

A Survey on Trustworthy LLM Agents: Threats and Countermeasures

Miao Yu^{1,†}, Fanci Meng^{1,†}, Xinyun Zhou⁴, Shilong Wang¹, Junyuan Mao¹, Linsey Pang², Tianlong Chen³, Kun Wang⁴, Xinfeng Li^{4,*}, Yongfeng Zhang⁵, Bo An⁴, Qingsong Wen^{1,*}

¹Squirrel AI Learning, ²Salesforce, ³The University of North Carolina at Chapel Hill,

⁴Nanyang Technological University, ⁵Rutgers University

Abstract

With the rapid evolution of Large Language Models (LLMs), LLM-based agents and Multi-agent Systems (MAS) have significantly expanded the capabilities of LLM ecosystems. This evolution stems from empowering LLMs with additional modules such as memory, tools, environment, and even other agents. However, this advancement has also introduced more complex issues of trustworthiness, which previous research focusing solely on LLMs could not cover. In this survey, we propose the TrustAgent framework, a comprehensive study on the trustworthiness of agents, characterized by **modular taxonomy, multi-dimensional annotations, and technical implementation**. By thoroughly investigating and summarizing newly emerged attacks, defenses, and evaluation methods for agents and MAS, we extend the concept of Trustworthy LLM to the emerging paradigm of Trustworthy Agent. In TrustAgent, we begin by deconstructing and introducing various components of the Agent and MAS. Then, we categorize their trustworthiness into intrinsic (brain, memory, and tool) and extrinsic (user, agent, and environment) aspects. Subsequently, we delineate the multifaceted meanings of trustworthiness and elaborate on the implementation techniques of existing research related to these internal and external modules. Finally, we present our insights and outlook on this domain, aiming to provide guidance for future endeavors. For easy reference, we categorize all the studies mentioned in this survey according to our taxonomy, available at: <https://github.com/Ymm-cll/TrustAgent>.

1 Introduction

The advent of large language models (LLMs) has catalyzed a paradigm shift in artificial intelligence systems [9, 66, 93, 137]. The integration of LLMs as backbone with extra modules (e.g. memory [130], tool [65, 75], and environment [109]) as extensions produces the concept of “LLM-based Agent”, transforming static neural network into dynamic cognitive subject capable of memory retrieval, tool utilization, and environmental interaction. Furthermore, the introduction of inter-agent communication has given rise to the more advanced concept of Multi-agent System (MAS), making the “hyper LLM ecosystem” become even more intricate, interactive and intelligent [36, 94, 106, 118, 124, 125].

Extensive academic research and industry practices have validated this performance hierarchy: MAS > Single Agent > LLM [8, 54, 84, 100]. However, the incorporation of additional modules is a double-edged sword, which has raised new concerns regarding trustworthiness across multiple dimensions, including safety,

Xinfeng Li and Qingsong Wen are the corresponding authors. [†] denotes equal contributions. Contact: Miao Yu (fishthreewater@gmail.com), Xinfeng Li (lxmakeit@gmail.com), and Qingsong Wen (qingsongedu@gmail.com).

Table 1: Comparison between TrustAgent and other surveys.

Survey	Object	Multi-Dimension	Modular	Technique [‡]	MAS*
Liu et al. [62]	LLM	✓	✗	Atk/Eval	✗
Huang et al. [41]	LLM	✓	✗	Eval	✗
He et al. [38]	Agent	✗	✗	Atk/Def	✗
Li et al. [57]	Agent	✓	✗	Atk	✗
Wang et al. [96]	Agent	✗	✗	Atk	✗
Deng et al. [24]	Agent	✗	✓	Atk/Def	✓
Gan et al. [31]	Agent	✓	✗	Atk/Def/Eval	✗
TrustAgent (Ours)	LLM + Agent	✓	✓	Atk/Def/Eval	✓

‡: Attack (Atk), Defense (Def), and Evaluation (Eval) denote TrustAgent’s broad technique view.

*: Multi-agent System (MAS).

privacy, fairness, and truthfulness [38, 96, 102]. From a risk perspective, introducing new modules expands the system’s attack surface, potentially leading to unforeseen vulnerabilities [31, 86, 116]. On the other hand, this integration poses new challenges to existing defense mechanisms and trustworthiness evaluations, necessitating the expansion and upgrade of previous research that focuses solely on the trustworthiness of LLMs or a single agent [40, 131].

Previous survey efforts have delved deeply into the realm of trustworthy LLMs. Liu et al. [62] dissect trustworthiness into seven major categories, focusing specially on establishing standards and guidelines for LLM alignment. Similarly, Huang et al. [41] interpret trustworthiness from 6 perspectives but concentrate on creating benchmarks to evaluate trustworthiness. However, these studies are only partially effective in agent scenarios, highlighting an urgent need to address new trustworthiness issues arising from the introduction of additional modules. Other specialized surveys on trustworthiness in agents primarily focus on the sub-aspects like security and privacy and largely overlap with the content of Trustworthy LLMs [31, 96, 96]. In fact, some works merely address the new issues arising from LLMs serving as the “brain” module of agents [38, 57], while overlooking the unexplored challenges introduced by other additional modules. To highlight our innovations, we compare TrustAgent with other surveys in Table 1.

To this end, we propose the TrustAgent framework, as illustrated in Figure 1, extending the research realm of previous trustworthiness surveys to the new context of agents and MAS. Our taxonomy in TrustAgent exhibits the following features: **(I) Modular**. TrustAgent rigorously categorizes trustworthiness issues based on the internal and external components of agents, specifically divided into intrinsic and extrinsic aspects. The former includes the trustworthiness of the brain, memory, and tools, while the latter encompasses the parts related to users, other agents, and the environment. **(II) Technical**. TrustAgent focuses on the implementation of trustworthy agents, providing a comprehensive summary and outlook on the relevant technology stack from three aspects: attack, defense, and evaluation (a comprehensive picture in Appendix B). **(III) Multi-dimensional**. TrustAgent expands the dimensions

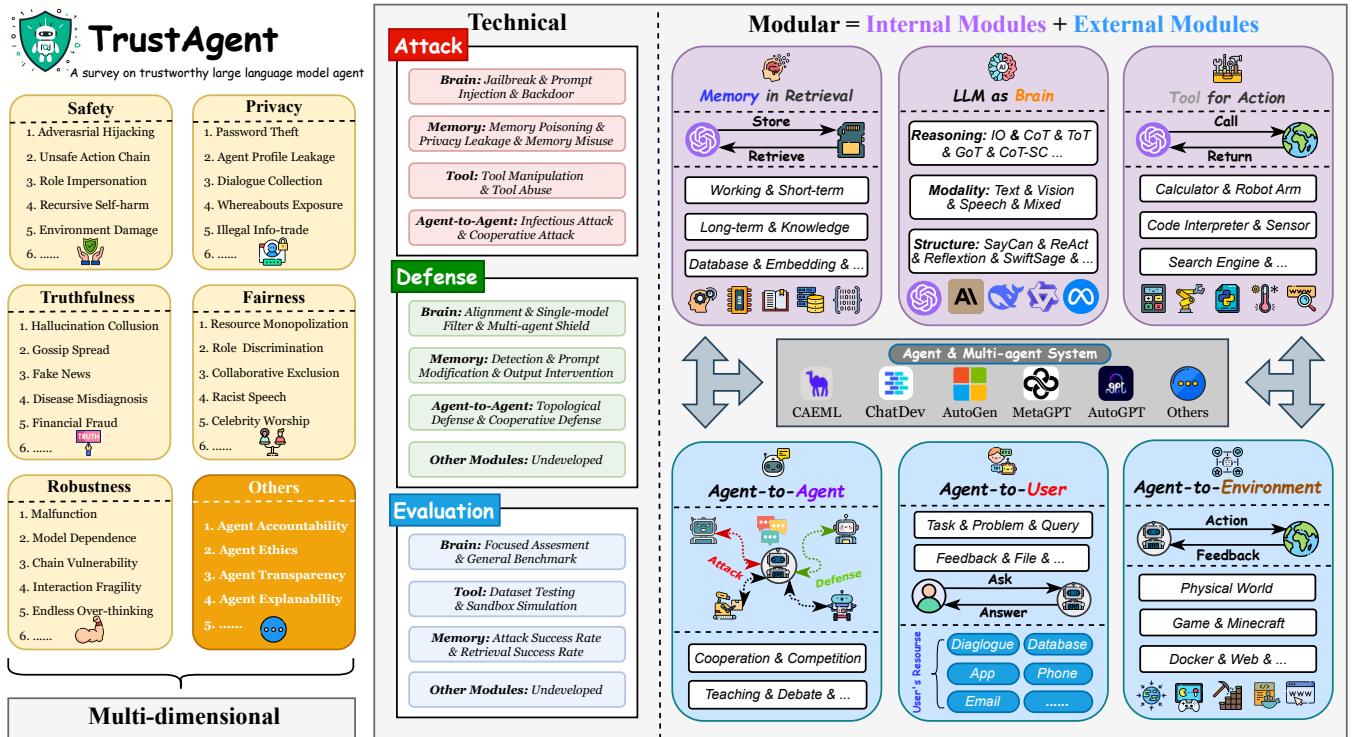


Figure 1: Overview of our TrustAgent taxonomy, featuring multi-dimensional (Left), technical (Middle), and modular (Right).

of LLM trustworthiness to the context of both single agent and MAS, specifically categorized into: safety, privacy, truthfulness, fairness and robustness (with specific definitions in Appendix A), by involving existing works in all these dimensions.

In each subsection, we begin with an overview of the current module’s mechanisms and roles within the agent system, then explore its trustworthiness issues from well-categorized perspectives such as attacks, defenses, and evaluations. Finally, we provide inspiring insights and outline potential future research avenues.

In summary, our contributions can be listed below:

- **Thorough and Latest Survey.** We present a thorough and contemporary analysis of the trustworthiness in LLM-based agent systems, covering a wide spectrum of architectures including single LLM, individual agent, and MAS framework.
- **New Technique-oriented Taxonomy.** Our taxonomy centers on techniques for compromising, achieving, and evaluating trustworthiness, updating old paradigms to the agent context and outlining new technical paradigms within the agent framework.
- **Insightful Future Directions.** For each module’s trustworthiness, we identify current vulnerabilities and outline future directions, urging researchers to delve deeper into this area.

2 Intrinsic Trustworthiness

In this section, we focus on the trustworthiness of the internal modules of agent systems. In our definition, the agent system is an independent entity with human-like cognition, composed of brain (Sec. 2.1) with memory (Sec. 2.2), and behavior in the tool form (Sec. 2.3). Due to the different functions and natures of these modules, the resulting trustworthiness issues vary. To provide in-depth analysis

and instructive insight, we first introduce the role and functions of each module in the agent system and then summarize technical paradigms for attack, defense, and evaluation methods.

2.1 Trustworthy LLM as Brain

LLM agents consist of a central LLM “brain” module [98]. The brain serves as the core reasoning and decision-making center, integrating inputs from various auxiliary modules to guide the agent towards its goal. Though the powerful brain connected with other modules enhances the agent’s ability, this integration expands the potential attack surface, making the system more vulnerable to trustworthiness threats. In this section, we center on brain-related attacks, defenses, and evaluations, with illustration in Figure 2.

