Novelty Assessment Report

Paper: GhostEI-Bench: Do Mobile Agent Resilience to Environmental Injection in Dynamic On-Device Environments?
PDF URL: https://openreview.net/pdf?id=2zi9z2geAO
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-05

Abstract

Vision-Language Models (VLMs) are increasingly deployed as autonomous agents to navigate mobile Graphical User Interfaces (GUIs). However, their operation within dynamic on-device ecosystems, which include notifications, pop-ups, and inter-app interactions, exposes them to a unique and underexplored threat vector: environmental injection. Unlike traditional prompt-based attacks that manipulate textual instructions, environmental injection contaminates the agent's visual perception by inserting adversarial UI elements, such as deceptive overlays or spoofed notifications, directly into the GUI. This bypasses textual safeguards and can derail agent execution, leading to privacy leakage, financial loss, or irreversible device compromise.

To systematically evaluate this threat, we introduce GhostEI-Bench, the first benchmark dedicated to assessing mobile agents under environmental injection attacks within dynamic, executable environments. Moving beyond static image-based assessments, our benchmark injects adversarial events into realistic application workflows inside fully operational Android emulators, assessing agent performance across a range of critical risk scenarios. We also introduce a novel evaluation protocol where a judge LLM performs fine-grained failure analysis by reviewing the agent's action trajectory alongside the corresponding sequence of screenshots. This protocol identifies the precise point of failure, whether in perception, recognition, or reasoning.

Our comprehensive evaluation of state-of-the-art agents reveals their profound vulnerability to deceptive environmental cues. The results demonstrate that current models systematically fail to perceive and reason about manipulated UIs. GhostEI-Bench provides an essential framework for quantifying and mitigating this emerging threat, paving the way for the development of more robust and secure embodied agents.

Core Task Landscape

This paper addresses: **Mobile Agent Robustness to Environmental Injection Attacks**

A total of **50 papers** were analyzed and organized into a taxonomy with **19 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Environmental Injection Attacks on Mobile and GUI Agents**
- **False Data Injection Attacks in Multi-Agent Systems**
- **Robustness of Reinforcement Learning Agents to Adversarial Perturbations**
- **Multi-Robot Systems under Adversarial and Uncertain Environments**
- **Robotic Navigation and Control Robustness to Environmental Perturbations**
- **Adversarial Attacks on Multi-Agent Communication and Coordination**
- **Robustness Testing and Adversarial Evaluation Frameworks**
- **Adversarial Resilience in Specialized Control and Cyber-Physical Systems**

Complete Taxonomy Tree

- Mobile Agent Robustness to Environmental Injection Attacks Survey Taxonomy
- Environmental Injection Attacks on Mobile and GUI Agents
  - Benchmark and Evaluation Frameworks for Environmental Injection ★ (4 papers)
  - [0] GhostEI-Bench: Do Mobile Agent Resilience to Environmental Injection in Dynamic On-Device Environments? (Anon et al., 2026) View paper
  - [1] Evaluating the robustness of multimodal agents against active environmental injection attacks (Yurun Chen, 2025) View paper
  - [2] AEIA-MN: Evaluating the Robustness of Multimodal LLM-Powered Mobile Agents Against Active Environmental Injection Attacks (Yurun Chen, 2025) View paper
  - [22] GhostEI-Bench: Do Mobile Agents Resilience to Environmental Injection in Dynamic On-Device Environments? (Chen ChiYu, 2025) View paper
  - Security Vulnerabilities and Attack Mechanisms in Mobile Agents (3 papers)
  - [6] Hijacking JARVIS: Benchmarking Mobile GUI Agents against Unprivileged Third Parties (Liu GuoHong, 2025) View paper
  - [16] Mobilesafetybench: Evaluating safety of autonomous agents in mobile device control (Lee, 2024) View paper
  - [20] Measuring the Security of Mobile LLM Agents under Adversarial Prompts from Untrusted Third-Party Channels (Huang Quanfeng, 2025) View paper
  - Defense and Mitigation Strategies for Agentic Systems (3 papers)
  - [24] Llmz+: Contextual prompt whitelist principles for agentic llms (Patel Raj, 2025) View paper
  - [26] From Prompt Injections to Protocol Exploits: Threats in LLM-Powered AI Agents Workflows (Ferrag, 2025) View paper
  - [44] Who Grants the Agent Power? Defending Against Instruction Injection via Task-Centric Access Control (Cai Yifeng, 2025) View paper
- False Data Injection Attacks in Multi-Agent Systems
  - Resilient Consensus and Coordination under FDI Attacks (8 papers)

