Position: Agent-Specific Trustworthiness Risk as a Research Priority

Zeming Wei [1] Tianlin Li [2] Xiaojun Jia [2] Yihao Zhang [1] Yang Liu [2] Meng Sun [1]

Abstract

The rapid development of Large Language Models (LLMs) has facilitated the development of AI agents for various applications. However, ensuring the trustworthiness of these LLM-based agents, encompassing aspects such as safety, robustness, and privacy, remains a critical challenge. While existing research predominantly addresses risks inherent to LLMs, the distinct vulnerabilities introduced by agent systems' design, including their perception, action, and interaction mechanisms, are insufficiently explored. These components expand the attack surface for adversaries, amplifying risks that demand urgent research attention. In this position paper, we comprehensively analyze trustworthiness risks specific to LLM-based agents, emphasizing threats arising from agent-specific modules going beyond standalone LLMs. Specifically, we summarize these risks across six dimensions, discuss their potential mitigation strategies, and highlight gaps in current attacks and defenses. Although preliminary studies have identified some of these risks, we argue that challenges stemming from agent systems still remain underprioritized and insufficiently addressed. Based on these discussions, we advocate for more research efforts to bridge this gap, ensuring the secure and responsible deployment of LLM-based agents in real-world scenarios.

1. Introduction

Large Language Models (LLMs) have made impressive strides across various applications. In particular, they have facilitated the wide development of **LLM-based AI agents** (***abbrev*. agents in this paper**) by serving as their internal reasoning brains (Xi et al., 2025; Li, 2024). A typical agent system can perceive information from the environment, reason and plan with the brain (LLMs), achieve goals with actions, and interact with other agents or humans, as outlined in Figure 1. Leveraging the LLMs' comprehension and reasoning capabilities as their brains, these agents can successfully handle complex real-world tasks (*e.g.*, web shopping or navigation) by combining other entities, *i.e.*, additional modules beyond LLMs. We provide further details on formulations of agents in Section 3.

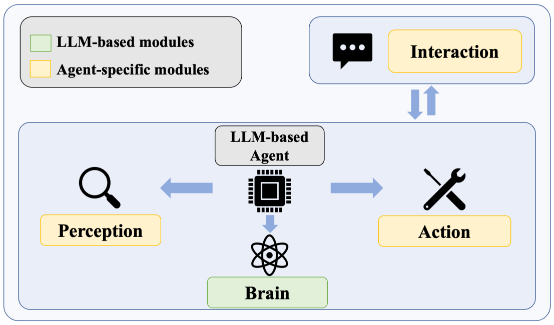

Figure 1. An illustration of agent systems. The brain is an LLM-based module (green), while perception, action, and interaction are agent-specific modules (yellow).

Despite remarkable success, a key challenge of deploying LLMs in real-world settings is their trustworthiness concerns, *e.g.* safety (Zou et al., 2023), privacy (Li et al., 2023), and fairness (Chu et al., 2024) issues. So far, such risks of LLMs have been relatively well-identified and formulated, establishing a solid research convention (Wang et al., 2023a; Huang et al., 2024c; Zhang et al., 2024e). However, the additional modules of agents, including perception, action, and interaction modules, expose broader attack surfaces than LLMs, introducing new risks and concerns about their trustworthy deployment. Unfortunately, the current research literature on this problem mostly focuses on the brains of agents (*i.e.*, LLMs), and agent-specific trustworthiness risk that primarily arises from other parts of agent systems seems to be underexplored. In this paper, we call ***agent-specific*** as related to at least one aspect of perception, action, and interaction of agents.

For instance, in a key paper list from a recent survey of agent risks (Gan et al., 2024), only 11 out of 41 papers are related to agent-specific risks, while the remaining 30 papers are purely LLM risk papers. In addition, we performed an analysis of the ICLR 2025 papers that utilize keyword

[1] Peking University [2] Nanyang Technological University. Correspondence to: Meng Sun <sunm@pku.edu.cn>.

Published at ICML 2025 Workshop on Reliable and Responsible Foundation Models. Copyright 2025 by the author(s).

matching in their titles, where we count the number of papers that include terms like *LLM* and *agent*. The results in Table 1 indicate that only a small proportion of papers in the trustworthiness area focus on agent-related issues, far less than the average of all areas. Such imbalances demonstrate that the current trustworthy research literature primarily emphasizes LLMs while somewhat neglecting other components of agents. However, given the extended capabilities of agents beyond LLMs, their newly introduced risks may lead to more severe outcomes, making trustworthy research on them an urgent priority.

Table 1. Paper numbers from different primary areas in ICLR 2025. '**Alignment etc.**' is abbreviated for **Alignment, fairness, safety, privacy, and societal considerations**.

Primary Area	LLM	Agent	$\frac{\text{Agent}}{\text{LLM}}$ %
All	724	181	25.0%
Alignment etc.	133	17	12.8%
$\frac{\text{Alignment etc.}}{\text{All}}$ %	18.4%	9.3%	/

In this position paper, we emphasize the trustworthy risks posed by agents that go beyond LLMs, and promote further research to identify and mitigate these issues. Following established conventions in trustworthy LLM research (Huang et al., 2024c), we discuss these risks from six dimensions: Truthfulness, Safety, Robustness, Fairness, Privacy, and Ethics. Unlike previous surveys (Gan et al., 2024; Cui et al., 2024; Deng et al., 2024b) that collectively summarize the threats from LLMs and other agent modules, this position paper focuses on the agent-specific parts to shed light on their unique risks. Therefore, this paper does **not** discuss risks that emerge exclusively from the LLMs. Instead, we highlight the underexplored yet urgent potential threats posed by these agents beyond LLMs, and call for more research to identify and mitigate them.

> **Our position: Research on agent-specific trustworthiness risk is urgent yet underexplored. We advocate prioritizing studies to systematically identify and mitigate these agent-specific risks arising from their perception, action, and interaction modules.**

Our contributions in this paper can be summarized as follows:

1. **A cutting-edge literature review.** Through six trustworthy dimensions, we systematically summarize the recent state-of-the-art research on identifying and mitigating agent-specific trustworthy risks.
2. **A roadmap of future research directions.** We also provide insights into unexplored potential aspects of agent risks, suggesting future directions for current research literature.
3. **A call for research priority.** By highlighting these underexplored aspects of agent-specific trustworthy risks, we call for more prioritized research on this topic to build more reliable agent systems.

2. Related Work

2.1. Related work on discussing agent risks

Recently, there have been some works on surveying the risks of agents (Deng et al., 2024b; He et al., 2024; Hua et al., 2024; Yu et al., 2025). The key difference between our work and theirs is that we only focus on highlighting the agent-specific risks, while these surveys discuss the risks from the LLMs and agents collectively. Though LLMs are a fundamental part of the agent system and jointly discussing agent-specific and LLM risks can provide a more comprehensive understanding of the agent trustworthiness risk problem, as a position paper, our work aims to highlight the additional risks arising from the agent-specific design rather than offer an exhaustive survey of the complete agent system. Besides, another difference between our work and existing ones is the taxonomy of the risks. Unlike these works that classify the risks mainly through different agent modules (Cui et al., 2024) or execution stages (Deng et al., 2024b; He et al., 2024), our work discusses the risks through different trustworthy dimensions, offering an intuitive view of agent risks.

3. Agent Formulations

Before delving into the trustworthiness risks in detail, this section provides a comprehensive formulation of agents to better understand their principles over simple LLMs.

3.1. Agent modules

AI Agents can be generally defined as artificial entities that perceive the environment, plan decisions, and take actions to achieve specific goals (Wooldridge & Jennings, 1995; Xi et al., 2025). Conventional AI Agents can be built by symbolic execution or reinforcement learning. Recently, the advanced capabilities of LLMs have positioned them as effective brains of agents, transforming the study of these systems. These agents can use LLMs as the brain for reasoning in various tasks, as illustrated in Figure 1. Meanwhile, these agents can leverage perception modules to acquire additional information from various resources and achieve complex goals through actions with additional tools. Furthermore, in an agent system, agents can collaborate with other agents or interact with users. We define the essential terminology related to agents in this paper as follows:

- **Perception** is all information gathered by the agent for reasoning or planning, including the user input, internal and external knowledge, *e.g.*, the RAG (Lewis et al., 2020) module.
- **Brain** is the base LLM of the agent.
- **Action** refers to the operations that the agent employs to achieve its objectives in various environments, *e.g.*, virtual tool execution and physical actions like web navigation (Deng et al., 2024a).
- **Interaction** covers all information exchange beyond a single agent, including agent-agent (Park et al., 2023; Du et al., 2023) and agent-human interactions (Gao et al., 2024b).