2.1.1 Attacks. Unlike single-LLM systems, the brain module in agent or MAS experiences more frequent and complex information dynamics. The textual and visual inputs from the internal and external provide attackers with more interfaces and methods to compromise the trustworthiness of the core brain module. We do not discuss Misalignment attacks (e.g., fine-tuning an LLM to break its alignment) on agents here, as attackers typically lack access to modify the internal parameters of the agents. According to the manipulation mechanisms (as illustrated in Figure 2), we categorize attacks into three paradigms:

Jailbreak attempts to bypass the aligned trustworthy mechanisms within the agent brain via human-designed or optimized adversarial prompts [47]. Early methods, such as GCG and its variants [43, 99, 147], optimize adversarial suffixes to manipulate responses,

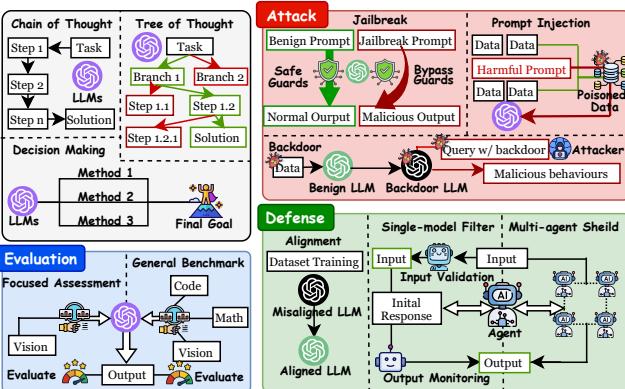


Figure 2: The framework of agent brain’s working mechanisms and its attack-defense-evaluation paradigm.

coercing a “Sure” reply to malicious queries. MRJ-Agent [91] designs a single attack agent to automatically generate covert jailbreak prompts, enhanced by information from multi-round dialogue. Pri-vAgent [69] trains a single LLM as an attack agent through reinforcement learning to generate jailbreak prompts that can induce the target model to leak system prompts or training data. Furthermore, by leveraging multi-agent collaboration, Evil Geniuses [88] and PANDORA [16] construct MAS as attack optimizers to perform role-specific jailbreak on the brain of target agents or MAS. They reinforced jailbreak effectiveness through Red-Blue exercises and MAS multi-step reasoning, respectively. Additionally, interactions between different brains give rise to viral jailbreak. Specifically, AgentSmith [34] and Tan et al. [85] optimize self-replicating images to attack the brain of a single agent and discover an exponential spread of infectious jailbreak at MAS-level.

Prompt Injection embeds malicious prompts to override the original instructions to the agent brain, thereby manipulating its output or behavior [61, 73]. Early research manually crafts instructions and injects them into data which may be retrieved during inference, potentially causing the agent brain to deviate from intended tasks [33]. Later studies automate this process via gradient-based or optimization-based methods [60, 80]. Beyond injecting prompts in text modality, Bagdasaryan et al. [6] explore that multi-modal adversarial perturbations embedded in images or audio can manipulate the agent’s brain to follow the attacker’s instructions. The interaction between the agent’s brain and other modules also provides more surfaces for prompt injection. Zhang et al. [122] exploit various interactions across multiple surfaces to mislead the agent’s brain into performing repetitive or irrelevant actions, inducing logical errors and causing malfunctions.

Backdoor attacks involve the insertion of malicious triggers during training, which will be retrieved during inference, forcing predefined generation in the agent’s brain for specific inputs [141]. Yang et al. [108] categorize agent backdoor attacks into two main types: The first type manipulates the final output distribution, by inserting triggers in different phases (query, observation), while the second one inserts triggers in agent’s thinking process to introduce malicious intermediate reasoning without altering the final output. The first type is exemplified by DemonAgent [143], which fragments the original backdoor into multiple sub-backdoor segments

and employs dynamic encryption, making the backdoor continuously change during execution. BLAST [114] follows the second type and enables “infectious backdoor”, implanting a backdoor in a single agent and using it to influence other agents’ reasoning.

2.1.2 Defenses. The agent brain is not only more complex than a single LLM in terms of interactions, but also faces more types of attacks. Therefore, designing corresponding defenses to ensure its trustworthiness across all dimensions deserves holistic study. Drawing on the defense mechanisms’ operational scopes, we organize defense approaches into three paradigms:

Alignment ensures that LLM agents operate in accordance with human values and ethical principles through fine-tuning, updating reward functions, etc. For instance, some studies embed human values directly into agents through intrinsic reward functions [87] or neuro-symbolic rule-learning frameworks [28], validated in open-world environments like Minecraft. Another work fine-tunes agents with the data generated in social simulation scenarios [71] to align agents with nuanced human behaviors and beliefs. Additionally, iterative alignment through user feedback and belief networks is leveraged to personalize agent behavior and reduce reliance on expert oversight, such as using historical edits [32] or empirically derived human belief structures [128].

Single-model Filter uses an external model to prevent attacks through input and output monitoring on the agent brain. Ayub et al. [5] employ traditional embedding classifiers to detect prompt injection by analyzing input semantic features, while Kwon et al. [50] use a small model for harmful query detection. As for LLM filter, StruQ [13] mitigates prompt injection by converting inputs into structured queries with a specially trained LLM. ShieldLM [134] fine-tunes an LLM safety detector that aligns with safety standards. Additionally, agent serves as an even more powerful filter. For example, GuardAgent [105] and AgentGuard [12] utilize a guardrail agent to protect target agents via safety constraint generation, action checking, and tool-use validation.

Multi-agent Shield leverages collaborative MAS to enhance the trustworthiness of the target agent’s brain. The main difference of these guard MAS lies in their communication pattern. Specifically, one pattern is multi-agent debate, where agents critique the reasoning processes of other components of the brain over multiple rounds to reach a consensus final answer [26]. Apart from debating, Kwartler et al. [49] use a reviewing agent to provide feedback and correct the target agent’s generation. AutoDefense [120] assigns different roles to each agent, allowing smaller models to collaborate and protect the brain of the targeted agent from jailbreak attacks.

2.1.3 Evaluation. Different from static LLMs, agents continuously adapt to their environment, making their behavior highly context-dependent and variable, which also complicates the evaluation. Current evaluation methodologies for agent brain trustworthiness can be categorized along two dimensions:

Focused Assessment assesses the trustworthiness of agent brain when facing certain types of attacks or in specific domains. InjecAgent [121] and AgentDojo [23] benchmark the brain’s vulnerability to *Indirect Prompt Injection*, while DemonAgent [143] introduces AgentBackdoorEval for *Backdoor* scenarios. RedAgent [107] addresses context-aware red teaming, enabling test cases of

context-sensitive *Jailbreak*. In addition to evaluating brain trustworthiness in specific domains, RiskAwareBench [144] introduces an automated framework for assessing *physical risk awareness* in LLM-based embodied agents. RedCode [35] provides a benchmark for evaluating the *code* agents under risky code execution and generation with challenging test cases.

General Benchmark literally involves systematic evaluations to test the agent brain in diverse domains and from multiple dimensions. S-Eval [117] automates multiple dimensional and open-ended safety evaluation for LLMs with four hierarchical levels consisting of eight risk dimensions. While BELLS [25] evaluates LLM Safe-guard with a structured collection of tests, including established failure tests, emerging failure tests, and next-gen architecture tests. Meanwhile, recent efforts have introduced comprehensive benchmarks for LLM agents. Agent-SafetyBench [132] and Agent Security Bench [126] are benchmarks that evaluate LLM agents with multiple scenarios and different types of attack/defense methods. AgentHarm [4] contains 110 explicitly malicious agent tasks to evaluate LLM agent misuse, covering multiple harmful categories, including fraud, cybercrime, and harassment. RJudge [116] benchmarks the trustworthiness of multi-turn interactions based on agent interaction records, with 27 risk scenarios across multiple application categories and risk types.

2.1.4 Insight. Current collaborative attacks in MAS can spread from a single compromised agent's brain to multiple agents' brain modules [34, 114]. Therefore, developing collaborative security mechanisms that enable agents to monitor and validate each other's actions is crucial. For example, a distributed consensus protocol can be implemented, where agents collectively verify and agree on critical decisions before execution. Additionally, current evaluations primarily rely on static datasets. However, given the agent brain module's frequently dynamic interactions with external information, such evaluations are clearly insufficient. To address this, dynamic evaluation mechanisms should be implemented to better simulate real-world scenarios. For example, a continuous learning-based evaluation framework can be designed, where agents are tested in real-time environments with evolving data streams.

2.2 Trustworthy Memory in Retrieval

Memory mechanisms in agent systems are crucial for interaction with the environment and user, categorized into long-term and short-term memory. Long-term memory is often associated with RAG, utilizing vector database to store real-world data for generation tasks, while short-term memory holds real-time interaction history, such as dialogue context or task logs. These mechanisms enhance agent capabilities but also introduce trustworthiness risks. In this section, we analyze the trustworthiness challenges posed by memory, focusing on attacks, defenses, and evaluation strategies. The overall framework is shown in Figure 3.

2.2.1 Attack. We categorize attacks related to memory into three types: Memory Poisoning, Privacy Leakage, and Memory Misuse.

Memory Poisoning refers to attackers injecting malicious data into the *long-term memory* [15, 34, 104, 131, 138, 148], which will be retrieved and then mislead the agent system to generate incorrect outputs, undermining its truthfulness. This attack's danger stems

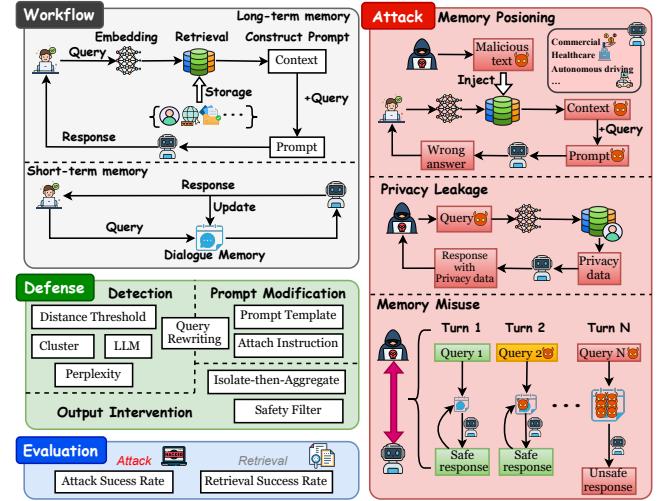


Figure 3: The framework of the agent's memory utilization workflow and its attack-defense-evaluation paradigm.

from its stealthiness: once malicious data is injected into memory, it may continuously influence the agent until detected and removed. Specifically, we summarize the following attack paradigms:

① Memory Injection. PoisonedRAG [148] optimizes malicious text through retrieval and generation conditions, enabling this text to be easily retrieved, while GARAG [20] continuously uses genetic algorithms to optimize adversarial examples for injection. Additionally, RobustRag [104] and Zhong et al. [138] demonstrate that injecting even a small amount of malicious information into vector database can successfully attack agents with high probability.

② Backdoor. AgentPoison [15] optimizes backdoor triggers and attaching them to queries, enhancing retrieval possibility of malicious samples.

Privacy Leakage refers to attackers leveraging the connection between agent and its *long-term memory* to steal the stored private data [3, 53, 119]. Beyond data theft, it enables criminal activities like phishing and identity trafficking, dramatically increasing privacy risks. Technically, we outline the following attack methods:

① Jailbreak. RAG-Thief [44] continuously optimizes anchor queries and adversarial commands, forming jailbreak prompts to conduct data stealing. RAG-MIA [3] designs specific jailbreak templates to extract private data from vector database.

