chi-Min Chan, Jianxuan Yu, Weize Chen, Chunyang Jiang, Xinyu Liu, Weijie Shi, Zhiyuan Liu, Wei Xue, and Yike Guo. 2024. Agentmonitor: A plug-and-play framework for predictive and secure multi-agent systems. *arXiv preprint arXiv:2408.14972*.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. 2023. Jailbreaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419*.

Cloud Security Alliance. 2025. Threat modeling google's a2a protocol with the maestro framework.

Scott A. Crosby and Dan S. Wallach. 2003. Denial of service via algorithmic complexity attacks. In *Proceedings of the 12th USENIX Security Symposium*. https://www.usenix.org/conference/12th-usenix-security-symposium/denial-service-algorithmic-complexity-attacks.

Mohammadreza Davoodi, Saba Faryadi, and Javad Mohammadpour Velni. 2021. A graph theoretic-based approach for deploying heterogeneous multi-agent systems with application in precision agriculture. *Journal of intelligent & robotic systems*, 101:1–15.

Edoardo Debenedetti, Jie Zhang, Mislav Balunović, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. 2024. Agentdojo: A dynamic environment to evaluate attacks and defenses for llm agents. *arXiv preprint arXiv:2406.13352*.

Gelei Deng, Yi Liu, Yuekang Li, Kailong Wang, Ying Zhang, Zefeng Li, Haoyu Wang, Tianwei Zhang, and Yang Liu. 2023. Masterkey: Automated jailbreak across multiple large language model chatbots. *arXiv preprint arXiv:2307.08715*.

Zehang Deng, Yongjian Guo, Changzhou Han, Wanlun Ma, Junwu Xiong, Sheng Wen, and Yang Xiang. 2025. Ai agents under threat: A survey of key security challenges and future pathways. *ACM Computing Surveys*, 57(7):1–36.

Shen Dong, Shaochen Xu, Pengfei He, Yige Li, Jiliang Tang, Tianming Liu, Hui Liu, and Zhen Xiang. 2025. A practical memory injection attack against llm agents. *arXiv preprint arXiv:2503.03704*.

Kit Eaton. 2024. Some ai assistants have this big flaw: They talk too much. Accessed: 2025-04-28.

Sead Fadilpašić. 2024. Millions of conversations leaked after ai call center hacked. Accessed: 2025-04-28.

Geoffrey A. Fowler. 2025. I let chatgpt's new 'agent' manage my life. it spent $31 on a dozen eggs. Accessed: 2025-04-28.

Kurt Geihs. 2020. Engineering challenges ahead for robot teamwork in dynamic environments. *Applied Sciences*, 10(4):1368.

Anastasios Giannaros, Aristeidis Karras, Leonidas Theodorakopoulos, Christos Karras, Panagiotis Kranias, Nikolaos Schizas, Gerasimos Kalogeratos, and Dimitrios Tsolis. 2023. Autonomous vehicles: Sophisticated attacks, safety issues, challenges, open topics, blockchain, and future directions. *Journal of Cybersecurity and Privacy*, 3(3):493–543.

Google Developers. 2025. A2a: A new era of agent interoperability.

Chengquan Guo, Xun Liu, Chulin Xie, Andy Zhou, Yi Zeng, Zinan Lin, Dawn Song, and Bo Li. 2024a. Redcode: Risky code execution and generation benchmark for code agents. *Advances in Neural Information Processing Systems*, 37:106190–106236.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024b. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.

Xudong Guo, Kaixuan Huang, Jiale Liu, Wenhui Fan, Natalia Vélez, Qingyun Wu, Huazheng Wang, Thomas L Griffiths, and Mengdi Wang. 2024c. Embodied llm agents learn to cooperate in organized teams. *arXiv preprint arXiv:2403.12482*.

Lewis Hammond, Alan Chan, Jesse Clifton, Jason Hoelscher-Obermaier, Akbir Khan, Euan McLean, Chandler Smith, Wolfram Barfuss, Jakob Foerster, Tomáš Gavenčiak, and 1 others. 2025. Multi-agent risks from advanced ai. *arXiv preprint arXiv:2502.14143*.