3.2. Agent configuations

In real-world deployments, agents require various operational configurations to facilitate their specific modules, like their interaction context, character design, and task specification. These configurations constitutionally shape their lifecycles, as well as their associated risks.

Context. The first defining configuration of agents is their context. Unlike input/output-oriented LLMs, agents typically operate in diverse contexts. These contexts may include internal knowledge bases or external environments, formalizing specifications for agents in this context. Knowledge bases can provide contextual information for agents to reason with retrieval mechanisms like memory vectors (Lewis et al., 2020), and environments can provide additional feedback for agents to plan further actions (Fan et al., 2022). For example, a research agent (Schmidgall et al., 2025; Kang & Xiong, 2024) might be built using scientific knowledge bases and code execution environments, while a shopping agent could integrate user preference databases and inventory trackers to dynamically adjust its strategies (Yao et al., 2022; Deng et al., 2024a). However, these contextual configurations can also expand the attack surface, allowing adversaries to exploit vulnerabilities present in environmental inputs.

Character. Another critical feature of agents is their particular character. The character of an agent can be defined as a set of cognitive and behavioral traits that shape how it interacts with environments, users, and other agents. Different from plain LLMs trained for generating neutral outputs, agents are often imbued with personalities to enhance user engagement and align with application-specific tasks. For instance, a customer service agent might adopt a friendly and empathetic tone, while a financial advisor agent could prioritize caution and analytical rigor. These personalities can be implemented through LLMs (fine-tuning or prompt engineering (Shinn et al., 2023)) or explicit specification files that guide agent operation. Furthermore, in multi-agent scenarios, different agents may have diverse roles linked to their characters, such as specific roles in research within an agent research laboratory (Schmidgall et al., 2025). These characters may also have vulnerabilities, where adversaries may exploit their character patterns to manipulate agent actions.

Goal. Finally, agents are typically task-specialized, with configurations tailored to targeted applications. Specifically, goal configurations determine the behavior patterns and decision-making logic of agents in specific environments. Such specification also involves particular access considerations and risk-utility trade-offs for agents, unlike unified standards for single LLMs. For example, a healthcare agent might prioritize privacy-preserving retrieval and strict ethical safeguards (Shojaei et al., 2024), while a financial agent emphasizes robustness against adversarial market data (Chen et al., 2025b). Together, these novel configurations of agents underscore the need for trustworthiness strategies for environmental contexts, agent characters, and domain-specific requirements.

4. Dissecting Trustworthiness Risks of Agents

In this paper, we summarize the current research on agent-specific risks across six dimensions, along with potential unexplored directions for further investigation. Due to space limitations, we exemplify the safety dimension in the main content and leave other dimensions in the Appendix A. For each dimension, we identify a few representative and realistic threats categorized within that dimension, where we begin by highlighting the respective attack surfaces and then outline possible mitigation strategies.

4.1. Safety

Safety risks encompass threats of harmful outcomes with societal consequences, a concern prominently exemplified by jailbreaking attacks in LLMs (Zou et al., 2023; Shen et al., 2024; Liu et al., 2023b). For agent systems, these weaknesses are intensified by exploiting broader input sources, untrusted interaction mechanisms, and undergeneralized safety in new environments. These dimensions broaden vulnerabilities beyond the input-output limitations of LLMs, enabling adversaries to exploit systemic weaknesses across agent pipelines.

4.1.1. Perception modules pose more jailbreaking attack surface

Attack surfaces. The rise of jailbreaking typically comes from the malicious inputs (*i.e.*, prompts) for LLMs. Extending the input-output turns of LLMs, agents leverage extra information from knowledge and memory modules for reasoning, which exposes broader surfaces for jailbreaking attackers. This was first realized by AgentPoison (Chen et al., 2024g), which proposes to optimize a malicious trigger that induces the agent to retrieve poisoned knowledge and finally return harmful results, posing new threats under autonomous and healthcare scenarios. Future explorations may include jailbreaking agents by more external retrieval models, where attackers can release these malicious instructions to public resources and induce the agents to retrieve them.

Potential mitigations. Monitoring the harmfulness of perception sources serves as a practical defense, requiring only a safeguard model (Dong et al., 2024b). This approach appears simpler than addressing knowledge poisoning attacks related to truthfulness, as it focuses on verifying the safety of sources instead of their correctness. Nevertheless, advanced attack techniques may target both the safeguard models and base LLMs simultaneously, which still challenges the effectiveness of this filtering solution (Zhang et al., 2024b; Mangaokar et al., 2024; Chen et al., 2025a) and demand more reliable filtering methods specialized for agent systems.

4.1.2. Infection risks in interaction mechanisms

Attack surfaces. The interaction mechanisms of multi-agent systems amplify risks by enabling adversarial inputs to propagate through shared memory modules or direct agent-to-agent infection. Shared memory vulnerabilities, akin to in-context attacks on standalone LLMs (Wei et al., 2023b; Anil et al., 2024; Zheng et al., 2024), allow attackers to inject harmful demonstrations into retrieval or memory pipelines, corrupting agent reasoning. For instance, Agent-Smith (Gu et al., 2024) demonstrates how a single malicious agent can jailbreak entire systems via adversarial chat interactions. On the other hand, direct infection exploits information exchange channels, where compromised agents transmit poisoned data (e.g., biased knowledge or malicious prompts) to peers, triggering cascading failures (Yu et al., 2024). These interaction-induced threats extend beyond a single text-based model. Additional threats, such as alternative communication methods among agents or adversaries, may also be explored in future work.

Potential mitigations. Two potential approaches can identify potential agent infections: monitoring interactions and assessing each agent individually. While examining all interaction logs can quickly reveal issues (Song et al., 2024a), it demands considerably higher computational resources. Alternatively, one can track each agent separately, for instance, by implementing a safety agent that continuously verifies the safety of all other agents (Xiang et al., 2024b).

4.1.3. Under-generalized safety in new environments

Attack surfaces. The operational versatility of agent systems, *i.e.* enabling complex task execution in virtual and physical environments, introduces novel safety risks as adversaries exploit scenario-specific vulnerabilities of these agents. While base LLMs may resist traditional jailbreaking, their safety alignment often fails to generalize across agent deployment context (Wei et al., 2023a). For example, BrowserART (Kumar et al., 2024) reveals that aligned LLMs' refusal capabilities collapse when integrated into browser agents, where simple query rephrasing bypasses safeguards. Such vulnerabilities also extend to physical systems, *e.g.* RoboPAIR (Robey et al., 2024) and BadRobot (Zhang et al., 2024a) demonstrate how agents execute harmful physical actions even in black-box settings. Similarly, RedCode (Guo et al., 2024) and Imprompter (Fu et al., 2024b) expose risks in code environments, where adversarial prompts induce unsafe code generation. These findings highlight the systemic undergeneralization of LLM safety in agent frameworks. Future attacks may consider targeting increasingly diverse environments, like spanning multi-modal interactions, IoT ecosystems, and embodied systems.

Potential mitigations. Similar to recent safety issues of LLMs related to out-of-distribution jailbreaking (Yuan et al., 2023; Handa et al., 2024), a potential cause of these vulnerabilities among agents in various environments stems from the safety under-generalization of the base LLM (Wei et al., 2023a). However, generalizing the safety of LLMs across all domains remains challenging (Wolf et al., 2023). Therefore, when considering a particular application of agents within a new environment, a viable mitigation is to employ incremental training or fine-tuning with task-specific safety data to fix the safety within this domain.

5. Conclusion

In this position paper, we assert that the unique trustworthiness risks of LLM-based agents, rooted in their perception, action, and interaction mechanisms, demand an urgent recalibration of research priorities, as current efforts primarily focus on standalone LLM risks while underexplore the agent-specific vulnerabilities. By categorizing these risks across six dimensions, we argue that agents' expanded attack surfaces pose novel threats requiring dedicated mitigation strategies distinct from LLM-centric defenses. We thus advocate for a paradigm shift: prioritizing research on identifying and mitigating agent-specific trustworthiness risk to address gaps in current literature.

Appendix

A. Other Dimensions of Trustworthiness Risks of Agents

A.1. Truthfulness

We start with the truthfulness risks introduced by agents. This aspect for LLMs is mostly related to the misinformation issue (Chen & Shu, 2024), which may be caused by hallucination (Huang et al., 2023b; Xu et al., 2024), backdoor attacks (Li et al., 2024e; Yang et al., 2024a) and data poisoning attacks (Fu et al., 2024a). In the context of agents, their additional attack surfaces include poisoned knowledge sources or backdoor triggers in new environments that result in providing misinformation to users.