② Embedding Inversion directly restores the original data from its embedding. Specifically,

Morris et al. [67] iteratively optimize the text hypothesis to bring its embedding closer to the target embedding, while Li et al. [55] reconstruct the data via a generative decoder.

Memory Misuse crafts specific query sequences to gradually bypass intrinsic safety alignment via multi-turn interaction, leveraging the storage property of agent *short-term memory* [1, 18, 56, 74, 77, 89]. Therefore, the safety of an agent system is reduced. From a technical perspective, we outline the following two attack methodologies:

① Jailbreak. Agarwal et al. [1] send an initial query along with an attack prompt in the first round, followed by a flattering challenge and a reiteration of the attack prompt in the second round, while Russinovich et al. [77] begin with harmless questions related to the target task and gradually steer harmful generation.

② Backdoor. Tong et al. [89] utilize multi-turn dialogue memory to conceal backdoor triggers, activating malicious responses only when all triggers appear, thus stealthily achieving the attack. As a

core component of agents, defense technologies against short-term memory misuse are even more important.

2.2.2 Defense. To address the aforementioned memory attacks, several corresponding defense strategies have been proposed [1, 3, 64, 103, 139], which can be classified into the following three types:

Detection typically involves identifying and removing harmful portions of text retrieved from long-term memory. TrustRAG [139] employs K-means clustering on the retrieved embeddings to distinguish between clean and potentially malicious documents based on the embedding distribution. Xian et al. [103] calculate the Mahalanobis distance between query and documents (poisoned or clean) retrieved from long-term memory and set a threshold to filter out malicious documents. Besides traditional methods for detection, Agarwal et al. [1] utilize LLMs to detect and filter out parts of the prompt that access private data. Additionally, ASB [131] uses both perplexity-based and LLM-based detection to identify whether the text retrieved from memory is compromised.

Prompt Modification refers to altering the query sent to the agent to make it safer. Anderson et al. [3] embed user queries into a designed prompt template, which enables the LLM to ignore direct requests for querying the content of vector database. Agarwal et al. [1] propose various modification strategies, such as adding security instructions to the original prompt, explicitly requiring the agent not to disclose data; or directly utilize the LLM to rewrite the query and filter out potentially privacy-leaking parts.

Output Intervention refers to intervening in the agent’s output before it generates the final response to prevent it from producing incorrect or unsafe replies. RobustRAG [104] employs an Isolate-then-Aggregate approach that independently generates responses for each retrieved passage and aggregates them via keyword, effectively reduce ASR as malicious passage are in the minority. Chen et al. [11] demonstrate that safety filter can prevent the generation of unsafe tokens in multi-turn dialogues when agents use memory.

2.2.3 Evaluation. Since there is currently no systematic and reliable evaluation of memory, we summarize some commonly used metrics in research. For memory poisoning and privacy leakage attacks, which rely on retrieving relevant data, evaluation metrics extend beyond ASR to include target text retrieval effectiveness. PoisonedRAG [148] uses precision, recall, and F1-score to assess the retrieval effectiveness of malicious text. AgentPoison [15] employs ASR for retrieval (ASR-r) to measure malicious trigger effectiveness. Zeng et al. [119] evaluate attacks by quantifying retrieved private data, and RAG-Thief [44] uses chunk recovery rate (CRR) and semantic similarity (SS) to gauge data theft from vector database.

2.2.4 Insight. Current memory-centric **attack** methods often lack generalization across tasks. Memory poisoning relies on task-specific malicious samples to ensure retrieval by relevant task queries. Memory misuse, on the other hand, heavily depends on dialogue context, requiring tailored designs for various scenarios. From the **defense** perspective, for memory poisoning defense, future defenses should focus on the vector database end to prevent the injection of toxic samples, thereby reducing the defense time during agent responses. For privacy leakage defense, privacy protection mechanisms such as query rewriting or fine-tuning LLMs must be implemented in

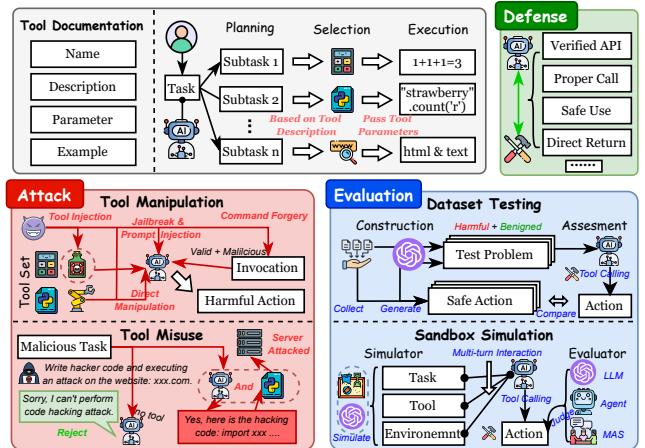


Figure 4: The workflow of agent tool calling with corresponding demonstrations on attack, defense, and evaluation.

applications where privacy breaches are possible. For memory misuse defense, developing a multi-round adversarial dialogue training paradigm for safety alignment is essential to enhance robustness of agent. For the **evaluation**, as there is currently no systematic and reliable benchmarks, we recommend establishing benchmarks for the paradigms of attacks and defenses mentioned above.

2.3 Trustworthy Tool for Action

In the components of the agent or MAS, the tool module serves as the medium for interaction between the system and the external world [59]. It can either gather information from outside (e.g., using a search engine for information retrieval [21]) or reflect the internally decided actions back to the external environment (e.g., mapping UI operations to the interface [123]). The typical forms of tools include API functions, sensors, embodied robots, etc [115]. However, different forms of tools exhibit varied properties and risks, making trustworthiness problems complicated and challenging. Simultaneously, although tools are also used in the context of LLMs [79, 145], the closer integration with agents imposes stricter challenges on their trustworthiness, especially since some agents’ actions can affect the real world. Risks associated with agent tools may lead to more severe negative impacts compared to LLMs [131]. In this section, we present agent tool trustworthiness from the attack, defense, and evaluation perspectives, illustrated in Figure 4.

2.3.1 Attack. The planning, selection, and execution phases of tool invocation expose additional threat interfaces to attackers. Besides, tools enable agents to conduct certain malicious actions. Based on tool’s role in attacks, we summarize the following two paradigms:

Tool Manipulation targets specifically at certain steps in the tool calling process to induce execution of malicious or sensitive behaviors. Specifically and technically, we can further list the following methods:

- 1 Jailbreak:** Cheng et al. [17] manually design jailbreak prompts to extract personal information from the training data of code generation agent. Furthermore, Imprompter [29] and Fu et al. [30] utilize gradient optimization to search for input prompts or images automatically, which induce agents to invoke tools to leak privacy from dialogues or execute destructive actions on users’ resources.
- 2 Prompt Injection:** BreakingAgents [122]

uses human-written prompt injection to achieve malfunction attacks, causing agents to perform repetitive or irrelevant actions, with further exploration on the attack propagation in MAS. **③ Tool Injection:** ToolCommander [92] introduces a two-stage attack strategy: injecting malicious tools to steal user queries first and then manipulating tool selection using the stolen data, achieving privacy theft and denial-of-service attacks. **④ Command Forgery:** AUTOCMD [45] utilizes another LLM, trained on tool calling datasets and enhanced by target-specific examples, to generate and mimic valid commands to deduce sensitive information from the tools. **⑤ Direct Manipulation:** Zhao et al. [136] manipulate the outputs of third-party APIs by injecting malicious content or deleting critical information, leading to incorrect or biased behaviors.

Tool Abuse refers to the attack type that exploits the tool using capabilities of agent or MAS to achieve or enhance attacks on external entities. For example, in web security, Fang et al. [27] explore how agents can autonomously hack websites when equipped with tools, while Fang et al. [27] demonstrate that tool-integrated agents can autonomously exploit one-day vulnerabilities in real-world systems. Besides deliberate guidance, BadAgent [97] and Kumar et.al [48] reveal that backdoor attacks or even refusal-based safety alignment can trigger agents to exploit tools for harmful actions.

2.3.2 Defense. Our investigation reveals that research on defenses against tool-related attacks is notably scarce, whether in the context of LLMs, agents, or MAS. GuardAgent [105] takes the first step to ensure the trustworthiness of target agents by invoking APIs to execute guardrail code for task plans. In addition, AgentGuard [12] autonomously identifies unsafe tool-use workflows via LLM orchestrators and generates safety constraints for secure tool using.

2.3.3 Evaluation. Due to the unique ability of agents to interact with the environment and their chatbot-like characteristics akin to LLMs, we categorize the paradigms for evaluating the trustworthiness of agent tool invocation into two paradigms:

Dataset Testing refers to the static evaluation using adversarial query datasets to test the trustworthiness of agent tool invocations. Specifically, ToolSword [111] evaluates the safety performance of mainstream LLMs during tool invocation via multiple attack methods: malicious queries and jailbreak attacks at the input phase, noisy misdirection and risky cues during execution, and harmful feedback and error conflicts at the output phase. InjectAgent [121] establishes a benchmark via data generation to test agent trustworthiness against indirect prompt injection attacks during tool usage, including datasets containing malicious queries involving tool invocation. AgentHarm [4] constructs behavioral datasets to assess the harmfulness of agents sequentially invoking multiple tools during task execution, finding that jailbreak templates can be adapted to effectively attack agents. PrivacyLens [78] focuses on privacy security assessments during tool execution processes.

Sandbox Simulation describes the dynamic evaluation simulating tool interactions in controlled environments to assess emergent risks, especially in multi-turn queries. For instance, ToolEmu [76] employs an emulator LLM to simulate tool execution and an evaluator LLM to assess trustworthiness and identify potential risks. In addition, HACosystem [140] establishes a sandbox environment through role-playing simulations and scenario-specific checklist evaluations to measure safety risks of tool-equipped AI agents in

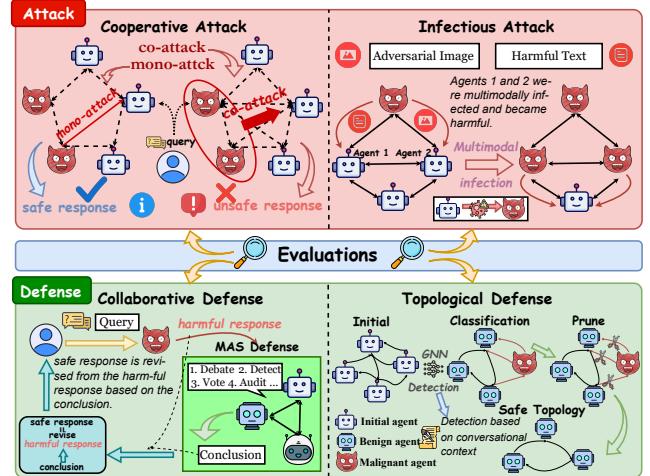


Figure 5: A framework for defining various attack, defense, and evaluation strategies in agent-to-agent interactions.

multi-turn interactions, revealing that LLMs exhibit higher risks when functioning as agents with tool access.