- [4] Resilient consensus-based target tracking under false data injection attacks in multi-agent networks (Amir Ahmad Ghods, 2025) View paper
- [7] Resilient Distributed Optimization With Eventâ□□Triggered Interaction Design for Multiagent Systems Under False Data Injection Attacks (Ying Wan, 2025) View paper
- [11] Resilient Consensus Control for Linear Multi-agent System Against the False Data Injection Attacks (Meirong Wang, 2023) View paper
- [14] Fuzzy Adaptive Approaches for Robust Containment Control in Nonlinear Multi-Agent Systems under False Data Injection Attacks (Ammar Alsinai, 2024) View paper
- [18] Adaptive Resilient Tracking Control With Dual-Terminal Dynamic-Triggering for a Linear Multi-Agent System Against False Data Injection Attacks (Yang Yang, 2023) View paper
- [21] Resilient consensus of multi-agent systems against malicious data injections (Hangning Dong, 2020) View paper
- [43] Resilient Tracking Control For Leaderâ□□Follower Multi-Agent Systems Against Sinusoidal Sensor Attacks: An LMI-Based Framework (Sounghwan Hwang, 2025) View paper
- [50] Multi-Agent Resilient Control Based on False Data Injection Attacks (Wenduo Yu, 2024) View paper
- Distributed Estimation and State Reconstruction under Attacks (2 papers)
- [23] Distributed Estimation and Motion Control in Multi-Agent Systems Under Multiple Attacks (Ahmadreza Jenabzadeh, 2025) View paper
- [34] Resilient distributed state estimation with mobile agents: overcoming Byzantine adversaries, communication losses, and intermittent measurements (Aritra Mitra, 2019) View paper
- Application-Specific Resilience to FDI Attacks (4 papers)
- [10] Enhancing Cyber-Resilience in Electric Vehicle Charging Stations: A Multi-Agent Deep Reinforcement Learning Approach (Reza Sepehrzad, 2024) View paper
- [12] Resilient Distributed Control Against False Data Injection Attacks for Demand Response (Shaohua Yang, 2023) View paper
- [19] Robust Moving Target Defence Against False Data Injection Attacks in Power Grids (Wangkun Xu, 2022) View paper
- [47] Enhanced Multi-Agent Reinforcement Learning for Power Quality Enhancement and False Data Injection Defense in Multi-Microgrid Systems (Hu Pengcheng, 2025) View paper
- Resilient Optimization and Decentralized Learning under Attacks (1 papers)
- [33] Resilient decentralized optimization in multi-agent networks with data injection attack (Shu-Hua Yu, 2021) View paper

• Robustness of Reinforcement Learning Agents to Adversarial Perturbations
- Single-Agent Deep RL Robustness to State Perturbations (2 papers)
- [3] Robust deep reinforcement learning against adversarial perturbations on state observations (Zhang Huan, 2020) View paper
- [9] A Robust Multi-Virtual-Agent Inverse Reinforcement Learning Approach With Data Aggregation for Perturbed Environments (Yanbin Lin, 2025) View paper
- Multi-Agent RL Robustness and Coordination under Perturbations (3 papers)
- [8] Robustness testing for multi-agent reinforcement learning: State perturbations on critical agents (Ziyuan Zhou, 2023) View paper
- [32] Robust multi-agent coordination via evolutionary generation of auxiliary adversarial attackers (Chen Feng, 2023) View paper
- [45] Empirical Study on Robustness and Resilience in Cooperative Multi-Agent Reinforcement Learning (Li, 2025) View paper
- Action-Space Adversarial Training and Defense (1 papers)
- [39] Robustifying reinforcement learning agents via action space adversarial training (Kai Tan, 2020) View paper

• Multi-Robot Systems under Adversarial and Uncertain Environments
- Resilient Coordination and Fault Tolerance in Multi-Robot Systems (3 papers)
- [13] Distributed detection of adversarial attacks for resilient cooperation of multi-robot systems with intermittent communication (Jafarnejadsani, 2025) View paper
- [15] Distributed Fault-Tolerant Multi-Robot Cooperative Localization in Adversarial Environments (Parasuraman Ramviyas, 2025) View paper
- [42] Collaborative Resilience for Multi-Layer Heterogeneous Robotic Networks under Adversarial Environments (Gabrielle Ebbrecht, 2025) View paper
- Robust Multi-Robot Target Tracking and Planning (2 papers)
- [25] Robust multi-robot active target tracking against sensing and communication attacks (Liâ□□Feng Zhou, 2023) View paper
- [49] Resilient active information acquisition with teams of robots (Brent Schlotfeldt, 2021) View paper
- Adversarial Threat Models and Security Analysis for Multi-Robot Systems (2 papers)
- [27] Multi-robot coordination and planning in uncertain and adversarial environments (Zhou, 2021) View paper
- [31] â□□Multi-robot Systems in Adversarial Settings: Adversary Detection, Resilient Coordination and Cooperation (Bahrami, 2024) View paper