A.1.1. Poisoned knowledge sources

Attack surfaces. The perception mechanisms of agents enable them to acquire extra information for reasoning, while posing new attack surfaces for poisoned knowledge. Even if the brain LLM does not exhibit hallucination issues, agents can still retrieve wrong information from knowledge databases, and such conflicts between internal and external knowledge may finally lead to misinformation in the agent output (Zhou et al., 2024b). Previous work (Zhong et al., 2023; Hu et al., 2024b; Zhang et al., 2024c) demonstrated that adversaries can inject adversarial passages that contain incorrect information into the retrieval corpus that induces RAG modules to retrieve them and provide wrong information to users. PoisonedRAG (Zou et al., 2024) further extends this threat from white-box to black-box settings, where attackers can poison retrieval modules by crafting poisoned corpora and submitting them to public databases like Wikipedia. These findings underscore how perception modules amplify poisoning risks beyond standalone LLM vulnerabilities. Future work may explore poisoning attacks targeting diverse modalities of perception and dynamic knowledge sources of agents to identify these threats.

Potential mitigations. The rise of this attack surface primarily occurs during the knowledge collection. A direct solution to this poisoning is adding filters to these public sources (Zhou et al., 2025), but checking their correctness with extra oracles is still challenging. Another way to resolve this is leveraging insights from defenses for classic data poisoning attacks like outlier removal (Steinhardt et al., 2017) or monitoring loss landscape (Liu et al., 2022). For instance, RobustRAG (Xiang et al., 2024a) proposes to collect responses from different retrieved passages individually to detect the poisoned ones. Further investigations can consider leveraging the base LLMs to inspect the conflict between knowledge sources, *e.g.* leveraging the hidden space of LLMs (Tan et al., 2024; Zeng et al., 2024c) or constructing an additional retrieval evaluator (Yan et al., 2024b).

A.1.2. Backdoor Triggers in New Environments

Attack surfaces. Backdoor attacks (Gu et al., 2017; Li et al., 2021a), where the backdoored model can provide incorrect outputs when adding triggers to the input, have threatened (Li et al., 2024e; Carlini et al., 2024) and been mitigated (Liu et al., 2024c) on LLMs. However, agent systems introduce novel attack surfaces through their perception and interaction mechanisms, enabling adversaries to exploit agent-specific triggers across broader contexts. Recent work demonstrates that even minimal perturbations to retrieval pipelines can compromise agent behavior, for example, inserting backdoored passages into a small fraction of training corpora can hijack RAG modules (Long et al., 2024; Xue et al., 2024; Cheng et al., 2024a). Besides, in real-world applications, the backdoor triggers may be embedded in diverse interaction environments (e.g., search results or e-commerce sites), which can stealthily manipulate agent outputs, such as inducing unintended purchases (Yang et al., 2024b). BadAgent (Wang et al., 2024c) further shows that this attack persists even after fine-tuning agents on benign data. Moreover, embodied agents in physical settings (*e.g.*, autonomous vehicles or domestic robots) face pose unique surfaces that can activate backdoor triggers for agents in driving, household, and manipulation contexts (Liu et al., 2024a). These findings highlight the urgent demand for defense strategies that address the unique interplay between agents and their diverse perceptions and interactions, extending beyond text-based backdoor triggers for LLMs. Besides, future attacks may formalize and characterize novel triggers for agents more systematically and comprehensively.

Potential mitigations. Trigger detection and anti-backdoor training are two possible approaches to mitigate backdoor attacks in large-scale ML systems like LLMs (Zhao et al., 2025). Trigger detection may leverage the base LLMs for rectifying these faults, *e.g.* check the consistency between the thought and actions (Li et al., 2024a). On the other hand, involving the potential trigger format of agents in anti-backdoor training techniques (Li et al., 2021b) during the LLM training is also possible.

A.2. Robustness

Unlike truthfulness risks, the robustness risk refers to the vulnerability against perturbations, particularly during inference time. In this section, we discuss the extended attack surfaces of robustness through two kinds of perturbations: optimization-based adversarial examples and manually designed prompt injection perturbations.

A.2.1. Adversarial Perception Sources

Attack surfaces. Conventional neural networks are known to be vulnerable against adversarial examples (Szegedy et al., 2013; Goodfellow et al., 2014), where they can be easily fooled to provide incorrect predictions. This issue was studied and enhanced in the LLM (Wang et al., 2023b) and multi-model LLMs (Zhao et al., 2024). However, the perception modules of agents pose more surface to adding such perturbations that attackers can exploit certain information in environmental or memory sources, such as inserting adversarial instructions into code documents. For operation environments, injecting only a few text or vision adversarial perturbations can achieve a notable attack success rate on manipulating the agent goals (*e.g.*, purchase a specific product (Wu et al., 2024a)), even if the agents have defense mechanisms like self-reflection. Besides these agent perception modules, AutoInject (Huang et al., 2024b) highlights that misinformation can diffuse between multi-agent interactions, showing attackers can even inject adversarial perturbation with malicious agents in a cross-user agent system. Future attacks may leverage advanced conventional adversarial attacks, like transfer attacks (Huang et al., 2023a; Chen et al., 2024c), to attack perception and interaction modules.

Potential mitigations. Incorporating classic adversarial purification methods (Nie et al., 2022; Chen et al., 2024b) during the perception process is a feasible way to filter the adversarial perturbations, especially for vision sources (Li et al., 2024d; Chen et al., 2024a). However, given the complex form of agent perception, more advanced pre-process modules specialized for given environments like web search or codebase need to be designed. Another direction is to inspect the thinking and interaction process of agents, as adversarial examples often cause inconsistency during execution (Grosse et al., 2017; Wang et al., 2019). Such inconsistency may also extend to inter-agents or internal planning processes (Song et al., 2024b). Besides, conducting adversarial training on perception modules like RAG may also enhance the robustness of these components (Zhu et al., 2024).

A.2.2. Indirect Prompt Injection

Attack surfaces. Beyond optimization-based adversarial perturbations, attackers can craft persuasive inputs to hijack agent objectives, a threat initially identified in LLMs as prompt injection (Perez & Ribeiro, 2022; Shi et al., 2024; Liu et al., 2024e) and extended to agents as indirect prompt injection (**IPI**) (Greshake et al., 2023; Liu et al., 2024a; 2023a). Unlike direct input-output attacks on LLMs, IPI exploits agent workflows (*e.g.*, retrieval, tool execution) to influence behaviors in a more stealthy manner. While IPI defenses have gained preliminary attention, some critical research gaps still persist. For example, recent benchmarks (Rossi et al., 2024; Yi et al., 2023; Zhan et al., 2024) reveal that most attacks focus on text or code inputs, overlooking risks posed by multi-modal triggers (*e.g.*, images, audio) or physical-environmental cues. For instance, adversarial sensory inputs could exploit vision or speech modules in embodied agents, while subtle environmental alterations might manipulate agent execution pipelines. Future research must prioritize these underexplored vectors to address the full scope of IPI vulnerabilities in agent systems.

Potential mitigations. Preliminary Defense against IPI attacks has been explored through reformating prompts (Hines et al., 2024; Chen et al., 2024e) or fine-tuning (Yi et al., 2023), but mostly from the LLM input-output perspectives. Considering the unique execution pipeline of agents, a potential defense paradigm is decomposing their workflow and step-by-step verifying the task (Wu et al., 2024b). Besides, designing specific query (Chen et al., 2024d) or parsing rules for a given task may also improve the robustness against such injections.

A.3. Fairness

This section discusses the fairness risks introduced by agent systems. A fair agent system should avoid biased or stereotypical actions, building on the fairness requirement for LLMs (Chu et al., 2024; Anthis et al., 2024), which focuses on ensuring fair text generation. However, the complex sources and environments for agent systems make addressing fairness issues more challenging than for LLMs.

A.3.1. Poisoning Biased Content in Perception

Attack surfaces. Just like previous risks, adversaries can compromise the integrity of knowledge or memory bases by injecting biased content, thereby inducing fairness violations through knowledge imbalance (e.g., skewed retrieval sources favoring specific groups) or direct propagation of prejudiced information (Dai et al., 2024). Although LLMs are aligned for fairness, this alignment of fairness may not be fully preserved during retrieval given the retrieved context, even if the database is moderately censored (Hu et al., 2024a). Leveraging this vulnerability, attackers can inject biased corpora into public sources to intervene in the perception modules. Furthermore, attackers can craft effective biased corpora in black-box settings through surrogate retrieval model imitation (Chen et al., 2024f). These discoveries pose realistic risks to these perception models for collecting information fairly. Besides the internal knowledge, the superficial alignment hypothesis of LLMs (Zhou et al., 2024a; Li & Kim, 2024; Raghavendra et al., 2024) has shown that they struggle to generalize their alignment to new scenarios, and this limitation also applies to fairness issues (Wei et al., 2024; Das et al., 2024). As agents rely on LLMs as their brains, this undergeneralization becomes critical in complex operational environments, yet this risk remains unexplored. For instance, while an LLM might equitably address group-related queries in simple chat interactions, it may exhibit bias when processing multifaceted inputs from contextually ambiguous environments. These scenarios reveal novel risks of input structures or environmental conditions that can inadvertently activate latent biases, even in ostensibly aligned systems. Mitigating these risks demands rigorous safeguards to inspect retrieval integrity and enhance alignment generalization across diverse agent deployment contexts.