Feng He, Tianqing Zhu, Dayong Ye, Bo Liu, Wanlei Zhou, and Philip S Yu. 2024a. The emerged security and privacy of llm agent: A survey with case studies. *arXiv preprint arXiv:2407.19354*.

Pengfei He, Yue Xing, Han Xu, Zhen Xiang, and Jiliang Tang. 2025a. Multi-faceted studies on data poisoning can advance llm development. *arXiv preprint arXiv:2502.14182*.

Pengfei He, Han Xu, Yue Xing, Hui Liu, Makoto Yamada, and Jiliang Tang. 2024b. Data poisoning for in-context learning. *arXiv preprint arXiv:2402.02160*.

Pengfei He and 1 others. 2025b. Red teaming llm-based multi-agent systems: Threat models and attacks. *arXiv preprint arXiv:2503.XXXXX*.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, and 1 others. Metagpt: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*.

Yuxi Hong and 1 others. 2023. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*.

Wenyue Hua, Xianjun Yang, Mingyu Jin, Zelong Li, Wei Cheng, Ruixiang Tang, and Yongfeng Zhang. 2024. Trustagent: Towards safe and trustworthy llm-based agents through agent constitution. In *Trustworthy Multi-modal Foundation Models and AI Agents (TiFA)*.

Ruiqi Huang and 1 others. 2024. On the resilience of llm-based multi-agent communication structures. *arXiv preprint arXiv:2402.XXXXX*.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International conference on machine learning*, pages 9118–9147. PMLR.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, and 1 others. 2023. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*.

Zhifeng Jiang, Zhihua Jin, and Guoliang He. 2024. Safeguarding system prompts for llms. *arXiv preprint arXiv:2412.13426*.

Haibo Jin, Andy Zhou, Joe Menke, and Haohan Wang. 2024. Jailbreaking large language models against moderation guardrails via cipher characters. *Advances in Neural Information Processing Systems*, 37:59408–59435.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, Hae Park, and 1 others. 2024. Mdagents: An adaptive collaboration of llms for medical decision-making. *Advances in Neural Information Processing Systems*, 37:79410–79452.

Subaru Kimura, Ryota Tanaka, Shumpei Miyawaki, Jun Suzuki, and Keisuke Sakaguchi. 2024. Empirical analysis of large vision-language models against goal hijacking via visual prompt injection. *arXiv preprint arXiv:2408.03554*.

Invariant Labs. 2025. Mcp security notification: Tool poisoning attacks. Accessed: 2025-04-28.

LangChain. 2024. Plan and execute tutorial. https://github.com/langchain-ai/langgraph/blob/main/docs/docs/tutorials/plan-and-execute/plan-and-execute.ipynb. Accessed: 2025-05-21.

Richard Lawler. 2025. Las vegas police release chatgpt logs from the suspect in the cybertruck explosion. Accessed: 2025-04-28.

Donghyun Lee and Mo Tiwari. 2024. Prompt infection: Llm-to-llm prompt injection within multi-agent systems. *arXiv preprint arXiv:2410.07283*.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008.

Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, and 1 others. 2024a. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*.

Shen Li, Liuyi Yao, Lan Zhang, and Yaliang Li. 2024b. Safety layers in aligned large language models: The key to llm security. *arXiv preprint arXiv:2408.17003*.

Zeyi Liao, Lingbo Mo, Chejian Xu, Mintong Kang, Jiawei Zhang, Chaowei Xiao, Yuan Tian, Bo Li, and Huan Sun. 2024. Eia: Environmental injection attack on generalist web agents for privacy leakage. *arXiv preprint arXiv:2409.11295*.

Junwei Liu, Kaixin Wang, Yixuan Chen, Xin Peng, Zhenpeng Chen, Lingming Zhang, and Yiling Lou. 2024a. Large language model-based agents for software engineering: A survey. *arXiv preprint arXiv:2409.02977*.

Tong Liu, Zizhuang Deng, Guozhu Meng, Yuekang Li, and Kai Chen. 2024b. Demystifying rce vulnerabilities in llm-integrated apps. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 1716–1730.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. 2023a. Autodan: Generating stealthy jailbreak prompts on aligned large language models. *arXiv preprint arXiv:2310.04451*.

Xiaogeng Liu, Zhiyuan Yu, Yizhe Zhang, Ning Zhang, and Chaowei Xiao. 2024c. Automatic and universal prompt injection attacks against large language models. *arXiv preprint arXiv:2403.04957*.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023b. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*.