2.3.4 Insight. A significant pain point in the current study on tool trustworthiness is the lack of defense mechanisms. Specifically, future research directions should focus on ensuring the security of Agents or MAS during tool invocation. Potential defensive measures could include reviewing tools before integration or simulating their execution in an isolated environment to verify safety before actual execution. On the other hand, from an attack perspective, research can follow traditional methods (e.g., jailbreak, prompt injection) while also leveraging tool characteristics to propose novel attack vectors, which may be a future research direction. For instance, since tools are often integrated into agent systems as third-party APIs, and there are currently no mechanisms to prevent malicious API providers, attackers could potentially provide legitimate and seemingly harmless tool descriptions that, when executed, lead to malicious outcomes. Furthermore, as agent systems become more complex, tool invocations may form chains, necessitating a shift in attack, defense, and evaluation research from single tool invocations to multiple invocations based on tool chains.

3 Extrinsic Trustworthiness

In this section, we focus on the trustworthiness of external modules interacting with the agent system. During operation, agents continuously engage with the external for purposes such as information gathering or decision execution. We categorize the interactions with the external modules into three types: agent-to-agent (**Sec. 3.1**), agent-to-environment (**Sec. 3.2**), and agent-to-user (**Sec. 3.3**). We first introduce the role of each type and then elaborate on corresponding trustworthiness research, ending up with our insights.

3.1 Agent-to-Agent Interaction

Interactions between agents are crucial to system functioning and take various patterns, such as cooperation, competition and debate [70]. These Agent-to-Agent Interaction, is the key to shaping the system's dynamics. However, they differ significantly in terms of their nature and associated trustworthiness risks [37]. To this

end, we introduce current works related to agent interactions from attacks, defenses, and evaluations, with illustration in Figure 5.

3.1.1 Attack. Unlike the simple information flow of a single agent, agent-agent interactions can not only be used to enhance attacks on a single agent via cooperation but also exploit its propagative nature to induce a trustworthiness crisis at the MAS level.

Cooperative Attack refers to malicious agents collaborateing to compromise the targeted agent-agent interactions, disrupting system safety. In terms of *truthfulness*, Ju et al. [46] coordinate agents in MAS to spread counterfactual and harmful information, amplified by persuasive manipulation. Regarding *robustness*, Agent-in-the-Middle [39] disrupts MAS coordination by using intermediary agent to intercept and manipulate communication, while Amayuelas et al. [2] exploit adversarial persuasion to further weaken system stability. As for *safety*, Evil Geniuses [88] uses compromised agents to create adversarial prompts, refining attack through iterative simulations.

Infectious Attack spreads malicious effects by infecting others or disrupt agents or components within MAS. Prompt Infection [52] and CORBA [142] exploit this attack in text modality. Specifically, Prompt Infection uses silent adversarial diffusion to facilitate data theft and misinformation, while CORBA introduces a self-propagating, topology-independent attack that drains resources. In the context of multi-modality, Agent Smith [34] and Tan et al. [85] extend this attack to image modality, enhancing stealth and contagion, which accelerates the collapse of MAS trustworthiness. Finally, NetSafe [113] analyzes how hallucinations and misinformation propagate across MAS topologies, revealing their structural dependencies and adversarial impacts.

3.1.2 Defense. Similar to Cooperative Attacks, we can enhance defense mechanisms by leveraging the cooperative ability of MAS. Besides, the interactions between agents allows modeling them as a graph, enabling defense strategies from a topological perspective.

Collaborative Defense uses agents cooperation via information sharing for a trusworthy analysis on the target agent response or action. Based on debate pattern, BlockAgents [10] ensures secure coordination through multi-round debate voting with Proof-of-Thought consensus, while Audit-LLM [19, 81] shifts to hallucination, safety and robustness dimension. Except for debating, adversarial techniques are used for defense. For example, AutoDefense [120] enhances safety through adversarial prompt filtering, while LLAMOS [58] establishes dynamic defense mechanisms to foster a robust equilibrium between attacker and defender. PsySafe [135] specifically targets dark property injection attacks, further solidifying the role of MAS in adversarial defense.

Topological Defense leverages network structure to isolate threats and limit their spread and impact. GPTSwarm [146] represents an initial exploration into enhancing MAS robustness through topology optimization. In contrast, G-Safeguard [95] delves deeper into combating adversarial attacks and misinformation in MAS across various topologies, employing graph neural networks (GNNs) to detect anomalies in discourse graphs.

3.1.3 Evaluation. The evaluation of agent-to-agent trustworthiness is still in its infancy with limited number of research. SafeAgent-Bench [112] introduces SafeAgentEnv, which enables multi-agent execution with various actions and baseline models, to evaluate

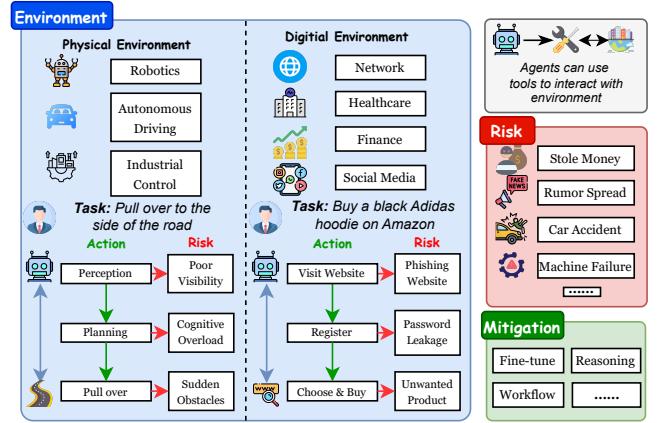


Figure 6: Framework of agent interaction with various environments and enhancement of safety and truthfulness.

defense success rate. R-judge [116] assesses agent interaction safety in multi-turn interactions across several domains and risk scenarios. [7] proposes JAILJUDGE, a benchmark covering synthetic, adversarial, outdoor, and multilingual risks and featuring a manually annotated dataset and multi-agent evaluation.

3.1.4 Insight. We can clearly observe that agent-agent interactions lead to a novel and severe threat to trustworthiness—Infectious Attack. Future research on the attack side could focus on how to automate such attacks against specific agents or MAS. Simultaneously, corresponding anti-propagation defense mechanisms and evaluation methods are also potential research topics. Additionally, some studies treat MAS as (temporal) graphs, with agents as nodes and interactions as directed edges, and have conducted preliminary explorations of topology-based defenses or trustworthiness evaluations. This perspective might also be a point worthy of further investigation. Moreover, we have not found research strictly focused on the trustworthiness evaluation of inter-agent interactions, a gap that could potentially be filled by using LLMs or agents as dynamic detectors and evaluators. Research on the evaluation side may appear more urgent and deserving of attention.

3.2 Agent-to-Environment Interaction

Agents face unique trustworthiness challenges when getting information and taking action in dynamic and heterogeneous environments. These environments, spanning physical and digital domains, require adaptive perception, reasoning, and action for effective deployment. Trustworthy risks, such as autonomous driving errors and network disruptions, are influenced by agent roles and environmental constraints. Given the diversity of dynamic scenarios and related issues, existing solutions to address these challenges are fragmented and lack a systematic framework. Therefore, we adopt an environment-centric classification approach for discussion with Figure 6, rather than technology-centric categorization.

3.2.1 Physical Environment. In this environment, agents are transforming sectors like robotics, autonomous vehicles, and industrial systems, bringing both new opportunities and challenges. Researchers are actively developing innovative solutions to address these challenges, aiming to optimize performance while maintaining the highest truthfulness standards across physical applications.

Robotics: Yang et al. [110] propose a constraint module based on Linear Temporal Logic, which enables safety violation reasoning and explanation, and unsafe action pruning, ensuring that robot agents meet global safety standards for industrial deployment. Additionally, SELP [101] integrates equivalence voting and domain-specific fine-tuning to ensure that robot agents generate safety task plans such as drone navigation and robot manipulation.

Autonomous Driving: Hudson [82] formalizes real-time perception data into natural language with attack detection instructions and analyzes causal reasoning to enhance safety and truthfulness during perception attacks. Furthermore, ChatScene [127] generates safety-critical scenarios for autonomous vehicles by transforming unstructured language instructions into domain-specific code for CARLA platform simulations.

Industrial Control: Vyas et al. [90] propose an agent framework for autonomous industrial control, featuring validation and reprompting architectures that ensure real-time error detection, recovery, and safe decision-making. Agents4PLC [63] automates PLC code generation and ensures truthfulness through rigorous code-level verification, improving correctness and reliability in industrial control systems by integrating RAG and COT.

3.2.2 Digital Environment: Research focusing on the applications of virtual data across digital environments aims to enhance task-specific performance such as improving navigation on the web, diagnostic accuracy in healthcare, and efficiency in financial tasks. However, the vulnerability of data leads to a growing focus on addressing critical security, ethical, and compliance issues.

Network: Fang et al. [27] point out that agents can uncover web vulnerabilities exploited for hacking attacks autonomously. Furthermore, researchers develop frameworks to evaluate the truthfulness of web agents. Debenedetti et al. [22] focus on the evaluation of defense in specific tasks like managing email clients.

Healthcare: For privacy, Xiang et al. [105] propose that LLM agents safeguard sensitive medical data through knowledge-driven reasoning. Regarding robustness, Polaris [68] suggests enhancing the reliability and robustness of LLM agents via multi-agent architectures for real-time patient-AI interactions.

Finance: Chen et al. [14] identify risks such as hallucinations, temporal misalignment, and adversarial vulnerabilities in financial applications for finance agents. Meanwhile, Park et al. [72] demonstrate how LLM agents can enhance anomaly detection via role-playing scenarios like manager-analyst communications, thus contributing significantly to risk management.

Social Media: Jeptoo et al. [42] show that agents can be utilized for fake news detection through multi-agent collaboration and automated workflows. La et al. [51] use agents to simulate the evolution of language patterns that try to evade social media regulations, contributing to content regulation.

3.2.3 Insight. In current research, the environment is often treated as a static backdrop, with safety efforts primarily focused on improving agent action outcomes. However, this approach overlooks the critical issue of trustworthy interactions between agents and environments, which represents a significant point of vulnerability for both attacks and defenses. Consequently, systematic attack and defense mechanisms are necessary for targeting environment-agent interactions to enhance system truthfulness and safety. For

example, adversaries might subtly alter environmental feedback to mislead agents into making suboptimal or unsafe decisions.

Current evaluations are predominantly focused on a limited set of domain-specific safety scenarios, failing to address the broader spectrum of interdisciplinary and cross-domain challenges. To advance the field, developing evaluations that address cross-domain and interdisciplinary safety challenges is essential for progress.

3.3 Agent-to-User Interaction

In this section, we discuss trustworthiness considerations in agent-user interaction, with particular emphasis on the **user perspective** [38]. We elaborate on key challenges in building reliable, human-aligned agent systems. In our survey, the works summarized in previous sections focus largely on agent-to-user interactions, which overlap significantly with the content discussed in this subsection. Therefore, this section will mainly offer discussions and insights.

3.3.1 Discussion. Current research on agent trustworthiness focuses on safety and reliability but overlooks trust mechanisms in **interactions** process. A key gap is how users adjust trust based on agent behavior. While personalization boosts engagement, it also risks manipulation, requiring a balance with system robustness. Transparency is essential, as clear explanations help users manage risks. As agents evolve, they enhance trust while protecting user data. However, most studies address single-agent interactions, leaving multi-agent trust dynamics largely unexplored.