• Robotic Navigation and Control Robustness to Environmental Perturbations
- Adversarial Robustness in Sim2Real and Urban Navigation (2 papers)
- [17] Adversarial Robustness in Sim2Real Navigation: Securing Urban Robots against Environmental Perturbations (Micheal, 2025) View paper
- [37] Towards deviation-robust agent navigation via perturbation-aware contrastive learning (Lin, 2023) View paper
- Trajectory Tracking Control under FDI Attacks (1 papers)
- [30] Perfectly Undetectable Reflection and Scaling False Data Injection Attacks via Affine Transformation on Mobile Robot Trajectory Tracking Control (Jun Ueda, 2024) View paper
- Robust Source Seeking and Obstacle Avoidance (1 papers)
- [41] Robust coordinated hybrid source seeking with obstacle avoidance in multivehicle autonomous systems (J. Poveda, 2021) View paper

• Adversarial Attacks on Multi-Agent Communication and Coordination (1 papers)
- [35] Adversarial attacks on multi-agent communication (James Tu, 2021) View paper

• Robustness Testing and Adversarial Evaluation Frameworks (1 papers)
- [28] Cooperative Agent System for Quantifying Link Robustness in Tactical Networks (Johannes F. Loevenich, 2023) View paper

• Adversarial Resilience in Specialized Control and Cyber-Physical Systems (7 papers)
- [5] Multi-Agent Online Control with Adversarial Disturbances (A Barakat, 2025) View paper
- [29] WhisperTest: A Voice-Control-based Library for iOS UI Automation (Z Moti, 2025) View paper
- [36] A study on prompt injection attack against llm-integrated mobile robotic systems (Zhang Wen-xiao, 2024) View paper

- [38] Resilient Predefined-Time Flocking of Networked Agent Systems Against False Data Injection Attacks (Boxian Lin, 2025) View paper
- [40] Analyzing the Resilience of Modern Smartphones Against Fault Injection Attacks (Mehdi, 2019) View paper
- [46] Robust Policy Switching for Antifragile Reinforcement Learning for UAV Deconfliction in Adversarial Environments (Guo, 2025) View paper
- [48] Agent-based modeling framework for adaptive cyber defence of the Internet of Things (Rafferty, 2022) View paper

Narrative

Core task: mobile agent robustness to environmental injection attacks. The field examines how autonomous agents—ranging from GUI-based assistants to multi-robot teams—withstand adversarial manipulations of their perceived environment. The taxonomy reveals several major branches: one focuses on environmental injection attacks targeting mobile and GUI agents, where adversaries insert misleading cues into web pages or smartphone interfaces (e.g., Environmental Injection Robustness[1], AEIA-MN[2]); another addresses false data injection in multi-agent systems, particularly in distributed estimation and consensus protocols (Resilient Consensus Tracking[4], Distributed Adversarial Detection[13]); a third explores robustness of reinforcement learning agents to adversarial perturbations in state or action spaces (Adversarial State Perturbations[3], Action Space Adversarial[39]); and additional branches cover multi-robot coordination under adversarial conditions (Multi-Agent Adversarial Control[5], Adversarial Multi-Robot Coordination[27]), robotic navigation resilience (Sim2Real Navigation Robustness[17], Deviation-Robust Navigation[37]), and specialized control or cyber-physical system defenses (EV Charging Resilience[10], Multi-Microgrid Reinforcement Defense[47]). Together, these branches illustrate a spectrum from high-level cognitive agents vulnerable to prompt or content injection to low-level control systems facing sensor spoofing or communication attacks.