Potential Mitigations. Pre-filtering retrieved content during the perception phase offers a direct approach to risk mitigation. However, unlike explicit harmful instructions that contravene established safety protocols, corpora containing subjective opinions or implicit biases pose unique detection challenges for safeguard models (Dong et al., 2024b), rendering direct filtering insufficient. To address this limitation, post-processing strategies during retrieval, such as verifying whether retrieved items holistically represent diverse perspectives across stakeholder groups (Lohia et al., 2019; Asai et al., 2023), is a potential way to complement pre-filtering.

A.3.2. Breaking Fairness through Long-term Interactions

Attack surfaces. Long-term interactions as lifelong learning is a new trend in agent learning (Zheng et al., 2025; Maharana et al., 2024; Hatalis et al., 2023). During long-term interaction, these agents can actively learn new facts and opinions from various sources (*e.g.*, web apps), communicate with other agents, or interact with users, where they can gradually absorb these contents into their long-term memory (Li et al., 2024b). Such exposure to open environments for agents brings new concerns regarding their fairness alignment since these environments may contain subjective opinions. Through long-term interaction with a certain group, agents may learn similar opinions or positions from them, yielding fairness issues.

Potential Mitigations. Regularly checking or auditing memories is a feasible way to ensure fairness in the character of lifelong agents. Adding proper demonstrations for fairness alignment (Lin et al., 2023; Huang et al., 2024a; Wang et al., 2024b) into memory modules can also mitigate this issue from a contextual perspective.

A.4. Privacy

The extension of LLMs to agents raises the risk of various privacy attacks. This includes not only their additional information databases, like memory models, that may lead to privacy leakage (Huang et al., 2023c), but also the increased interaction surfaces that can trigger such leakages. In this section, we identify two representative privacy risks introduced by the agent attack surfaces, including membership inference (Shokri et al., 2017; Hu et al., 2022) and private information leakage (Huang et al., 2022; Yu et al., 2023; Kim et al., 2024).

A.4.1. Membership Inference on Databases

Attack surfaces. Integrating knowledge bases into agents poses a clear risk of membership inference attacks on this data. Unlike traditional membership inference attacks on LLMs (Fu et al., 2023; Song et al., 2024c) that determine if data is part of the training set, which are relatively challenging (Meeus et al., 2024; Duan et al., 2024), this attack on agent knowledge bases can directly reveal whether specific data resides in the memory module. This process appears to be simpler, as the agent can readily access the data. Existing research has shown that agents can expose the existence of a private data point even with direct request (Anderson et al., 2024), generation similarity comparison (Li et al., 2024f), or masking and reconstruction (Liu et al., 2024b), showing significant vulnerability against such attacks.

Potential mitigations. Inspired by the notable success of differential privacy algorithms in defending membership inference (Chaudhuri et al., 2011; Hu et al., 2023), adding proper noises into the knowledge data or the retrieval process can mitigate this threat (Cheng et al., 2024b).

A.4.2. Private Information Leakage

Attack surfaces. Various execution stages of agents can reveal private information to attackers, including extracting local private data (Zeng et al., 2024b) or instruction prompts from agents (Hui et al., 2024). For example, private information extraction can be achieved by iteratively querying the agent system, where adversaries can optimize queries (Jiang et al., 2024) or backdoor triggers (Peng et al., 2024) to induce the reasoning process to reveal the related information in the retrieval databases. Furthermore, the agents' interactions with various environments also make them vulnerable to these attacks, resembling indirect prompt injection attacks. EIA (Liao et al., 2024) implements website injection techniques that deceive agents into revealing private information by entering certain APIs. Besides the internal knowledge bases, prompt leakage becomes another privacy threat. As current agents are facilitated with LLMs by instruction prompts for planning and reasoning (Li, 2024), these prompts become critical and private in commercial usage (Yan et al., 2024a). Unfortunately, attack techniques (Sha & Zhang, 2024; Liang et al., 2024; Perez & Ribeiro, 2022) have shown the vulnerability of LLMs in protecting the system prompts, and this threat is deepened in the agent scenarios (Hui et al., 2024). Further investigations may explore how to steal agent-instruction prompts through diverse environments, as well as collectively steal different prompts in a multi-agent system.

Potential mitigations. To defend against these privacy threats to agents, previous research has explored restricting the data access (Bagdasaryan et al., 2024) and generating synthetic data (Zeng et al., 2024a). Future work could extend beyond these data-centric approaches, such as designing and integrating privacy-aware agents that can monitor internal data communication sources, providing an additional layer of privacy protection.

A.5. Ethics

This section discusses a few representative ethical perspectives of agent risks by mapping existing taxonomies (Huang et al., 2024c; Liu et al., 2023c) to agent scenarios, *e.g.*, aspects like regulation and interpretability.

A.5.1. Misuse of agents

Attack surfaces. The strong capability agents deepened concerns regarding their potential misuses (Anderljung et al., 2024), where attackers may apply or design adversarial agents to achieve malicious goals. This typically happens when the LLM is sufficiently aligned, but the goal of the designed agents is nasty. Recent research has explored the possibility of leveraging agents for malicious uses, like jailbreaking LLMs (Liu et al., 2024d; Dong et al., 2024a; Wang et al., 2024a) and privacy leakage (Nie et al., 2024; Jiang et al., 2024). Their active reasoning and planning abilities make these concerns deeper than a single chat LLM, and this threat may be further extended to physical environments to cause more severe harm.

Potential mitigations. Preventing open-sourced agents from improper usage is an urgent requirement for agent developers. This alignment ability of agents should also be robust even under harmful manipulations, e.g., harmful fine-tuning on LLMs (Qi et al., 2023; Zhang et al., 2024f) or modifying internal interaction logics. Another way to regulate this threat is to add agent-specific watermarks where imperceptible patterns are injected into environmental actions (Yang et al., 2024b).

A.5.2. Transparency-induced threats

Attack surfaces. This aspect emphasizes the risks from the black-box mechanisms of complex agent systems. The internal planning and execution processes render the decision-making procedure unclear for users and even system developers, which poses threats for adversaries to attack the model more imperceptible (Zhang et al., 2024d; Wu et al., 2024c). For instance, attackers can insert imperceptible injections into websites or fulfill a harmful objective (like purchasing a specific item) while still achieving the original user's goal. Besides, the interaction mechanism of agents also remains unclear, making agent collusion (Campedelli et al., 2024; Lin et al., 2024; Wu et al., 2024d) a possible way for adversaries to utilize. These issues urge a more transparent decision logic for agent system execution.

Potential mitigations. Enhancing transparency in agent system mechanisms offers a practical solution to these threats. For instance, the agent developer can outline the execution steps and emphasize key information retrieved for users while performing a task. Furthermore, designing automatic verification protocols for these steps can decrease the oversight costs.

A.5.3. Dishonesty risks of agents

Attack surfaces. Recent work suggests that LLMs may exhibit dishonesty problems (Li et al., 2024c; Chern et al., 2024). Unlike hallucinations discussed above that arise from unintended errors, dishonesty refers to the intentional generation of misleading or inaccurate information, typically caused by the agents' optimization goal, *e.g.* a model may deliberately generate an incorrect result that has higher user satisfaction. Although dishonesty in LLMs can be mitigated through prompting or fine-tuning (Gao et al., 2024a), the dishonesty of agents under diverse application scenarios raises new ethical risks, but has not been systematically investigated yet, typically emerging in the action modules. For example, a service agent may violate the model developer's specification to satisfy the user's instruction, opening attack surfaces for adversaries to exploit agents that betray developers for malicious goals.

Potential mitigations. The unique character and operational context of agents make it impossible for LLMs to detect their dishonesty issues by training alone. One way to mitigate this agent-specific risk is to implement robust verification protocols that assess whether the agent's actions align with its intended character traits. This can involve periodic audits to ensure that the agent's behavior does not deviate from its specified role or personality. Cross-checking within multi-agent systems is also a feasible way to oversee the honesty of individual agents.