3.3.2 Insight. To advance research on trust in agent-to-user interactions, future work should focus on developing adaptive trust calibration frameworks that enable users to dynamically adjust trust thresholds based on real-time interactions. Feedback mechanisms should be optimized to reinforce beneficial agent behaviors while preventing trust erosion due to errors or biases. Additionally, a unified paradigm can be established to regulate and ensure fairness and reliability across different users and personalized agents.

A promising direction to improve transparency is by developing explainable agents that provide decisions and interpretable reasons to users. In the context of multi-agent, maintaining trust requires innovative monitoring—using supervisors agents to oversee interactions with user and ensure consistent responses.

4 Conclusion

In this survey, we introduce the six key components of agent systems, both internal and external. Focusing on each module, we summarize current research on trustworthy agents across multiple dimensions such as safety and privacy, from the perspectives of attack, defense, and evaluation. We summarize and define the typical method paradigm in each perspective to standardize and inspire further study. Notably, we identify several research and technical gaps in existing works, such as defense mechanisms for trustworthy tool invocation and the evaluation of trustworthiness in memory and agent interactions. Based on this, we highlight future research directions and key insights about the trustworthiness of agent and MAS, emphasizing their distinctive attributes and implications. We believe that this survey may encourage and inspire further research and exploration of trustworthy agent by both researchers and developers.

References

- [1] Divyansh Agarwal, Alexander R Fabbri, Ben Risher, Philippe Laban, Shafiq Joty, and Chien-Sheng Wu. 2024. Prompt Leakage effect and defense strategies for multi-turn LLM interactions. *arXiv preprint arXiv:2404.16251* (2024).
- [2] Alfonso Amayuelas, Xianjun Yang, Antonis Antoniades, Wenyue Hua, Liangming Pan, and William Wang. 2024. Multiagent collaboration attack: Investigating adversarial attacks in large language model collaborations via debate. *arXiv preprint arXiv:2406.14711* (2024).
- [3] Maya Anderson, Guy Amit, and Abigail Goldstein. 2024. Is my data in your retrieval database? membership inference attacks against retrieval augmented generation. *arXiv preprint arXiv:2405.20446* (2024).
- [4] Maksym Andriushchenko, Alexandra Souly, Mateusz Dziemian, Derek Dueñas, Maxwell Lin, Justin Wang, Dan Hendrycks, Andy Zou, Zico Kolter, Matt Fredrikson, et al. 2024. Agenthamr: A benchmark for measuring harmfulness of llm agents. *arXiv preprint arXiv:2410.09024* (2024).
- [5] Md Ahsan Ayub and Subhabrata Majumdar. 2024. Embedding-based classifiers can detect prompt injection attacks. *arXiv preprint arXiv:2410.22284* (2024).
- [6] Eugena Bagdasaryan, Tsung-Yin Hsieh, Ben Nassi, and Vitaly Shmatikov. 2023. Abusing images and sounds for indirect instruction injection in multi-modal LLMs. *arXiv preprint arXiv:2307.10490* (2023).
- [7] JUDGE BENCHMARK [n. d.]. JAILJUDGE: AComprehensive JAILBREAK JUDGE BENCHMARK WITH MULTI-AGENT ENHANCED EXPLANATION EVALUATION FRAMEWORK. ([n. d.]).
- [8] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201* (2023).
- [9] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kajie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology* 15, 3 (2024), 1–45.
- [10] Bei Chen, Gaolei Li, Xi Lin, Zheng Wang, and Jianhua Li. 2024. BlockAgents: Towards Byzantine-Robust LLM-Based Multi-Agent Coordination via Blockchain. In *Proceedings of the ACM Turing Award Celebration Conference-China 2024*. 187–192.
- [11] Bocheng Chen, Guangjing Wang, Hanqing Guo, Yuanda Wang, and Qiben Yan. 2023. Understanding multi-turn toxic behaviors in open-domain chatbots. In *Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses*. 282–296.
- [12] Jizhou Chen and Samuel Lee Cong. 2025. AgentGuard: Repurposing Agen-tic Orchestrator for Safety Evaluation of Tool Orchestration. *arXiv preprint arXiv:2502.09809* (2025).
- [13] Sizhe Chen, Julien Piet, Chawin Sitawarin, and David Wagner. 2024. Struq: Defending against prompt injection with structured queries. *arXiv preprint arXiv:2402.06363* (2024).
- [14] Zichen Chen, Jiaao Chen, Jianda Chen, and Misha Sra. 2025. Position: Standard Benchmarks Fail—LLM Agents Present Overlooked Risks for Financial Applications. *arXiv preprint arXiv:2502.15865* (2025).
- [15] Zhaorun Chen, Zhen Xiang, Chaowei Xiao, Dawn Song, and Bo Li. 2025. Agent-poison: Red-teaming llm agents via poisoning memory or knowledge bases. *Advances in Neural Information Processing Systems* 37 (2025), 130185–130213.
- [16] Zhaorun Chen, Zhuokai Zhao, Wenjie Qu, Zichen Wen, Zhiqiang Han, Zhihong Zhu, Jiaheng Zhang, and Huaxiu Yao. 2024. Pandora: Detailed llm jailbreaking via collaborated phishing agents with decomposed reasoning. In *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*.
- [17] Wen Cheng, Ke Sun, Xinyu Zhang, and Wei Wang. 2024. Security Attacks on LLM-based Code Completion Tools. *arXiv preprint arXiv:2408.11006* (2024).
- [18] Yixin Cheng, Markos Georgopoulos, Volkan Cevher, and Grigoris G Chrysos. 2024. Leveraging the context through multi-round interactions for jailbreaking attacks. *arXiv preprint arXiv:2402.09177* (2024).
- [19] Steffi Chern, Zhen Fan, and Andy Liu. 2024. Combating adversarial attacks with multi-agent debate. *arXiv preprint arXiv:2401.05998* (2024).
- [20] Sukmin Cho, Soyeong Jeong, Jeongyeon Seo, Taeho Hwang, and Jong C Park. 2024. Typo that Broke the RAG’s Back: Genetic Attack on RAG Pipeline by Simulating Documents in the Wild via Low-level Perturbations. *arXiv preprint arXiv:2404.13948* (2024).
- [21] Gobinda Chowdhury and Sudatta Chowdhury. 2024. AI-and LLM-driven search tools: A paradigm shift in information access for education and research. *Journal of Information Science* (2024), 01655515241284046.
- [22] 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* (2024).
- [23] Edoardo Debenedetti, Jie Zhang, Mislav Balunovic, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. 2025. Agentdojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. *Advances in Neural Information Processing Systems* 37 (2025), 82895–82920.
- [24] Zehang Deng, Yongjian Guo, Changzhou Han, Wanlun Ma, Junwu Xiong, Sheng Wen, and Yang Xiang. 2024. Ai agents under threat: A survey of key security challenges and future pathways. *Comput. Surveys* (2024).
- [25] Diego Dorn, Alexandre Variengien, Charbel-Raphaël Segéry, and Vincent Corruble. 2024. Bells: A framework towards future proof benchmarks for the evaluation of llm safeguards. *arXiv preprint arXiv:2406.01364* (2024).
- [26] Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*.
- [27] Richard Fang, Rohan Bindu, Akul Gupta, Qiusi Zhan, and Daniel Kang. 2024. Llm agents can autonomously hack websites. *arXiv preprint arXiv:2402.06664* (2024).
- [28] Ivar Frisch and Mario Giulianelli. 2024. LLM agents in interaction: Measuring personality consistency and linguistic alignment in interacting populations of large language models. *arXiv preprint arXiv:2402.02896* (2024).
- [29] Xiaohan Fu, Shuheng Li, Zihan Wang, Yihao Liu, Rajesh K Gupta, Taylor Berg-Kirkpatrick, and Earlene Fernandes. 2024. Imprompter: Tricking LLM Agents into Improper Tool Use. *arXiv preprint arXiv:2410.14923* (2024).
- [30] Xiaohan Fu, Zihan Wang, Shuheng Li, Rajesh K Gupta, Niloofer Mireshghal-lah, Taylor Berg-Kirkpatrick, and Earlene Fernandes. 2023. Misusing tools in large language models with visual adversarial examples. *arXiv preprint arXiv:2310.03185* (2023).
- [31] Yuyou Gan, Yong Yang, Zhe Ma, Ping He, Rui Zeng, Yiming Wang, Qingming Li, Chunyi Zhou, Songze Li, Ting Wang, et al. 2024. Navigating the risks: A survey of security, privacy, and ethics threats in llm-based agents. *arXiv preprint arXiv:2411.09523* (2024).
- [32] Ge Gao, Alexey Taymanov, Eduardo Salinas, Paul Mineiro, and Dipendra Misra. 2024. Aligning llm agents by learning latent preference from user edits. *arXiv preprint arXiv:2404.15269* (2024).
- [33] Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. 2023. Not what you’ve signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*. 79–90.
- [34] Xiangming Gu, Xiaosen Zheng, Tianyu Pang, Chao Du, Qian Liu, Ye Wang, Jing Jiang, and Min Lin. 2024. Agent smith: A single image can jailbreak one million multimodal llm agents exponentially fast. *arXiv preprint arXiv:2402.08567* (2024).
- [35] Chengquan Guo, Xun Liu, Chulin Xie, Andy Zhou, Yi Zeng, Zinan Lin, Dawn Song, and Bo Li. 2025. RedCode: Risky Code Execution and Generation Benchmark for Code Agents. *Advances in Neural Information Processing Systems* 37 (2025), 106190–106236.
- [36] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680* (2024).
- [37] Lewis Hammond, Alan Chan, Jesse Clifton, Jason Hoelscher-Obermaier, Akbir Khan, Euan McLean, Chandler Smith, Wolfram Barfuss, Jakob Foerster, Tomáš Gavenčík, et al. 2025. Multi-Agent Risks from Advanced AI. *arXiv preprint arXiv:2502.14143* (2025).
- [38] Feng He, Tianqiang Zhu, Dayong Ye, Bo Liu, Wanlei Zhou, and Philip S Yu. 2024. The emerged security and privacy of llm agent: A survey with case studies. *arXiv preprint arXiv:2407.19354* (2024).
- [39] Pengfei He, Yupin Lin, Shen Dong, Han Xu, Yue Xing, and Hui Liu. 2025. Red-Teaming LLM Multi-Agent Systems via Communication Attacks. *arXiv preprint arXiv:2502.14847* (2025).
- [40] Wenyue Hua, Xianjun Yang, Mingyu Jin, Zelong Li, Wei Cheng, Ruixiang Tang, and Yongfeng Zhang. 2024. Trustagent: Towards safe and trustworthy llm-based agents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 10000–10016.
- [41] Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, et al. 2024. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561* (2024).
- [42] Korir Nancy Jeptoo and Chengjie Sun. 2024. Enhancing Fake News Detection with Large Language Models Through Multi-agent Debates. In *CCF International Conference on Natural Language Processing and Chinese Computing*. Springer, 474–486.
- [43] Xiaojun Jia, Tianyu Pang, Chao Du, Yihao Huang, Jindong Gu, Yang Liu, Xi-aochun Cao, and Min Lin. 2024. Improved techniques for optimization-based jailbreaking on large language models. *arXiv preprint arXiv:2405.21018* (2024).
- [44] Changyue Jiang, Xudong Pan, Geng Hong, Chenfu Bao, and Min Yang. 2024. Rag-thief: Scalable extraction of private data from retrieval-augmented generation applications with agent-based attacks. *arXiv preprint arXiv:2411.14110* (2024).
- [45] Ziyou Jiang, Mingyang Li, Guowei Yang, Junjie Wang, Yuekai Huang, Zhiyuan Chang, and Qing Wang. 2025. Mimicking the Familiar: Dynamic Command Generation for Information Theft Attacks in LLM Tool-Learning System. *arXiv preprint arXiv:2502.11358* (2025).