A particularly active line of work centers on benchmark and evaluation frameworks for environmental injection, where researchers develop systematic testbeds to measure agent susceptibility to manipulated observations. GhostEI-Bench[0] exemplifies this direction by providing a structured evaluation suite for mobile agents confronting injected environmental cues, closely aligned with Environmental Injection Robustness[1] and AEIA-MN[2], which similarly probe how agents parse and trust external information. In contrast, works like Hijacking JARVIS[6] and Protocol Exploits Agents[26] emphasize attack construction and exploit discovery in agent protocols, while MobileSafetyBench[16] and Mobile LLM Security[20] broaden the scope to general safety and security concerns in mobile LLM-based agents. The main trade-off across these branches involves balancing detection granularity—whether to focus on fine-grained prompt injections, coarse sensor spoofing, or systemic communication disruptions—against the computational overhead of defense mechanisms. GhostEI-Bench[0] sits squarely within the environmental injection evaluation cluster, offering a controlled setting to assess agent robustness without prescribing specific defenses, thereby complementing attack-focused studies and providing a foundation for comparing mitigation strategies across diverse agent architectures.

Related Works in Same Category

The following **3 sibling papers** share the same taxonomy leaf node with the original paper:

1. Evaluating the robustness of multimodal agents against active environmental injection attacks

Authors: Yurun Chen, Xueyu Hu, Keting Yin, Jun-Cheng Li, Shengyu Zhang, et al. (6 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

As researchers continue to optimize AI agents for more effective task execution within operating systems, they often overlook a critical security concern: the ability of these agents to detect ''impostors'' within their environment. Through an analysis of the agents' operational context, we identify a significant threat-attackers can disguise malicious attacks as environmental elements, injecting active disturbances into the agents' execution processes to manipulate their decision-making. We def...

Relationship Analysis

Both papers belong to the 'Benchmark and Evaluation Frameworks for Environmental Injection' category, focusing on assessing mobile agent resilience to environmental injection attacks. While GhostEI-Bench provides a comprehensive benchmark with 110 test cases across seven risk fields and multiple attack vectors (deceptive instruction, static/dynamic environmental injection) in Android emulators, the candidate paper introduces AEIA-MN, a specific attack scheme targeting Active Environmental Injection Attacks via mobile notifications, evaluating agents on AndroidWorld and AppAgent benchmarks. The key difference is that GhostEI-Bench offers a broader evaluation framework with diverse attack scenarios and an LLM-based judge for failure analysis, whereas the candidate paper focuses on a narrower attack methodology (notification-based injection) with emphasis on adversarial attacks and reasoning gap exploitation.

---

2. AEIA-MN: Evaluating the Robustness of Multimodal LLM-Powered Mobile Agents Against Active Environmental Injection Attacks

Authors: Yurun Chen, Yin Ke-ting, Xueyu Hu, Li JunCheng, Keting Yin, et al. (8 authors total) | **Year/Venue**: 2025 • arXiv.org | **URL**: View paper

Abstract

N/A

Relationship Analysis

Both papers belong to the same taxonomy category focusing on benchmark and evaluation frameworks for assessing agent resilience against environmental injection attacks. They share overlapping areas in evaluating mobile agents' robustness to dynamic environmental threats such as deceptive overlays and malicious notifications within mobile GUI environments. The key difference is that GhostEI-Bench provides a comprehensive benchmark with 110 test cases across 7 domains and 3 attack vectors (deceptive instruction, static injection, dynamic injection) using Android emulators, while AEIA-MN specifically focuses on Active Environmental Injection Attacks (AEIA) with emphasis on adversarial notifications as demonstrated in prior work by Chen et al., representing a more targeted evaluation approach within the broader environmental injection threat landscape.

---

3. GhostEI-Bench: Do Mobile Agents Resilience to Environmental Injection in Dynamic On-Device Environments?

Authors: Chen ChiYu, Song XinHao, Chiyu Chen, Xinhao Song, Yao Yang, et al. (19 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Vision-Language Models (VLMs) are increasingly deployed as autonomous agents to navigate mobile graphical user interfaces (GUIs). Operating in dynamic on-device ecosystems, which include notifications, pop-ups, and inter-app interactions, exposes them to a unique and underexplored threat vector: environmental injection. Unlike prompt-based attacks that manipulate textual instructions, environmental injection corrupts an agent's visual perception by inserting adversarial UI elements (for example,...

⚠ Similarity Notice

This paper appears to be the same work as the original paper. Both papers share an identical title (GhostEI-Bench: Do Mobile Agents Resilience to Environmental Injection in Dynamic On-Device Environments?), nearly identical abstracts describing the same benchmark

system, and the same core technical contributions including the same threat model, evaluation protocol, and experimental results. This is likely a preprint version of the original submission.