B. Alternative Views

View 1: The risks from LLMs are prioritized above other aspects of agents. We acknowledge that trustworthiness risks posed by LLMs are an important part of the agent system, as there are truly vulnerabilities of agents inherited from LLMs (Deng et al., 2024b) and may be solved from the LLM side. For instance, training a super-aligned LLM whose safety generalizes across all environments could potentially mitigate unsafe actions of agents. However, focusing on the agent-specific components is still a viable and efficient approach to addressing these concerns, and whether it is possible to solve all risks from a single LLM remains controversial (Wolf et al., 2023; Wei et al., 2023a). Given the urgent agent-specific risks discussed in this paper, we maintain that the trustworthiness of agents should be a research priority.

View 2: Current research has adequately focused on the agent-specific risks. We recognize that some of the perspectives discussed in this paper, such as indirect prompt injection (Zhan et al., 2024), have garnered notable research attention and have established preliminary foundations. However, we argue that most of the risks highlighted in this paper are still underexplored within the research community and warrant further investigation. Additionally, the potential directions outlined in this paper can guide further research on these formulated aspects. Overall, this position paper asserts that agent-specific risks are not yet sufficiently investigated and can be explored from various perspectives.

C. Summary and Discussion

Based on the dissection of the urgent yet underexplored agent risks above, we advocate for prioritizing research to identify and mitigate those risks. Below, we briefly summarize the key takeaways through the three agent modules.

C.1. Perception modules

Perception modules introduce novel risks beyond conventional LLM input vulnerabilities, primarily through compromised internal and external knowledge. Adversaries can poison internal memory or knowledge bases, which are retrieved for agent planning to propagate misinformation, unsafe actions, or biased outcomes, expanding attack surfaces beyond direct prompt manipulation for LLMs. Meanwhile, sensitive data within these modules becomes susceptible to extraction via membership inference or adversarial queries. Externally sourced knowledge further exposes agents to indirect prompt injection and backdoor triggers, enabling multifaceted adversarial control. To counter these threats, advancing robust perception frameworks, such as RAG systems (Zhou et al., 2024b), is a potential and critical research direction to ensure trustworthy knowledge acquisition and processing in agent perception modules.

C.2. Action modules

The advanced action capabilities of agents, which enable execution in sandboxed, virtual, or physical environments, raise critical trustworthiness concerns beyond basic text or image generation. A primary vulnerability involves adversaries manipulating these functions to trigger unsafe actions, such as executing harmful code or inducing physical harm. Furthermore, such actions risk privacy breaches by inadvertently exposing sensitive data to untrusted entities. To address risks stemming from agent action modules, implementing hierarchical access controls (*e.g.*, privilege restriction) and formal verification techniques (*e.g.*, runtime safety checks) could reduce the likelihood of unintended or adversarial outcomes.

C.3. Interaction mechanisms

The unique interaction mechanisms of agents pose new attack surfaces for infectious risks. In a multi-agent interaction scenario, malicious agents can spread adversarial resources to other agents or shared knowledge modules, compromising the truthfulness, safety, or robustness of the system. In the context of human-agent interaction, an under-inspected agent may be tricked into fairness or privacy issues. More supervision techniques regarding the interaction mechanism, like run-time verification or designing safety-aware agents, are potential directions to mitigate the interaction risks.

References

Anderljung, M., Hazell, J., and von Knebel, M. Protecting society from ai misuse: when are restrictions on capabilities warranted? *AI & SOCIETY*, pp. 1–17, 2024. 8

Anderson, M., Amit, G., and Goldsteen, A. Is my data in your retrieval database? membership inference attacks against retrieval augmented generation. *arXiv preprint arXiv:2405.20446*, 2024. 7

Anil, C., Durmus, E., Rimsky, N., et al. Many-shot jailbreaking. In *NeurIPS*, 2024. 4

Anthis, J., Lum, K., Ekstrand, M., Feller, A., D'Amour, A., and Tan, C. The impossibility of fair llms. *arXiv preprint arXiv:2406.03198*, 2024. 6

Asai, A., Wu, Z., Wang, Y., Sil, A., and Hajishirzi, H. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*, 2023. 7

Bagdasaryan, E., Yi, R., Ghalebikesabi, S., Kairouz, P., Gruteser, M., Oh, S., Balle, B., and Ramage, D. Air gap: Protecting privacy-conscious conversational agents. *arXiv preprint arXiv:2405.05175*, 2024. 7

Campedelli, G. M., Penzo, N., Stefan, M., et al. I want to break free! persuasion and anti-social behavior of llms in multi-agent settings with social hierarchy. *arXiv preprint arXiv:2410.07109*, 2024. 8

Carlini, N., Jagielski, M., Choquette-Choo, C. A., et al. Poisoning web-scale training datasets is practical. In *S&P*, 2024. 5

Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011. 7

Chen, C. and Shu, K. Combating misinformation in the age of llms: Opportunities and challenges. *AI Magazine*, 45 (3):354–368, 2024. 5

Chen, H., Dong, Y., Shao, S., Hao, Z., Yang, X., Su, H., and Zhu, J. Diffusion models are certifiably robust classifiers. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024a. 6

Chen, H., Dong, Y., Wang, Z., Yang, X., Duan, C., Su, H., and Zhu, J. Robust classification via a single diffusion model. In *Forty-first International Conference on Machine Learning*, 2024b. 6

Chen, H., Zhang, Y., Dong, Y., Yang, X., Su, H., and Zhu, J. Rethinking model ensemble in transfer-based adversarial attacks. In *The Twelfth International Conference on Learning Representations*, 2024c. 6

Chen, H., Dong, Y., Wei, Z., Su, H., and Zhu, J. Towards the worst-case robustness of large language models. *arXiv preprint arXiv:2501.19040*, 2025a. 4

Chen, S., Piet, J., Sitawarin, C., and Wagner, D. Struq: Defending against prompt injection with structured queries. *arXiv preprint arXiv:2402.06363*, 2024d. 6

Chen, Y., Li, H., Zheng, Z., et al. Defense against prompt injection attack by leveraging attack techniques. *arXiv preprint arXiv:2411.00459*, 2024e. 6

Chen, Z., Liu, J., Liu, H., et al. Black-box opinion manipulation attacks to retrieval-augmented generation of large language models. *arXiv preprint arXiv:2407.13757*, 2024f. 6

Chen, Z., Xiang, Z., Xiao, C., et al. Agentpoison: Red-teaming llm agents via poisoning memory or knowledge bases. In *NeurIPS*, 2024g. 4

Chen, Z., Chen, J., Chen, J., and Sra, M. Position: Standard benchmarks fail–llm agents present overlooked risks for financial applications. *arXiv preprint arXiv:2502.15865*, 2025b. 3

Cheng, P., Ding, Y., Ju, T., et al. Trojanrag: Retrieval-augmented generation can be backdoor driver in large language models. *arXiv preprint arXiv:2405.13401*, 2024a. 5

Cheng, Y., Zhang, L., Wang, J., Yuan, M., and Yao, Y. Remoterag: A privacy-preserving llm cloud rag service. *arXiv preprint arXiv:2412.12775*, 2024b. 7

Chern, S., Hu, Z., Yang, Y., Chern, E., Guo, Y., Jin, J., Wang, B., and Liu, P. Behonest: Benchmarking honesty in large language models. *arXiv preprint arXiv:2406.13261*, 2024. 8

Chu, Z., Wang, Z., and Zhang, W. Fairness in large language models: A taxonomic survey. *ACM SIGKDD explorations newsletter*, 26(1):34–48, 2024. 1, 6

Cui, T., Wang, Y., Fu, C., et al. Risk taxonomy, mitigation, and assessment benchmarks of large language model systems. *arXiv preprint arXiv:2401.05778*, 2024. 2

Dai, S., Xu, C., Xu, S., Pang, L., Dong, Z., and Xu, J. Unifying bias and unfairness in information retrieval: A survey of challenges and opportunities with large language models. *arXiv preprint arXiv:2404.11457*, 2024. 6

Das, S., Romanelli, M., Tran, C., Reza, Z., Kailkhura, B., and Fioretto, F. Low-rank finetuning for llms: A fairness perspective. *arXiv preprint arXiv:2405.18572*, 2024. 6

Deng, X., Gu, Y., Zheng, B., et al. Mind2web: Towards a generalist agent for the web. *NeurIPS*, 2024a. 3

Deng, Z., Guo, Y., Han, C., et al. Ai agents under threat: A survey of key security challenges and future pathways. *arXiv preprint arXiv:2406.02630*, 2024b. 2, 8

Dong, Y., Li, Z., Meng, X., Yu, N., and Guo, S. Jailbreaking text-to-image models with llm-based agents. *arXiv preprint arXiv:2408.00523*, 2024a. 8

Dong, Y., Mu, R., Zhang, Y., et al. Safeguarding large language models: A survey. *arXiv preprint arXiv:2406.02622*, 2024b. 4, 7

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023. 3

Duan, M., Suri, A., Mireshghallah, N., et al. Do membership inference attacks work on large language models? *arXiv preprint arXiv:2402.07841*, 2024. 7