- [46] Tianjie Ju, Yiting Wang, Xinbei Ma, Pengzhou Cheng, Haodong Zhao, Yulong Wang, Lifeng Liu, Jian Xie, Zhuosheng Zhang, and Gongshen Liu. 2024. Flooding spread of manipulated knowledge in llm-based multi-agent communities. *arXiv preprint arXiv:2407.07791* (2024).
- [47] Aounon Kumar, Chirag Agarwal, Suraj Srinivas, Aaron Jiaxun Li, Soheil Feizi, and Himabindu Lakkaraju. 2023. Certifying llm safety against adversarial prompting. *arXiv preprint arXiv:2309.02705* (2023).
- [48] Priyanshu Kumar, Elaine Lau, Saranya Vijayakumar, Tu Trinh, Scale Red Team, Elaine Chang, Vaughn Robinson, Sean Hendryx, Shuyan Zhou, Matt Fredrikson, et al. 2024. Refusal-trained llms are easily jailbroken as browser agents. *arXiv preprint arXiv:2410.13886* (2024).
- [49] Ted Kwartler, Matthew Berman, and Alan Aqrawi. 2024. Good Parenting is all you need—Multi-agentic LLM Hallucination Mitigation. *arXiv preprint arXiv:2410.14262* (2024).
- [50] Ohjoon Kwon, Donghyeon Jeon, Nayoung Choi, Gyu-Hwung Cho, Changbong Kim, Hyunwoo Lee, Inho Kang, Sun Kim, and Taiwoo Park. 2024. SLM as Guardian: Pioneering AI Safety with Small Language Models. *arXiv preprint arXiv:2405.19795* (2024).
- [51] Lucio La Cava and Andrea Tagarelli. 2024. Safeguarding Decentralized Social Media: LLM Agents for Automating Community Rule Compliance. *arXiv preprint arXiv:2409.08963* (2024).
- [52] Donghyun Lee and Mo Tiwari. 2024. Prompt infection: Llm-to-llm prompt injection within multi-agent systems. *arXiv preprint arXiv:2410.07283* (2024).
- [53] Ang Li, Yin Zhou, Vethavikashini Chithrra Raguram, Tom Goldstein, and Micah Goldblum. 2025. Commercial LLM Agents Are Already Vulnerable to Simple Yet Dangerous Attacks. *arXiv preprint arXiv:2502.08586* (2025).
- [54] Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbulin, and Bernard Ghanem. 2023. Camel: Communicative agents for "mind" exploration of large language model society. *Advances in Neural Information Processing Systems* 36 (2023), 51991–52008.
- [55] Haoran Li, Mingshi Xu, and Yangqiu Song. 2023. Sentence embedding leaks more information than you expect: Generative embedding inversion attack to recover the whole sentence. *arXiv preprint arXiv:2305.03010* (2023).
- [56] Nathaniel Li, Ziwen Han, Ian Steneker, Willow Primack, Riley Goodside, Hugh Zhang, Zifan Wang, Cristina Menghini, and Summer Yue. 2024. Llm defenses are not robust to multi-turn human jailbreaks yet. *arXiv preprint arXiv:2408.15221* (2024).
- [57] Yuanchun Li, Hao Wen, Weijun Wang, Xiangyu Li, Yizhen Yuan, Guohong Liu, Jiacheng Liu, Wenxing Xu, Xiang Wang, Yi Sun, et al. 2024. Personal llm agents: Insights and survey about the capability, efficiency and security. *arXiv preprint arXiv:2401.05459* (2024).
- [58] Guang Lin and Qibin Zhao. 2024. Large Language Model Sentinel: LLM Agent for Adversarial Purification. *arXiv preprint arXiv:2405.20770* (2024).
- [59] Bing Liu, Zhou Jianxiang, Dan Meng, and Haonan Lu. 2024. An Evaluation Mechanism of LLM-based Agents on Manipulating APIs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 4649–4662.
- [60] Xiaogeng Liu, Zhiyuan Yu, Yizhe Zhang, Ning Zhang, and Chaowei Xiao. 2024. Automatic and universal prompt injection attacks against large language models. *arXiv preprint arXiv:2403.04957* (2024).
- [61] Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Zihao Wang, Xiaofeng Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yan Zheng, et al. 2023. Prompt Injection attack against LLM-integrated Applications. *arXiv preprint arXiv:2306.05499* (2023).
- [62] Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023. Trustworthy LLMs: A survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374* (2023).
- [63] Zihan Liu, Ruinan Zeng, Dongxia Wang, Gengyun Peng, Jingyi Wang, Qiang Liu, Peiyu Liu, and Wenhui Wang. 2024. Agents4PLC: Automating Closed-loop PLC Code Generation and Verification in Industrial Control Systems using LLM-based Agents. *arXiv preprint arXiv:2410.14209* (2024).
- [64] Junyuan Mao, Fanci Meng, Yifan Duan, Miao Yu, Xiaojun Jia, Junfeng Fang, Yuxuan Liang, Kun Wang, and Qingsong Wen. 2025. AgentSafe: Safeguarding Large Language Model-based Multi-agent Systems via Hierarchical Data Management. *arXiv preprint arXiv:2503.04392* (2025).
- [65] Tula Masterman, Sandi Besen, Mason Sawtell, and Alex Chao. 2024. The landscape of emerging ai agent architectures for reasoning, planning, and tool calling: A survey. *arXiv preprint arXiv:2404.11584* (2024).
- [66] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196* (2024).
- [67] John X Morris, Volodymyr Kuleshov, Vitaly Shmatikov, and Alexander M Rush. 2023. Text embeddings reveal (almost) as much as text. *arXiv preprint arXiv:2310.06816* (2023).
- [68] Subhabrata Mukherjee, Paul Gamble, Markel Sanz Ausin, Neel Kant, Kriti Aggarwal, Neha Manjunath, Debajyoti Datta, Zhengliang Liu, Jiayuan Ding, Sophia Busacca, et al. 2024. Polaris: A safety-focused llm constellation architecture for healthcare. *arXiv preprint arXiv:2403.13313* (2024).
- [69] Yuzhou Nie, Zhun Wang, Ye Yu, Xian Wu, Xuandong Zhao, Wenbo Guo, and Dawn Song. 2024. PrivAgent: Agentic-based Red-teaming for LLM Privacy Leakage. *arXiv preprint arXiv:2412.05734* (2024).
- [70] Liviu Panait and Sean Luke. 2005. Cooperative multi-agent learning: The state of the art. *Autonomous agents and multi-agent systems* 11 (2005), 387–434.
- [71] Xianghe Pang, Shuo Tang, Rui Ye, Yuxin Xiong, Bolun Zhang, Yanfeng Wang, and Siheng Chen. 2024. Self-alignment of large language models via multi-agent social simulation. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.
- [72] Taejin Park. 2024. Enhancing anomaly detection in financial markets with an llm-based multi-agent framework. *arXiv preprint arXiv:2403.19735* (2024).
- [73] Fábio Perez and Ian Ribeiro. 2022. Ignore previous prompt: Attack techniques for language models. *arXiv preprint arXiv:2211.09527* (2022).
- [74] Aman Priyanshu and Supriti Vijay. 2024. FRACTURED-SORRY-Bench: Framework for Revealing Attacks in Conversational Turns Undermining Refusal Efficacy and Defenses over SORRY-Bench (Automated Multi-shot Jailbreaks). *arXiv preprint arXiv:2408.16163* (2024).
- [75] Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Xuanhe Zhou, Yufei Huang, Chaojun Xiao, et al. 2024. Tool learning with foundation models. *Comput. Surveys* 57, 4 (2024), 1–40.
- [76] 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* (2023).
- [77] Mark Russinovich, Ahmed Salem, and Ronen Eldan. 2024. Great, now write an article about that: The crescendo multi-turn llm jailbreak attack. *arXiv preprint arXiv:2404.01833* (2024).
- [78] Yijia Shao, Tianshi Li, Weiyan Shi, Yanchen Liu, and Diyi Yang. 2025. Privacylens: Evaluating privacy norm awareness of language models in action. *Advances in Neural Information Processing Systems* 37 (2025), 89373–89407.
- [79] Zhuocheng Shen. 2024. Llm with tools: A survey. *arXiv preprint arXiv:2409.18807* (2024).
- [80] 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*. 660–674.
- [81] Chengyu Song, Linru Ma, Jianming Zheng, Jinzhili Liao, Hongyu Kuang, and Lin Yang. 2024. Audit-LLM: Multi-Agent Collaboration for Log-based Insider Threat Detection. *arXiv preprint arXiv:2408.08902* (2024).
- [82] Ruoyu Song, Muslim Ozgur, Hyungsuk Kim, Antonio Bianchi, and Z Berkay Celik. 2024. Enhancing llm-based autonomous driving agents to mitigate perception attacks. *arXiv preprint arXiv:2409.14488* (2024).
- [83] Bolun Sun, Yifan Zhou, and Haiyun Jiang. 2024. Empowering Users in Digital Privacy Management through Interactive LLM-Based Agents. *arXiv preprint arXiv:2410.11906* (2024).
- [84] Yashar Talebirad and Amirhossein Nadiri. 2023. Multi-agent collaboration: Harnessing the power of intelligent llm agents. *arXiv preprint arXiv:2306.03314* (2023).
- [85] Zhen Tan, Chengshuai Zhao, Raha Moraffah, Yifan Li, Yu Kong, Tianlong Chen, and Huan Liu. 2024. The wolf within: Covert injection of malice into mllm societies via an mllm operative. *arXiv preprint arXiv:2402.14859* (2024).
- [86] Xiangru Tang, Qiao Jin, Kunlun Zhu, Tongxin Yuan, Yichi Zhang, Wangchunshu Zhou, Meng Qu, Yilun Zhao, Jian Tang, Zhuseng Zhang, et al. 2024. Prioritizing safeguarding over autonomy: Risks of llm agents for science. *arXiv preprint arXiv:2402.04247* (2024).
- [87] Elizabeth Tennant, Stephen Hailes, and Mirco Musolesi. 2024. Moral Alignment for LLM Agents. *arXiv preprint arXiv:2410.01639* (2024).
- [88] Yu Tian, Xiao Yang, Jingyuan Zhang, Yinping Dong, and Hang Su. 2023. Evil geniuses: Delving into the safety of llm-based agents. *arXiv preprint arXiv:2311.11855* (2023).
- [89] Terry Tong, Jiashu Xu, Qin Liu, and Muhaoo Chen. 2024. Securing Multi-turn Conversational Language Models From Distributed Backdoor Triggers. *arXiv preprint arXiv:2407.04151* (2024).
- [90] Javal Vyas and Mehmet Mercangöz. 2024. Autonomous Industrial Control using an Agentic Framework with Large Language Models. *arXiv preprint arXiv:2411.05904* (2024).
- [91] Fengxiang Wang, Ranjie Duan, Peng Xiao, Xiaojun Jia, YueFeng Chen, Chongwen Wang, Jialing Tao, Hang Su, Jun Zhu, and Hui Xue. 2024. Mrj-agent: An effective jailbreak agent for multi-round dialogue. *arXiv preprint arXiv:2411.03814* (2024).
- [92] Haowei Wang, Rupeng Zhang, Junjie Wang, Mingyang Li, Yuekai Huang, Dandan Wang, and Qing Wang. 2024. From Allies to Adversaries: Manipulating LLM Tool-Calling through Adversarial Injection. *arXiv preprint arXiv:2412.10198* (2024).
- [93] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science* 18, 6 (2024), 186345.