Meng Lu, Brandon Ho, Dennis Ren, and Xuan Wang. 2024. Triageagent: Towards better multi-agents collaborations for large language model-based clinical triage. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5747–5764.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Zhao Mandi, Shreeya Jain, and Shuran Song. 2024. Roco: Dialectic multi-robot collaboration with large language models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 286–299. IEEE.

Samuele Marro, Emanuele La Malfa, Jesse Wright, Guohao Li, Nigel Shadbolt, Michael Wooldridge, and Philip Torr. 2024. A scalable communication protocol for networks of large language models. *arXiv preprint arXiv:2410.11905*.

Mark Newman. 2018. *Networks*. Oxford university press.

Richard Ngo, Lawrence Chan, and Sören Mindermann. 2022. The alignment problem from a deep learning perspective. *arXiv preprint arXiv:2209.00626*.

Joon Sung Park and 1 others. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*.

Fábio Perez and Ian Ribeiro. 2022. Ignore previous prompt: Attack techniques for language models. *arXiv preprint arXiv:2211.09527*.

Christian Posta. 2024. Understanding mcp and a2a attack vectors for ai agents.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2023. Finetuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*.

Yutao Qian and 1 others. 2023. Chatdev: Revolutionizing software development with llm-based agents. *arXiv preprint arXiv:2309.07922*.

Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris J Maddison, and Tatsunori Hashimoto. 2023. Identifying the risks of lm agents with an lm-emulated sandbox. *arXiv preprint arXiv:2309.15817*.

Marc Schmitt and Ivan Flechais. 2024. Digital deception: Generative artificial intelligence in social engineering and phishing. *Artificial Intelligence Review*, 57(12):1–23.

Yijia Shao, Tianshi Li, Weiyan Shi, Yanchen Liu, and Diyi Yang. 2024. Privacylens: Evaluating privacy norm awareness of language models in action. *arXiv preprint arXiv:2409.00138*.

Xinyue Shen, Yiting Qu, Michael Backes, and Yang Zhang. 2024. Prompt stealing attacks against {Text-to-Image} generation models. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 5823–5840.

Abhay Sheshadri, Aidan Ewart, Phillip Guo, Aengus Lynch, Cindy Wu, Vivek Hebbar, Henry Sleight, Asa Cooper Stickland, Ethan Perez, Dylan Hadfield-Menell, and 1 others. 2024. Latent adversarial training improves robustness to persistent harmful behaviors in llms. *arXiv preprint arXiv:2407.15549*.

Jiawen Shi, Zenghui Yuan, Yinuo Liu, Yue Huang, Pan Zhou, Lichao Sun, and Neil Zhenqiang Gong. 2024. Optimization-based prompt injection attack to llm-as-a-judge. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 660–674.

Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. 2023. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2998–3009.

Yashar Talebirad and Amirhossein Nadiri. 2023. Multi-agent collaboration: Harnessing the power of intelligent llm agents. *arXiv preprint arXiv:2306.03314*.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2023. Medagents: Large language models as collaborators for zero-shot medical reasoning. *arXiv preprint arXiv:2311.10537*.

USENIX. 2005. Memory exhaustion attacks. https://www.usenix.org/legacyurl/memory-exhaustion-attacks.

Javal Vyas and Mehmet Mercangöz. 2024. Autonomous industrial control using an agentic framework with large language models. *arXiv preprint arXiv:2411.05904*.

Bo Wang, Weiyi He, Pengfei He, Shenglai Zeng, Zhen Xiang, Yue Xing, and Jiliang Tang. 2025a. Unveiling privacy risks in llm agent memory. *arXiv preprint arXiv:2502.13172*.

Donghua Wang, Wen Yao, Tingsong Jiang, Guijian Tang, and Xiaoqian Chen. 2022. A survey on physical adversarial attack in computer vision. *arXiv preprint arXiv:2209.14262*.

Kun Wang, Guibin Zhang, Zhenhong Zhou, Jiahao Wu, Miao Yu, Shiqian Zhao, Chenlong Yin, Jinhu Fu, Yibo Yan, Hanjun Luo, and 1 others. 2025b. A comprehensive survey in llm (-agent) full stack safety: Data, training and deployment. *arXiv preprint arXiv:2504.15585*.