Fan, L., Wang, G., Jiang, Y., Mandlekar, A., Yang, Y., Zhu, H., Tang, A., Huang, D.-A., Zhu, Y., and Anandkumar, A. Minedojo: Building open-ended embodied agents with internet-scale knowledge. In *NeurIPS*, 2022. 3

Fu, T., Sharma, M., Torr, P., Cohen, S. B., Krueger, D., and Barez, F. Poisonbench: Assessing large language model vulnerability to data poisoning. *arXiv preprint arXiv:2410.08811*, 2024a. 5

Fu, W., Wang, H., Gao, C., Liu, G., Li, Y., and Jiang, T. Practical membership inference attacks against fine-tuned large language models via self-prompt calibration. *arXiv preprint arXiv:2311.06062*, 2023. 7

Fu, X., Li, S., Wang, Z., et al. Imprompter: Tricking llm agents into improper tool use. *arXiv preprint arXiv:2410.14923*, 2024b. 4

Gan, Y., Yang, Y., Ma, Z., et al. Navigating the risks: A survey of security, privacy, and ethics threats in llm-based agents. *arXiv preprint arXiv:2411.09523*, 2024. 1, 2

Gao, C., Wu, S., Huang, Y., Chen, D., Zhang, Q., Fu, Z., Wan, Y., Sun, L., and Zhang, X. Honestllm: Toward an honest and helpful large language model. In *NeurIPS*, 2024a. 8

Gao, J., Gebreegziabher, S. A., Choo, K. T. W., Li, T. J.-J., Perrault, S. T., and Malone, T. W. A taxonomy for human-llm interaction modes: An initial exploration. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pp. 1–11, 2024b. 3

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 5

Greshake, K., Abdelnabi, S., Mishra, S., et al. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In *ACM Workshop on Artificial Intelligence and Security*, 2023. 6

Grosse, K., Manoharan, P., Papernot, N., et al. On the (statistical) detection of adversarial examples. *arXiv preprint arXiv:1702.06280*, 2017. 6

Gu, T., Dolan-Gavitt, B., and Garg, S. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017. 5

Gu, X., Zheng, X., Pang, T., et al. Agent smith: A single image can jailbreak one million multimodal llm agents exponentially fast. In *ICML*, 2024. 4

Guo, C., Liu, X., Xie, C., et al. Redcode: Risky code execution and generation benchmark for code agents. *arXiv preprint arXiv:2411.07781*, 2024. 4

Handa, D., Chirmule, A., Gajera, B., and Baral, C. Jailbreaking proprietary large language models using word substitution cipher. *arXiv preprint*, 2024. 4

Hatalis, K., Christou, D., Myers, J., et al. Memory matters: The need to improve long-term memory in llm-agents. In *AAAI Symposium Series*, 2023. 7

He, F., Zhu, T., Ye, D., Liu, B., Zhou, W., and Yu, P. S. The emerged security and privacy of llm agent: A survey with case studies. *arXiv preprint arXiv:2407.19354*, 2024. 2

Hines, K., Lopez, G., Hall, M., et al. Defending against indirect prompt injection attacks with spotlighting. *arXiv preprint arXiv:2403.14720*, 2024. 6

Hu, H., Salcic, Z., Sun, L., Dobbie, G., Yu, P. S., and Zhang, X. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, 54(11s):1–37, 2022. 7

Hu, L., Yan, A., Yan, H., Li, J., Huang, T., Zhang, Y., Dong, C., and Yang, C. Defenses to membership inference attacks: A survey. *ACM Computing Surveys*, 56(4):1–34, 2023. 7

Hu, M., Wu, H., Guan, Z., et al. No free lunch: Retrieval-augmented generation undermines fairness in llms, even for vigilant users. *arXiv preprint arXiv:2410.07589*, 2024a. 6

Hu, Z., Wang, C., Shu, Y., et al. Prompt perturbation in retrieval-augmented generation based large language models. In *KDD*, 2024b. 5

Hua, W., Yang, X., Jin, M., Li, Z., Cheng, W., Tang, R., and Zhang, Y. Trustagent: Towards safe and trustworthy llm-based agents. *arXiv preprint arXiv:2402.01586*, 2024. 2

Huang, H., Chen, Z., Chen, H., Wang, Y., and Zhang, K. T-sea: Transfer-based self-ensemble attack on object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20514–20523, 2023a. 6

Huang, H., Li, Y., Sun, H., Bai, Y., and Gao, Y. How far can in-context alignment go? exploring the state of in-context alignment. *arXiv preprint arXiv:2406.11474*, 2024a. 7

Huang, J., Shao, H., and Chang, K. C.-C. Are large pre-trained language models leaking your personal information? *arXiv preprint arXiv:2205.12628*, 2022. 7

Huang, J.-t., Zhou, J., Jin, T., et al. On the resilience of multi-agent systems with malicious agents. *arXiv preprint arXiv:2408.00989*, 2024b. 6

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023b. 5

Huang, Y., Gupta, S., Zhong, Z., Li, K., and Chen, D. Privacy implications of retrieval-based language models. *arXiv preprint arXiv:2305.14888*, 2023c. 7

Huang, Y., Sun, L., Wang, H., et al. Position: Trustllm: Trustworthiness in large language models. In *ICML*, 2024c. 1, 2, 8

Hui, B., Yuan, H., Gong, N., et al. Pleak: Prompt leaking attacks against large language model applications. In *CCS*, 2024. 7

Jiang, C., Pan, X., Hong, G., Bao, C., and Yang, M. Rag-thief: Scalable extraction of private data from retrieval-augmented generation applications with agent-based attacks. *arXiv preprint arXiv:2411.14110*, 2024. 7, 8

Kang, H. and Xiong, C. Researcharena: Benchmarking llms' ability to collect and organize information as research agents. *arXiv preprint arXiv:2406.10291*, 2024. 3

Kim, S., Yun, S., Lee, H., Gubri, M., Yoon, S., and Oh, S. J. Propile: Probing privacy leakage in large language models. In *NeurIPS*, 2024. 7

Kumar, P., Lau, E., Vijayakumar, S., et al. Refusal-trained llms are easily jailbroken as browser agents. *arXiv preprint arXiv:2410.13886*, 2024. 4

Lewis, P., Perez, E., Piktus, A., et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *NeurIPS*, 2020. 3

Li, C., Liang, J., Cao, B., Chen, J., and Wang, T. Your agent can defend itself against backdoor attacks, 2024a. 5

Li, H., Chen, Y., Luo, J., Wang, J., Peng, H., Kang, Y., Zhang, X., Hu, Q., Chan, C., Xu, Z., et al. Privacy in large language models: Attacks, defenses and future directions. *arXiv preprint arXiv:2310.10383*, 2023. 1

Li, H., Yang, C., Zhang, A., Deng, Y., Wang, X., and Chua, T.-S. Hello again! llm-powered personalized agent for long-term dialogue. *arXiv preprint arXiv:2406.05925*, 2024b. 7

Li, J. and Kim, J.-E. Superficial safety alignment hypothesis. *arXiv preprint arXiv:2410.10862*, 2024. 6

Li, S., Yang, C., Wu, T., Shi, C., Zhang, Y., Zhu, X., Cheng, Z., Cai, D., Yu, M., Liu, L., et al. A survey on the honesty of large language models. *arXiv preprint arXiv:2409.18786*, 2024c. 8

Li, X. A review of prominent paradigms for llm-based agents: Tool use (including rag), planning, and feedback learning. *arXiv preprint arXiv:2406.05804*, 2024. 1, 7

Li, X., Sun, W., Chen, H., et al. Adbm: Adversarial diffusion bridge model for reliable adversarial purification. *arXiv preprint arXiv:2408.00315*, 2024d. 6

Li, Y., Li, Y., Wu, B., et al. Invisible backdoor attack with sample-specific triggers. In *ICCV*, 2021a. 5

Li, Y., Lyu, X., Koren, N., Lyu, L., Li, B., and Ma, X. Anti-backdoor learning: Training clean models on poisoned data. *Advances in Neural Information Processing Systems*, 34:14900–14912, 2021b. 5

Li, Y., Huang, H., Zhao, Y., Ma, X., and Sun, J. Backdoorllm: A comprehensive benchmark for backdoor attacks on large language models. *arXiv preprint arXiv:2408.12798*, 2024e. 5

Li, Y., Liu, G., Wang, C., and Yang, Y. Generating is believing: Membership inference attacks against retrieval-augmented generation. *arXiv preprint arXiv:2406.19234*, 2024f. 7

Liang, Z., Hu, H., Ye, Q., Xiao, Y., and Li, H. Why are my prompts leaked? unraveling prompt extraction threats in customized large language models. *arXiv preprint arXiv:2408.02416*, 2024. 7