- [94] Shilong Wang, Guibin Zhang, Miao Yu, Guancheng Wan, Fanci Meng, Chongye Guo, Kun Wang, and Yang Wang. 2025. G-Safeguard: A Topology-Guided Security Lens and Treatment on LLM-based Multi-agent Systems. *arXiv:2502.11127 [cs.CR]* <https://arxiv.org/abs/2502.11127>
- [95] Shilong Wang, Guibin Zhang, Miao Yu, Guancheng Wan, Fanci Meng, Chongye Guo, Kun Wang, and Yang Wang. 2025. G-Safeguard: A Topology-Guided Security Lens and Treatment on LLM-based Multi-agent Systems. *arXiv preprint arXiv:2502.11127* (2025).
- [96] Yuntao Wang, Yanghe Pan, Quan Zhao, Yi Deng, Zhou Su, Linkang Du, and Tom H Luan. 2024. Large Model Agents: State-of-the-Art, Cooperation Paradigms, Security and Privacy, and Future Trends. *arXiv preprint arXiv:2409.14457* (2024).
- [97] Yifei Wang, Dizhan Xue, Shengjie Zhang, and Shengsheng Qian. 2024. Badagent: Inserting and activating backdoor attacks in llm agents. *arXiv preprint arXiv:2406.03007* (2024).
- [98] Zihao Wang, Shaofei Cai, Guanzhou Chen, Anji Liu, Xiaojian Shawn Ma, and Yitao Liang. 2023. Describe, explain, plan and select: interactive planning with llms enables open-world multi-task agents. *Advances in Neural Information Processing Systems* 36 (2023), 34153–34189.
- [99] Zijun Wang, Haoqin Tu, Jieru Mei, Bingchen Zhao, Yisen Wang, and Cihang Xie. 2024. AttnGGC: Enhancing jailbreaking attacks on LLMs with attention manipulation. *arXiv preprint arXiv:2410.09040* (2024).
- [100] 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* (2023).
- [101] Yi Wu, Zikang Xiong, Yiran Hu, Shreyash S Iyengar, Nan Jiang, Aniket Bera, Lin Tan, and Suresh Jagannathan. 2024. SELP: Generating safe and efficient task plans for robot agents with large language models. *arXiv preprint arXiv:2409.19471* (2024).
- [102] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2025. The rise and potential of large language model based agents: A survey. *Science China Information Sciences* 68, 2 (2025), 121101.
- [103] Xun Xian, Ganghua Wang, Xuan Bi, Jayanth Srinivas, Ashish Kundu, Charles Fleming, Mingyi Hong, and Jie Ding. 2024. On the Vulnerability of Applying Retrieval-Augmented Generation within Knowledge-Intensive Application Domains. *arXiv preprint arXiv:2409.17275* (2024).
- [104] Chong Xiang, Tong Wu, Zexuan Zhong, David Wagner, Danqi Chen, and Prateek Mittal. 2024. Certifiably robust rag against retrieval corruption. *arXiv preprint arXiv:2405.15556* (2024).
- [105] Zhen Xiang, Linzhi Zheng, Yanjie Li, Junyuan Hong, Qinbin Li, Han Xie, Jiawei Zhang, Zidi Xiong, Chulin Xie, Carl Yang, et al. 2024. Guardagent: Safeguard llm agents by a guard agent via knowledge-enabled reasoning. *arXiv preprint arXiv:2406.09187* (2024).
- [106] Junlin Xie, Zhihong Chen, Ruifei Zhang, Xiang Wan, and Guanbin Li. 2024. Large multimodal agents: A survey. *arXiv preprint arXiv:2402.15116* (2024).
- [107] Huiyu Xu, Wenhui Zhang, Zhibo Wang, Feng Xiao, Rui Zheng, Yunhe Feng, Zhongjie Ba, and Kui Ren. 2024. Redagent: Red teaming large language models with context-aware autonomous language agent. *arXiv preprint arXiv:2407.16667* (2024).
- [108] Wenkai Yang, Xiaohan Bi, Yankai Lin, Sishuo Chen, Jie Zhou, and Xu Sun. 2024. Watch out for your agents! investigating backdoor threats to llm-based agents. *arXiv preprint arXiv:2402.11208* (2024).
- [109] Yijun Yang, Tianyi Zhou, Kanxue Li, Dapeng Tao, Lusong Li, Li Shen, Xiaodong He, Jing Jiang, and Yuhu Shi. 2024. Embodied multi-modal agent trained by an llm from a parallel textworld. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26275–26285.
- [110] Ziyi Yang, Shreyas S Raman, Ankit Shah, and Stefanie Tellex. 2024. Plug in the safety chip: Enforcing constraints for llm-driven robot agents. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 14435–14442.
- [111] Junjie Ye, Sixian Li, Guanyu Li, Caishuang Huang, Songyang Gao, Yilong Wu, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Toolsword: Unveiling safety issues of large language models in tool learning across three stages. *arXiv preprint arXiv:2402.10753* (2024).
- [112] Sheng Yin, Xianghe Pang, Yuanzhuo Ding, Menglan Chen, Yutong Bi, Yichen Xiong, Wenhao Huang, Zhen Xiang, Jing Shao, and Siheng Chen. 2024. SafeAgentBench: A Benchmark for Safe Task Planning of Embodied LLM Agents. *arXiv preprint arXiv:2412.13178* (2024).
- [113] 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* (2024).
- [114] Yinbo Yu, Saihao Yan, Xueyu Yin, Jing Fang, and Jiajia Liu. 2025. BLAST: A Stealthy Backdoor Leverage Attack against Cooperative Multi-Agent Deep Reinforcement Learning based Systems. *arXiv preprint arXiv:2501.01593* (2025).
- [115] Siyu Yuan, Kaitao Song, Jiangjie Chen, Xu Tan, Yongliang Shen, Ren Kan, Dongsheng Li, and Deqing Yang. 2024. Easytool: Enhancing llm-based agents with concise tool instruction. *arXiv preprint arXiv:2401.06201* (2024).
- [116] Tongxin Yuan, Zhiwei He, Lingzhong Dong, Yiming Wang, Ruijie Zhao, Tian Xia, Lizhen Xu, Binglin Zhou, Fangqi Li, Zhusheng Zhang, et al. 2024. R-judge: Benchmarking safety risk awareness for llm agents. *arXiv preprint arXiv:2401.10019* (2024).
- [117] Xiaohan Yuan, Jinfeng Li, Dongxia Wang, Yuefeng Chen, Xiaofeng Mao, Longtao Huang, Hui Xue, Wenhui Wang, Kui Ren, and Jingyi Wang. 2024. S-eval: Automatic and adaptive test generation for benchmarking safety evaluation of large language models. *arXiv preprint arXiv:2405.14191* (2024).
- [118] Yanwei Yue, Guibin Zhang, Boyang Liu, Guancheng Wan, Kun Wang, Dawei Cheng, and Yiyuan Qi. 2025. MasRouter: Learning to Route LLMs for Multi-Agent Systems. *arXiv:2502.11133 [cs.LG]* <https://arxiv.org/abs/2502.11133>
- [119] Shenglai Zeng, Jiankun Zhang, Pengfei He, Yu Xing, Yiding Liu, Han Xu, Jie Ren, Shuaiqiang Wang, Dawei Yin, Yi Chang, et al. 2024. The good and the bad: Exploring privacy issues in retrieval-augmented generation (rag). *arXiv preprint arXiv:2402.16893* (2024).
- [120] Yifan Zeng, Yiran Wu, Xiao Zhang, Huazheng Wang, and Qingyun Wu. 2024. Autodefense: Multi-agent llm defense against jailbreak attacks. *arXiv preprint arXiv:2403.04783* (2024).
- [121] 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* (2024).
- [122] Boyang Zhang, Yicong Tan, Yun Shen, Ahmed Salem, Michael Backes, Savvas Zannettou, and Yang Zhang. 2024. Breaking agents: Compromising autonomous llm agents through malfunction amplification. *arXiv preprint arXiv:2407.20859* (2024).
- [123] Chaoyun Zhang, Liquan Li, Shilin He, Xu Zhang, Bo Qiao, Si Qin, Minghua Ma, Yu Kang, Qingwei Lin, Saravan Rajmohan, et al. 2024. Ufo: A ui-focused agent for windows os interaction. *arXiv preprint arXiv:2402.07939* (2024).
- [124] Guibin Zhang, Kaijie Chen, Guancheng Wan, Heng Chang, Hong Cheng, Kun Wang, Shuyue Hu, and Lei Bai. 2025. EvoFlow: Evolving Diverse Agentic Workflows On The Fly. *arXiv preprint arXiv:2502.07373* (2025).
- [125] Guibin Zhang, Luyang Niu, Junfeng Fang, Kun Wang, Lei Bai, and Xiang Wang. 2025. Multi-agent Architecture Search via Agentic Supernet. *arXiv preprint arXiv:2502.04180* (2025).
- [126] Hanrong Zhang, Jingyuan Huang, Kai Mei, Yifei Yao, Zhenting Wang, Chenlu Zhan, Hongwei Wang, and Yongfeng Zhang. 2024. Agent security bench (asb): Formalizing and benchmarking attacks and defenses in llm-based agents. *arXiv preprint arXiv:2410.02644* (2024).
- [127] Jiawei Zhang, Chejian Xu, and Bo Li. 2024. Chatscene: Knowledge-enabled safety-critical scenario generation for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15459–15469.
- [128] Shiyao Zhang, Yuji Dong, Yichuan Zhang, Terry R Payne, and Jie Zhang. 2024. Large Language Model Assisted Multi-Agent Dialogues for Ontology Alignment. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, 2594–2596.
- [129] Shuning Zhang, Lyumanshan Ye, Xin Yi, Jingyu Tang, Bo Shui, Haobin Xing, Pengfei Liu, and Hewu Li. 2024. "Ghost of the past": identifying and resolving privacy leakage from LLM's memory through proactive user interaction. *arXiv preprint arXiv:2410.14931* (2024).
- [130] Zeyu Zhang, Xiaobo Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhui Dong, and Ji-Rong Wen. 2024. A survey on the memory mechanism of large language model based agents. *arXiv preprint arXiv:2404.13501* (2024).
- [131] Zhexin Zhang, Shiyao Cui, Yida Lu, Jingzhou Zhou, Junxiao Yang, Hongning Wang, and Minlie Huang. 2024. Agent-SafetyBench: Evaluating the Safety of LLM Agents. *arXiv preprint arXiv:2412.14470* (2024).
- [132] Zhexin Zhang, Shiyao Cui, Yida Lu, Jingzhou Zhou, Junxiao Yang, Hongning Wang, and Minlie Huang. 2024. Agent-SafetyBench: Evaluating the Safety of LLM Agents. *arXiv preprint arXiv:2412.14470* (2024).
- [133] Zhiping Zhang, Bingcan Guo, and Tianshi Li. 2024. Privacy Leakage Overshadowed by Views of AI: A Study on Human Oversight of Privacy in Language Model Agent. *arXiv preprint arXiv:2411.01344* (2024).
- [134] Zhexin Zhang, Yida Lu, Jingyuan Ma, Di Zhang, Rui Li, Pei Ke, Hao Sun, Lei Sha, Zhipang Sui, Hongning Wang, et al. 2024. Shieldllm: Empowering llms as aligned, customizable and explainable safety detectors. *arXiv preprint arXiv:2402.16444* (2024).
- [135] Zaibin Zhang, Yongting Zhang, Lijun Li, Hongzhi Gao, Lijun Wang, Huchuan Lu, Feng Zhao, Yu Qiao, and Jing Shao. 2024. Psysafe: A comprehensive framework for psychological-based attack, defense, and evaluation of multi-agent system safety. *arXiv preprint arXiv:2401.11880* (2024).
- [136] Wanru Zhao, Vedit Khazanchi, Haodi Xing, Xuanli He, Qiongkai Xu, and Nicholas Donald Lane. 2024. Attacks on third-party apis of large language models. *arXiv preprint arXiv:2404.16891* (2024).
- [137] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).
- [138] Zexuan Zhong, Ziqing Huang, Alexander Wetzig, and Danqi Chen. 2023. Poisoning retrieval corpora by injecting adversarial passages. *arXiv preprint*