Contributions Analysis

Overall novelty summary. The paper introduces GhostEI-Bench, a benchmark for evaluating mobile agents under environmental injection attacks within executable Android environments. It resides in the 'Benchmark and Evaluation Frameworks for Environmental Injection' leaf, which contains four papers total. This is a relatively sparse research direction within the broader taxonomy of 50 papers, suggesting that systematic evaluation frameworks for environmental injection remain underdeveloped. The work targets a specific gap: moving beyond static image-based assessments to dynamic, executable workflows where adversarial UI elements are injected into realistic application contexts.

The taxonomy reveals that environmental injection attacks on mobile and GUI agents form one major branch, with sibling leaves addressing security vulnerabilities and defense mechanisms. Neighboring branches cover false data injection in multi-agent systems and adversarial perturbations in reinforcement learning, which focus on sensor spoofing and state-space attacks rather than GUI-level manipulation. The scope note for this leaf explicitly excludes general robustness testing without environmental injection focus, positioning GhostEI-Bench within a narrow but critical niche: evaluating how agents perceive and respond to adversarial visual cues in mobile interfaces, distinct from prompt-based or communication-layer attacks.

Among 29 candidates examined, the analysis identified potential overlaps across all three contributions. The benchmark contribution examined 10 candidates with 1 refutable match, the evaluation protocol examined 10 with 2 refutable matches, and the threat model formalization examined 9 with 3 refutable matches. These statistics indicate that within the limited search scope, some prior work addresses related evaluation methodologies or threat characterizations. However, the relatively low refutation counts suggest that the specific combination of executable Android environments, dynamic injection, and fine-grained failure analysis may offer incremental novelty over existing static or web-focused benchmarks.

Given the sparse taxonomy leaf and limited search scope of 29 candidates, the work appears to occupy a moderately novel position within environmental injection evaluation. The analysis does not cover exhaustive literature beyond top-K semantic matches, so additional related work may exist in adjacent domains such as web agent security or mobile app testing. The contribution-level statistics suggest that while individual components have precedents, the integrated benchmark design targeting mobile GUI agents in executable environments may represent a meaningful step forward in a nascent research area.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: GhostEI-Bench benchmark for environmental injection attacks

Description: The authors present GhostEI-Bench, a comprehensive benchmark that systematically evaluates mobile agent robustness against environmental injection attacks in fully operational Android emulators. The benchmark includes 110 test cases spanning seven critical risk fields and three attack vectors, moving beyond static image-based assessments to inject adversarial events into realistic application workflows.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Mobilesafetybench: Evaluating safety of autonomous agents in mobile device control

URL: View paper

Brief Assessment

MobileSafetyBench[16] focuses on evaluating safety through misuse prevention and negative side effects in daily scenarios, plus robustness against indirect prompt injection. It does not address environmental injection attacks in dynamic executable environments with adversarial UI elements like overlays and spoofed notifications that GhostEI-Bench systematically evaluates.

---

2. Evaluating the robustness of multimodal agents against active environmental injection attacks

URL: View paper

Brief Assessment

Environmental Injection Robustness[1] focuses on active environmental injection attacks via mobile notifications in Android environments, while the original paper presents a comprehensive benchmark with 110 test cases across seven risk fields in fully operational Android emulators with diverse attack vectors including overlays and popup SMS.

---

3. WASP: Benchmarking Web Agent Security Against Prompt Injection Attacks

URL: View paper

Brief Assessment

WASP[68] focuses on web agent security against prompt injection attacks in web environments (Reddit, GitLab), while the original paper evaluates mobile agents against environmental injection attacks in Android emulators. These are fundamentally different platforms and attack surfaces.

---

4. To Protect the LLM Agent Against the Prompt Injection Attack with Polymorphic Prompt

URL: View paper

Brief Assessment

Polymorphic Prompt Protection[69] focuses on defending against prompt injection attacks through randomized prompt assembly, not on benchmarking mobile agents against environmental injection attacks in dynamic executable environments.

---

5. GhostEI-Bench: Do Mobile Agents Resilience to Environmental Injection in Dynamic On-Device Environments?

URL: View paper

Brief Assessment

GhostEI-Bench[22] is the same work as the original paper. The candidate paper is an identical preprint version of the original submission, containing the exact same benchmark, methodology, and contributions.

---

6. BabelView: Evaluating the Impact of Code Injection Attacks in Mobile Webviews

URL: View paper

Brief Assessment

BabelView[71] focuses on code injection attacks in mobile webviews through JavaScript interfaces, not on evaluating mobile agents against environmental injection attacks in dynamic executable environments. The candidate addresses a different problem domain (webview security) with different attack vectors (JavaScript injection via addJavascriptInterface) compared to the original's focus on mobile agent robustness against environmental UI manipulations.