Liao, Z., Mo, L., Xu, C., et al. Eia: Environmental injection attack on generalist web agents for privacy leakage. *arXiv preprint arXiv:2409.11295*, 2024. 7

Lin, B. Y., Ravichander, A., Lu, X., Dziri, N., Sclar, M., Chandu, K., Bhagavatula, C., and Choi, Y. The unlocking spell on base llms: Rethinking alignment via in-context learning. In *ICLR*, 2023. 7

Lin, R. Y., Ojha, S., Cai, K., and Chen, M. F. Strategic collusion of llm agents: Market division in multi-commodity competitions. *arXiv preprint arXiv:2410.00031*, 2024. 8

Liu, A., Zhou, Y., Liu, X., et al. Compromising embodied agents with contextual backdoor attacks. *arXiv preprint arXiv:2408.02882*, 2024a. 5, 6

Liu, M., Zhang, S., and Long, C. Mask-based membership inference attacks for retrieval-augmented generation. *arXiv preprint arXiv:2410.20142*, 2024b. 7

Liu, Q., Mo, W., Tong, T., Xu, J., Wang, F., Xiao, C., and Chen, M. Mitigating backdoor threats to large language models: Advancement and challenges. In *2024 60th Annual Allerton Conference on Communication, Control, and Computing*, pp. 1–8. IEEE, 2024c. 5

Liu, T. Y., Yang, Y., and Mirzasoleiman, B. Friendly noise against adversarial noise: a powerful defense against data poisoning attack. *NeurIPS*, 2022. 5

Liu, X., Li, P., Suh, E., Vorobeychik, Y., Mao, Z., Jha, S., McDaniel, P., Sun, H., Li, B., and Xiao, C. Autodan-turbo: A lifelong agent for strategy self-exploration to jailbreak llms. *arXiv preprint arXiv:2410.05295*, 2024d. 8

Liu, Y., Deng, G., Li, Y., Wang, K., Wang, Z., Wang, X., Zhang, T., Liu, Y., Wang, H., Zheng, Y., et al. Prompt injection attack against llm-integrated applications. *arXiv preprint arXiv:2306.05499*, 2023a. 6

Liu, Y., Deng, G., Xu, Z., Li, Y., Zheng, Y., Zhang, Y., Zhao, L., Zhang, T., Wang, K., and Liu, Y. Jailbreaking chatgpt via prompt engineering: An empirical study. *arXiv preprint arXiv:2305.13860*, 2023b. 3

Liu, Y., Yao, Y., Ton, J.-F., et al. Trustworthy llms: A survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*, 2023c. 8

Liu, Y., Jia, Y., Geng, R., Jia, J., and Gong, N. Z. Formalizing and benchmarking prompt injection attacks and defenses. In *USENIX Security*, 2024e. 6

Lohia, P. K., Ramamurthy, K. N., Bhide, M., et al. Bias mitigation post-processing for individual and group fairness. In *ICASSP*, 2019. 7

Long, Q., Deng, Y., Gan, L., Wang, W., and Pan, S. J. Backdoor attacks on dense passage retrievers for disseminating misinformation. *arXiv preprint arXiv:2402.13532*, 2024. 5

Maharana, A., Lee, D.-H., Tulyakov, S., Bansal, M., Barbieri, F., and Fang, Y. Evaluating very long-term conversational memory of llm agents. *arXiv preprint arXiv:2402.17753*, 2024. 7

Mangaokar, N., Hooda, A., Choi, J., et al. Prp: Propagating universal perturbations to attack large language model guard-rails. *arXiv preprint arXiv:2402.15911*, 2024. 4

Meeus, M., Shilov, I., Jain, S., et al. Sok: Membership inference attacks on llms are rushing nowhere (and how to fix it). *arXiv preprint arXiv:2406.17975*, 2024. 7

Nie, W., Guo, B., Huang, Y., et al. Diffusion models for adversarial purification. *arXiv preprint arXiv:2205.07460*, 2022. 6

Nie, Y., Wang, Z., Yu, Y., Wu, X., Zhao, X., Guo, W., and Song, D. Privagent: Agentic-based red-teaming for llm privacy leakage. *arXiv preprint arXiv:2412.05734*, 2024. 8

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023. 3

Peng, Y., Wang, J., Yu, H., and Houmansadr, A. Data extraction attacks in retrieval-augmented generation via backdoors. *arXiv preprint arXiv:2411.01705*, 2024. 7

Perez, F. and Ribeiro, I. Ignore previous prompt: Attack techniques for language models. In *NeurIPS ML Safety Workshop*, 2022. 6, 7

Qi, X., Zeng, Y., Xie, T., Chen, P.-Y., Jia, R., Mittal, P., and Henderson, P. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*, 2023. 8

Raghavendra, M., Nath, V., and Hendryx, S. Revisiting the superficial alignment hypothesis. *arXiv preprint arXiv:2410.03717*, 2024. 6

Robey, A., Ravichandran, Z., Kumar, V., et al. Jailbreaking llm-controlled robots. *arXiv preprint arXiv:2410.13691*, 2024. 4

Rossi, S., Michel, A. M., Mukkamala, R. R., and Thatcher, J. B. An early categorization of prompt injection attacks on large language models. *arXiv preprint arXiv:2402.00898*, 2024. 6

Schmidgall, S., Su, Y., Wang, Z., Sun, X., Wu, J., Yu, X., Liu, J., Liu, Z., and Barsoum, E. Agent laboratory: Using llm agents as research assistants. *arXiv preprint arXiv:2501.04227*, 2025. 3

Sha, Z. and Zhang, Y. Prompt stealing attacks against large language models. *arXiv preprint arXiv:2402.12959*, 2024. 7

Shen, X., Chen, Z., Backes, M., et al. " do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In *CCS*, 2024. 3

Shi, J., Yuan, Z., Liu, Y., et al. Optimization-based prompt injection attack to llm-as-a-judge. In *CCS*, 2024. 6

Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., and Yao, S. Reflexion: Language agents with verbal reinforcement learning. In *NeurIPS*, 2023. 3

Shojaei, P., Vlahu-Gjorgievska, E., and Chow, Y.-W. Security and privacy of technologies in health information systems: A systematic literature review. *Computers*, 13 (2):41, 2024. 3

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. In *S&P*, 2017. 7

Song, C., Ma, L., Zheng, J., Liao, J., Kuang, H., and Yang, L. Audit-llm: Multi-agent collaboration for log-based insider threat detection. *arXiv preprint arXiv:2408.08902*, 2024a. 4

Song, R., Ozmen, M. O., Kim, H., Bianchi, A., and Celik, Z. B. Enhancing llm-based autonomous driving agents to mitigate perception attacks. *arXiv preprint arXiv:2409.14488*, 2024b. 6

Song, Z., Huang, S., and Kang, Z. Em-mias: Enhancing membership inference attacks in large language models through ensemble modeling. *arXiv preprint arXiv:2412.17249*, 2024c. 7

Steinhardt, J., Koh, P. W. W., and Liang, P. S. Certified defenses for data poisoning attacks. *NeurIPS*, 2017. 5

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 5

Tan, X., Luan, H., Luo, M., Sun, X., Chen, P., and Dai, J. Knowledge database or poison base? detecting rag poisoning attack through llm activations. *arXiv preprint arXiv:2411.18948*, 2024. 5

Wang, B., Chen, W., Pei, H., et al. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. In *NeurIPS*, 2023a. 1

Wang, F., Duan, R., Xiao, P., Jia, X., Chen, Y., Wang, C., Tao, J., Su, H., Zhu, J., and Xue, H. Mrj-agent: An effective jailbreak agent for multi-round dialogue. *arXiv preprint arXiv:2411.03814*, 2024a. 8

Wang, J., Dong, G., Sun, J., et al. Adversarial sample detection for deep neural network through model mutation testing. In *ICSE*, 2019. 6

Wang, J., Hu, X., Hou, W., Chen, H., Zheng, R., Wang, Y., Yang, L., Huang, H., Ye, W., Geng, X., et al. On the robustness of chatgpt: An adversarial and out-of-distribution perspective. *arXiv preprint arXiv:2302.12095*, 2023b. 6

Wang, Y., Wu, Y., Wei, Z., Jegelka, S., and Wang, Y. A theoretical understanding of self-correction through in-context alignment. In *NeurIPS*, 2024b. 7

Wang, Y., Xue, D., Zhang, S., and Qian, S. Badagent: Inserting and activating backdoor attacks in llm agents. In *ACL*, 2024c. 5

Wei, A., Haghtalab, N., and Steinhardt, J. Jailbroken: How does LLM safety training fail? In *NeurIPS*, 2023a. 4, 8