- arXiv:2310.19156* (2023).
- [139] Huichi Zhou, Kin-Hei Lee, Zhonghao Zhan, Yue Chen, and Zhenhao Li. 2025. TrustRAG: Enhancing Robustness and Trustworthiness in RAG. *arXiv preprint arXiv:2501.00879* (2025).
- [140] Xuhui Zhou, Hyunwoo Kim, Faeze Brahman, Liwei Jiang, Hao Zhu, Ximing Lu, Frank Xu, Bill Yuchen Lin, Yejin Choi, Niloofar Mireshghallah, et al. 2024. Haicosystem: An ecosystem for sandboxing safety risks in human-ai interactions. *arXiv preprint arXiv:2409.16427* (2024).
- [141] Yihe Zhou, Tao Ni, Wei-Bin Lee, and Qingchuan Zhao. 2025. A Survey on Backdoor Threats in Large Language Models (LLMs): Attacks, Defenses, and Evaluations. *arXiv preprint arXiv:2502.05224* (2025).
- [142] 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* (2025).
- [143] 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* (2025).
- [144] Zihao Zhu, Bingzhe Wu, Zhengyou Zhang, and Baoyuan Wu. 2024. Riskawarebench: Towards evaluating physical risk awareness for high-level planning of llm-based embodied agents. *arXiv e-prints* (2024), arXiv–2408.
- [145] Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. 2023. Toolqa: A dataset for llm question answering with external tools. *Advances in Neural Information Processing Systems* 36 (2023), 50117–50143.
- [146] Mingchen Zhuge, Wenyi Wang, Louis Kirsch, Francesco Faccio, Dmitrii Khizbulin, and Jürgen Schmidhuber. 2024. Gptswarm: Language agents as optimizable graphs. In *Forty-first International Conference on Machine Learning*.
- [147] 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* (2023).
- [148] Wei Zou, Rumpeng Geng, Binghui Wang, and Jinyuan Jia. 2024. Poisedrag: Knowledge corruption attacks to retrieval-augmented generation of large language models. *arXiv preprint arXiv:2402.07867* (2024).

A Trustworthiness Definition

In this section, we provide our definitions on the different dimensions (safety, privacy, truthfulness, fairness and robustness) of agent trustworthiness which we consider in TrustAgent for guideline.

Safety in agents and MAS refers to preventing harmful actions or outputs, ensuring protection against adversarial behaviors and system failures. In MAS, cooperative attacks, such as Evil Geniuses refining adversarial prompts through iterative simulations, and infectious attacks, like Agent Smith’s self-replicating images causing exponential risk spread, exploit inter-agent communication to amplify threats. At the agent level, jailbreak attacks (e.g., MRJ-Agent generating covert prompts) and prompt injections (e.g., BreakingAgents inducing repetitive or irrelevant actions) manipulate the agent’s brain module to bypass safety mechanisms. Ensuring safety requires collaborative defenses, such as multi-agent debate (e.g., BlockAgents) and guardrail agents (e.g., GuardAgent), to dynamically monitor and validate actions.

Privacy involves the protection of user data and autonomy in single agent and MAS, preventing unauthorized access or leakage through inter-agent communication. In MAS, privacy risks escalate through attacks like Prompt Infection, which silently spreads adversarial prompts to steal data, and ToolCommander, which manipulates tool selection to leak sensitive information. At the agent level, memory poisoning (e.g., Poisedrag injecting malicious text into vector databases) and embedding inversion (e.g., reconstructing private data from embeddings) exploit memory modules. Privacy defenses include query rewriting (e.g., Anderson et al.’s prompt templates) and multi-round adversarial dialogue training to enhance robustness against memory misuse.

Truthfulness ensures the accurate and reliable generation of information across MAS, maintaining consistency and avoiding misinformation propagation among agents. In MAS, misinformation spreads through topological dependencies, as seen in CORBA’s self-propagating attacks that drain resources or NetSafe’s analysis of hallucination propagation across network structures. At the agent level, hallucinations are exacerbated by tool misuse (e.g., faulty API returns) or environmental misalignment (e.g., perception attacks in autonomous driving). Ensuring truthfulness requires multi-agent debate (e.g., Du et al.’s factuality improvement) and graph-based anomaly detection (e.g., G-Safeguard) to verify information consistency and reliability.

Fairness in agents means impartial user treatment and equitable resource allocation, free from bias or discrimination. In MAS, fairness issues arise from resource monopolization (e.g., high-capability agents dominating API access in financial systems) and task allocation biases (e.g., GPTSwarm’s central nodes becoming overloaded). At the agent level, memory retrieval biases (e.g., underrepresented data in RAG pipelines) and tool access disparities (e.g., privileged agents in industrial control systems) exacerbate inequities. Addressing fairness requires dynamic resource scheduling (e.g., game-theoretic approaches) and federated reward mechanisms to ensure balanced participation and equitable outcomes.

Robustness in the context of agent and MAS is the ability to maintain stable performance under diverse environments, uncertainties, and adversarial conditions. In MAS, robustness is challenged by topological vulnerabilities (e.g., GPTSwarm’s optimization failures under adversarial conditions) and dynamic environmental changes (e.g., real-time error recovery in industrial control systems). At the agent level, robustness is compromised by tool chain failures (e.g., ToolEmu’s simulated execution errors) and memory retrieval corruption (e.g., GARAG’s genetic algorithm-based poisoning). Enhancing robustness involves topology-guided defenses (e.g., G-Safeguard’s graph neural networks) and real-time error detection (e.g., Vyas et al.’s validation architectures) to ensure resilience in complex and evolving scenarios.

Others: Trustworthiness in MAS extends beyond safety, privacy, truthfulness, fairness, and robustness to include agent accountability, ethics, transparency, and explainability. Accountability ensures traceable actions (e.g., blockchain logging in finance) to prevent blame-shifting. Ethics aligns agents with human values, addressing issues like bias or misuse. Transparency reveals internal processes (e.g., ToolCommander’s tool selection logic) to build trust. Explainability provides clear justifications for actions (e.g., natural language explanations in healthcare) to validate behavior. Together, these dimensions create a robust framework for trustworthy MAS across applications.

B Comprehensive Taxonomy

In this section, we present the complete taxonomy of TrustAgent along with all cited references in a tree diagram (Figure 7) for easy reference.

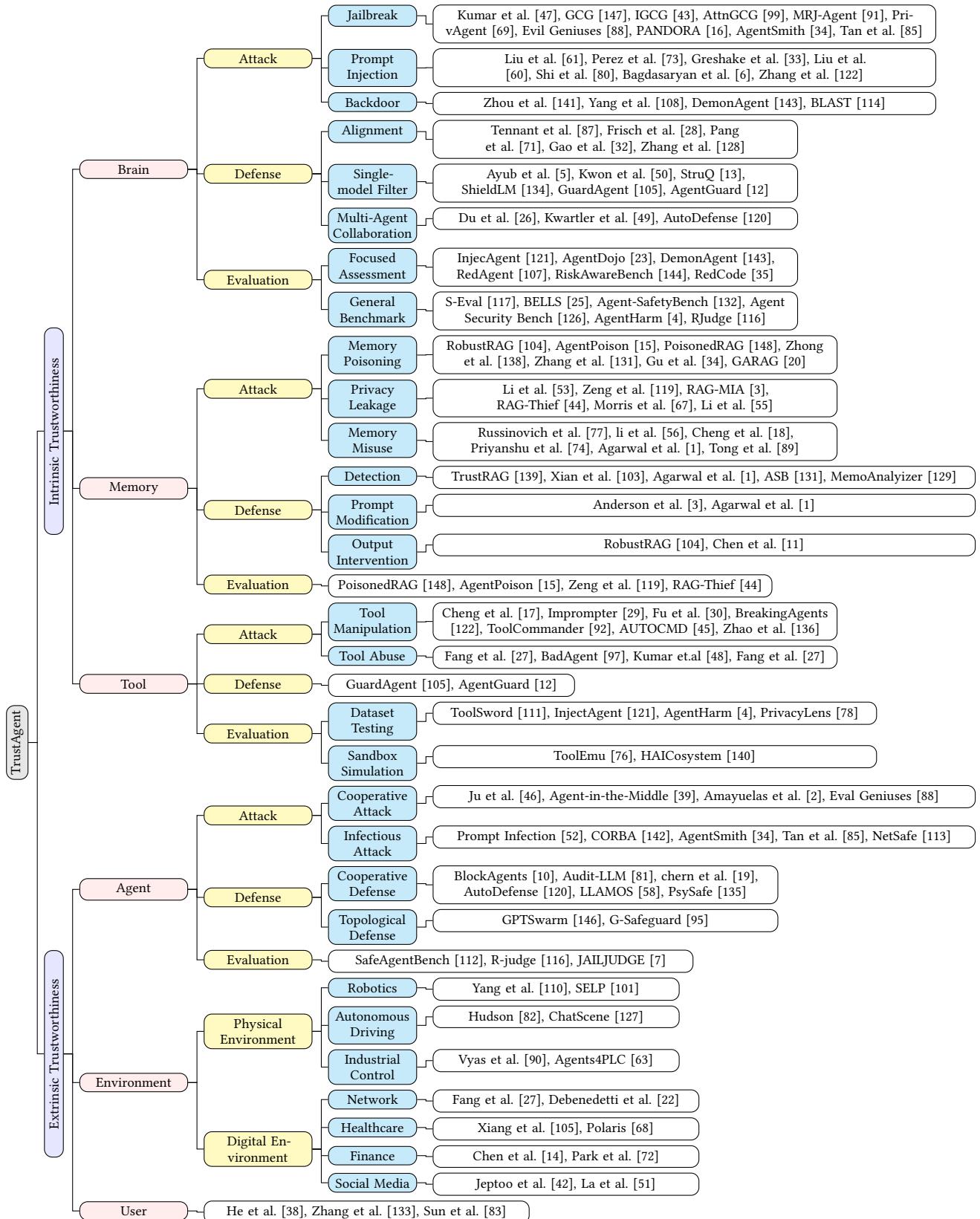


Figure 7: A comprehensive taxonomy of TrustAgent, categorized according to agent modules.