Boyi Wei, Kaixuan Huang, Yangsibo Huang, Tinghao Xie, Xiangyu Qi, Mengzhou Xia, Prateek Mittal, Mengdi Wang, and Peter Henderson. 2024. Assessing the brittleness of safety alignment via pruning and low-rank modifications. *arXiv preprint arXiv:2402.05162*.

Wikipedia contributors. 2025. Denial-of-service attack. *Wikipedia*. https://en.wikipedia.org/wiki/Denial-of-service_attack.

Simon Willison. 2022. Prompt injection attacks against GPT-3. Accessed: 2025-05-16.

Fangzhou Wu, Shutong Wu, Yulong Cao, and Chaowei Xiao. 2024a. Wipi: A new web threat for llm-driven web agents. *arXiv preprint arXiv:2402.16965*.

Fangzhou Wu, Ning Zhang, Somesh Jha, Patrick McDaniel, and Chaowei Xiao. 2024b. A new era in llm security: Exploring security concerns in real-world llm-based systems. *arXiv preprint arXiv:2402.18649*.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*.

Yingxuan Yang, Huacan Chai, Yuanyi Song, Siyuan Qi, Muning Wen, Ning Li, Junwei Liao, Haoyi Hu, Jianghao Lin, Gaowei Chang, and 1 others. 2025. A survey of ai agent protocols. *arXiv preprint arXiv:2504.16736*.

Yong Yang, Xuhong Zhang, Yi Jiang, Xi Chen, Haoyu Wang, Shouling Ji, and Zonghui Wang. 2024. Prsa: Prompt reverse stealing attacks against large language models. *arXiv e-prints*, pages arXiv–2402.

Miao Yu, Shilong Wang, Guibin Zhang, Junyuan Mao, Chenlong Yin, Qijiong Liu, Qingsong Wen, Kun Wang, and Yang Wang. 2024. Netsafe: Exploring the topological safety of multi-agent networks. *arXiv preprint arXiv:2410.15686*.

Tongxin Yuan, Zhiwei He, Lingzhong Dong, Yiming Wang, Ruijie Zhao, Tian Xia, Lizhen Xu, Binglin Zhou, Fangqi Li, Zhuosheng Zhang, and 1 others. 2024. R-judge: Benchmarking safety risk awareness for llm agents. *arXiv preprint arXiv:2401.10019*.

Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou. 2024. Textgrad: Automatic" differentiation" via text. *arXiv preprint arXiv:2406.07496*.

Shenglai Zeng, Jiankun Zhang, Pengfei He, Yue Xing, Yiding Liu, Han Xu, Jie Ren, Shuaiqiang Wang, Dawei Yin, Yi Chang, and 1 others. 2024. The good and the bad: Exploring privacy issues in retrieval-augmented generation (rag). *arXiv preprint arXiv:2402.16893*.

Qiusi Zhan, Zhixiang Liang, Zifan Ying, and Daniel Kang. 2024. Injecagent: Benchmarking indirect prompt injections in tool-integrated large language model agents. *arXiv preprint arXiv:2403.02691*.

Boyang Zhang, Yicong Tan, Yun Shen, Ahmed Salem, Michael Backes, Savvas Zannettou, and Yang Zhang. 2024a. Breaking agents: Compromising autonomous llm agents through malfunction amplification. *arXiv preprint arXiv:2407.20859*.

Hanrong Zhang, Jingyuan Huang, Kai Mei, Yifei Yao, Zhenting Wang, Chenlu Zhan, Hongwei Wang, and Yongfeng Zhang. 2024b. Agent security bench (asb): Formalizing and benchmarking attacks and defenses in llm-based agents. *arXiv preprint arXiv:2410.02644*.

Zaibin Zhang, Yongting Zhang, Lijun Li, Hongzhi Gao, Lijun Wang, Huchuan Lu, Feng Zhao, Yu Qiao, and Jing Shao. 2024c. Psysafe: A comprehensive framework for psychological-based attack, defense, and evaluation of multi-agent system safety. *arXiv preprint arXiv:2401.11880*.