Wei, Q., Chan, A. J., Goetz, L., Watson, D., and van der Schaar, M. Actions speak louder than words: Superficial fairness alignment in llms. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*, 2024. 6

Wei, Z., Wang, Y., Li, A., Mo, Y., and Wang, Y. Jailbreak and guard aligned language models with only few in-context demonstrations. *arXiv preprint arXiv:2310.06387*, 2023b. 4

Wolf, Y., Wies, N., Avnery, O., Levine, Y., and Shashua, A. Fundamental limitations of alignment in large language models. *arXiv preprint arXiv:2304.11082*, 2023. 4, 8

Wooldridge, M. and Jennings, N. R. Intelligent agents: Theory and practice. *The knowledge engineering review*, 10(2):115–152, 1995. 2

Wu, C. H., Shah, R. R., Koh, J. Y., Salakhutdinov, R., Fried, D., and Raghunathan, A. Dissecting adversarial robustness of multimodal lm agents. In *NeurIPS 2024 Workshop on Open-World Agents*, 2024a. 6

Wu, F., Cecchetti, E., and Xiao, C. System-level defense against indirect prompt injection attacks: An information flow control perspective. *arXiv preprint arXiv:2409.19091*, 2024b. 6

Wu, F., Wu, S., Cao, Y., and Xiao, C. Wipi: A new web threat for llm-driven web agents. *arXiv preprint arXiv:2402.16965*, 2024c. 8

Wu, Z., Peng, R., Zheng, S., Liu, Q., Han, X., Kwon, B., Onizuka, M., Tang, S., and Xiao, C. Shall we team up: Exploring spontaneous cooperation of competing llm agents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5163–5186, 2024d. 8

Xi, Z., Chen, W., Guo, X., et al. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101, 2025. 1, 2

Xiang, C., Wu, T., Zhong, Z., et al. Certifiably robust rag against retrieval corruption. *arXiv preprint arXiv:2405.15556*, 2024a. 5

Xiang, Z., Zheng, L., Li, Y., Hong, J., Li, Q., Xie, H., Zhang, J., Xiong, Z., Xie, C., Yang, C., et al. Guardagent: Safeguard llm agents by a guard agent via knowledge-enabled reasoning. *arXiv preprint arXiv:2406.09187*, 2024b. 4

Xu, Z., Jain, S., and Kankanhalli, M. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*, 2024. 5

Xue, J., Zheng, M., Hu, Y., Liu, F., Chen, X., and Lou, Q. Badrag: Identifying vulnerabilities in retrieval augmented generation of large language models. *arXiv preprint arXiv:2406.00083*, 2024. 5

Yan, B., Li, K., Xu, M., Dong, Y., Zhang, Y., Ren, Z., and Cheng, X. On protecting the data privacy of large language models (llms): A survey. *arXiv preprint arXiv:2403.05156*, 2024a. 7

Yan, S.-Q., Gu, J.-C., Zhu, Y., and Ling, Z.-H. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*, 2024b. 5

Yang, H., Xiang, K., Ge, M., et al. A comprehensive overview of backdoor attacks in large language models within communication networks. *IEEE Network*, 2024a. 5

Yang, W., Bi, X., Lin, Y., et al. Watch out for your agents! investigating backdoor threats to llm-based agents. *arXiv preprint arXiv:2402.11208*, 2024b. 5, 8

Yao, S., Chen, H., Yang, J., and Narasimhan, K. Webshop: Towards scalable real-world web interaction with grounded language agents. In *NeurIPS*, 2022. 3

Yi, J., Xie, Y., Zhu, B., et al. Benchmarking and defending against indirect prompt injection attacks on large language models. *arXiv preprint arXiv:2312.14197*, 2023. 6

Yu, M., Meng, F., Zhou, X., Wang, S., Mao, J., Pang, L., Chen, T., Wang, K., Li, X., Zhang, Y., et al. A survey on trustworthy llm agents: Threats and countermeasures. *arXiv preprint arXiv:2503.09648*, 2025. 2

Yu, W., Pang, T., Liu, Q., Du, C., Kang, B., Huang, Y., Lin, M., and Yan, S. Bag of tricks for training data extraction from language models. In *ICML*, 2023. 7

Yu, W., Hu, K., Pang, T., Du, C., Lin, M., and Fredrikson, M. Infecting llm agents via generalizable adversarial attack. In *Red Teaming GenAI Workshop*, 2024. 4

Yuan, Y., Jiao, W., Wang, W., Huang, J.-t., He, P., Shi, S., and Tu, Z. Gpt-4 is too smart to be safe: Stealthy chat with llms via cipher. *arXiv preprint arXiv:2308.06463*, 2023. 4

Zeng, S., Zhang, J., He, P., Ren, J., Zheng, T., Lu, H., Xu, H., Liu, H., Xing, Y., and Tang, J. Mitigating the privacy issues in retrieval-augmented generation (rag) via pure synthetic data. *arXiv preprint arXiv:2406.14773*, 2024a. 7

Zeng, S., Zhang, J., He, P., et al. The good and the bad: Exploring privacy issues in retrieval-augmented generation (rag). *arXiv preprint arXiv:2402.16893*, 2024b. 7

Zeng, S., Zhang, J., Li, B., et al. Towards knowledge checking in retrieval-augmented generation: A representation perspective. *arXiv preprint arXiv:2411.14572*, 2024c. 5

Zhan, Q., Liang, Z., Ying, Z., and Kang, D. Injecagent: Benchmarking indirect prompt injections in tool-integrated large language model agents. *arXiv preprint arXiv:2403.02691*, 2024. 6, 9

Zhang, H., Zhu, C., Wang, X., et al. Badrobot: Jailbreaking llm-based embodied ai in the physical world. *arXiv preprint arXiv:2407.20242*, 2024a. 4

Zhang, Q., Xiong, Z., and Mao, Z. Safeguard is a double-edged sword: Denial-of-service attack on large language models, 2024b. 4

Zhang, Q., Zeng, B., Zhou, C., et al. Human-imperceptible retrieval poisoning attacks in llm-powered applications. In *Companion Proceedings of FSE*, 2024c. 5

Zhang, Q., Zhou, C., Go, G., Zeng, B., Shi, H., Xu, Z., and Jiang, Y. Imperceptible content poisoning in llm-powered applications. In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering*, pp. 242–254, 2024d. 8

Zhang, Y., Huang, Y., Sun, Y., Liu, C., Zhao, Z., Fang, Z., Wang, Y., Chen, H., Yang, X., Wei, X., et al. Benchmarking trustworthiness of multimodal large language models: A comprehensive study. *arXiv preprint arXiv:2406.07057*, 2024e. 1

Zhang, Y., Wei, Z., Sun, J., and Sun, M. Adversarial representation engineering: A general model editing framework for large language models. In *NeurIPS*, 2024f. 8

Zhao, S., Jia, M., Guo, Z., et al. A survey of recent backdoor attacks and defenses in large language models. *Transactions on Machine Learning Research*, 2025. 5

Zhao, Y., Pang, T., Du, C., et al. On evaluating adversarial robustness of large vision-language models. *NeurIPS*, 2024. 6

Zheng, J., Shi, C., Cai, X., Li, Q., Zhang, D., Li, C., Yu, D., and Ma, Q. Lifelong learning of large language model based agents: A roadmap. *arXiv preprint arXiv:2501.07278*, 2025. 7

Zheng, X., Pang, T., Du, C., et al. Improved few-shot jailbreaking can circumvent aligned language models and their defenses. In *NeurIPS*, 2024. 4

Zhong, Z., Huang, Z., Wettig, A., and Chen, D. Poisoning retrieval corpora by injecting adversarial passages. In *EMNLP*, 2023. 5

Zhou, C., Liu, P., Xu, P., et al. Lima: Less is more for alignment. In *NeurIPS*, 2024a. 6

Zhou, H., Lee, K.-H., Zhan, Z., Chen, Y., and Li, Z. Trustrag: Enhancing robustness and trustworthiness in rag. *arXiv preprint arXiv:2501.00879*, 2025. 5

Zhou, Y., Liu, Y., Li, X., et al. Trustworthiness in retrieval-augmented generation systems: A survey. *arXiv preprint arXiv:2409.10102*, 2024b. 5, 9

Zhu, J., Yan, L., Shi, H., Yin, D., and Sha, L. Atm: Adversarial tuning multi-agent system makes a robust retrieval-augmented generator. *arXiv preprint arXiv:2405.18111*, 2024. 6

Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J. Z., and Fredrikson, M. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023. 1, 3

Zou, W., Geng, R., Wang, B., and Jia, J. Poisonedrag: Knowledge corruption attacks to retrieval-augmented generation of large language models. *arXiv preprint arXiv:2402.07867*, 2024. 5