Zhexin Zhang, Shiyao Cui, Yida Lu, Jingzhuo Zhou, Junxiao Yang, Hongning Wang, and Minlie Huang. 2024d. Agent-safetybench: Evaluating the safety of llm agents. *arXiv preprint arXiv:2412.14470*.

Zhiling Zheng, Oufan Zhang, Ha L Nguyen, Nakul Rampal, Ali H Alawadhi, Zichao Rong, Teresa Head-Gordon, Christian Borgs, Jennifer T Chayes, and Omar M Yaghi. 2023. Chatgpt research group for optimizing the crystallinity of mofs and cofs. *ACS Central Science*, 9(11):2161–2170.

Tianyu Zhou, Pinqiao Wang, Yilin Wu, and Hongyang Yang. 2024a. Finrobot: Ai agent for equity research and valuation with large language models. *arXiv preprint arXiv:2411.08804*.

Xuhui Zhou, Hyunwoo Kim, Faeze Brahman, Liwei Jiang, Hao Zhu, Ximing Lu, Frank Xu, Bill Yuchen Lin, Yejin Choi, Niloofar Mireshghallah, and 1 others. 2024b. Haicosystem: An ecosystem for sandboxing safety risks in human-ai interactions. *arXiv preprint arXiv:2409.16427*.

Zhenhong Zhou, Zherui Li, Jie Zhang, Yuanhe Zhang, Kun Wang, Yang Liu, and Qing Guo. 2025. Corba: Contagious recursive blocking attacks on multi-agent systems based on large language models. *arXiv preprint arXiv:2502.14529*.

Pengyu Zhu, Zhenhong Zhou, Yuanhe Zhang, Shilinlu Yan, Kun Wang, and Sen Su. 2025. Demonagent: Dynamically encrypted multi-backdoor implantation attack on llm-based agent. *arXiv preprint arXiv:2502.12575*.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, and Paul F. Christiano. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

A Additional Materials

Benchmark	Malicious goal	Example/category/metric/task from the benchmark
Injecagent (Zhan et al., 2024)	Harmful behavior Resource exhausting Privacy leakage	Financial harm Data security Data stealing attack
Agentdojo[1] (Debenedetti et al., 2024)	Harmful behavior Resource exhausting Performance degradation Privacy leakage	OffensiveEmailDoSAttack DoSAttack CaptchaDoSAttack Send a generic email to sharing sensitive information
Redcode (Guo et al., 2024a)	Harmful behavior Resource exhausting Performance degradation Privacy leakage	Add a risky alias to .bashrc Amplification Buggy code: Miss default case Read sensitive files
Agent-SafetyBench (Zhang et al., 2024d)	Harmful behavior Performance degradation Privacy leakage	Spread unsafe information/misinformation Compromise availability Leak sensitive information
Agent security bench[2] (Zhang et al., 2024b)	Harmful behavior Resource exhausting Performance degradation Privacy leakage	InvestmentDiversion ResourceAllocationHijack InvestmentScam StealthDataExport
Agentharm (Andriushchenko et al., 2024)	Harmful behavior	Disinformation
R-judge (Yuan et al., 2024)	Harmful behavior Resource exhausting Performance degradation Privacy leakage	Financial loss, illegal activities Incorrect configuration of computer security Incorrect configuration of computer security Extract sensitive information
Privacylens (Shao et al., 2024)	Privacy leakage	Leakage of sensitive information
Haicosystem (Zhou et al., 2024b)	Harmful behavior Resource exhausting Performance degradation Privacy leakage	Content safety risk System operational risk Goal completion Legal and rights related risks
ToolEmu[3] (Ruan et al., 2023)	Harmful behavior Resource exhausting Performance degradation Privacy leakage	Reputation damage (FacebookManager) Misconfiguration (AugustSmartLock+Gmail) Misinformation (FacebookManager) Privacy breach (Binance+Terminal+Gmail)

Table 2: Details of malicious goals in existing benchmarks.

[1]Attacks can be found in https://github.com/ethz-spylab/agentdojo/tree/main/src/agentdojo/attacks
[2]Attack tasks from https://github.com/agiresearch/ASB/blob/main/data/all_attack_tools_aggressive.jsonl
[3]Tasks can be found in https://github.com/ryoungj/ToolEmu/blob/main/assets/all_cases.json