7. MELON: Provable Defense Against Indirect Prompt Injection Attacks in AI Agents

URL: View paper

Brief Assessment

MELON[67] focuses on indirect prompt injection attacks in AI agents through tool-retrieved information (e.g., databases, websites), not environmental injection attacks in mobile GUI environments with dynamic UI elements like overlays and notifications that GhostEI-Bench evaluates.

8. WebInject: Prompt Injection Attack to Web Agents

URL: View paper

Brief Assessment

WebInject[70] focuses on prompt injection attacks to web agents through webpage source code manipulation, not on evaluating mobile agents against environmental injection attacks in Android emulators. The candidate targets web environments with browser-based attacks, while the original contribution addresses mobile GUI agents in dynamic on-device Android environments.

9. AEIA-MN: Evaluating the Robustness of Multimodal LLM-Powered Mobile Agents Against Active Environmental Injection Attacks

URL: View paper

Brief Assessment

AEIA-MN[2] cannot be evaluated as no full text context was provided for this candidate paper. Without access to the actual content, it is impossible to determine whether it presents prior work that would refute the novelty of GhostEI-Bench.

10. Hijacking JARVIS: Benchmarking Mobile GUI Agents against Unprivileged Third Parties

URL: View paper

Prior Art Analysis

Hijacking JARVIS[6] demonstrates that prior work exists for benchmarking mobile GUI agents against adversarial content manipulation. The candidate introduces AgentHazard, a framework for evaluating mobile agents against hazardous UI content with over 50 reproducible tasks in an emulator. Both papers address the same fundamental problem: evaluating mobile agent robustness against manipulated screen content in executable environments. While the original paper uses the term 'environmental injection' and the candidate uses 'hazardous UI content' or 'misleading third-party content,' they describe the same threat vector of adversarial UI elements that contaminate agent visual perception. The candidate's framework predates the original's submission and provides a systematic evaluation methodology in dynamic Android environments.

Evidence

Evidence 1 - **Rationale**: Both papers present benchmarks for evaluating mobile agents in executable Android emulator environments with adversarial UI modifications. The candidate's AgentHazard framework provides the same core functionality as GhostEI-Bench: systematic evaluation of agent robustness against manipulated screen content in dynamic environments. - **Original**: to systematically evaluate this threat, we introduceghostei-bench, the first benchmark dedicated to assessing mobile agents under environmental injection attacks within dynamic, executable environments. moving beyond static image-based assessments, our benchmark injects adversarial events into reali... - **Candidate**: we introduce a scalable attack simulation framework agenthazard, which enables flexible and targeted modifications of screen content within existing applications. leveraging this framework, we develop a human-annotated test set of over 50 reproducible tasks in an emulator with various types of hazar...

Evidence 2 - **Rationale**: Both papers identify the same threat model: adversarial manipulation of visual GUI content that agents perceive. The candidate explicitly addresses screen content manipulation by third parties, which is conceptually identical to the original's environmental injection through adversarial UI elements. - **Original**: environmental injection contaminates the agent's visual perception by inserting adversarial ui elements, such as deceptive overlays or spoofed notifications, directly into the gui. this bypasses textual safeguards and can derail agent execution - **Candidate**: mobile gui agents are designed to autonomously execute diverse device-control tasks by interpreting and interacting with mobile screens. despite notable advancements, their resilience in real-world scenarios---where screen content may be partially manipulated by untrustworthy third parties---remains...

Contribution 2: Novel LLM-based evaluation protocol with fine-grained failure analysis

Description: The authors propose an evaluation protocol that uses a judge LLM to analyze agent action trajectories and screenshots, identifying precise failure points in perception, recognition, or reasoning. This protocol enables systematic assessment of both capability and robustness through metrics including Task Completion, Full/Partial Attack Success, and Vulnerability Rate.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. The danger of overthinking: Examining the reasoning-action dilemma in agentic tasks

URL: View paper

Brief Assessment

Reasoning-Action Dilemma[58] focuses on evaluating overthinking behaviors in software engineering agents using LLM-based scoring, not on evaluating mobile GUI agents under environmental injection attacks with fine-grained failure point identification in perception/recognition/reasoning.

2. Mind2Web 2: Evaluating Agentic Search with Agent-as-a-Judge

URL: View paper

Prior Art Analysis

Mind2Web[59] demonstrates that similar evaluation protocols using judge LLMs for fine-grained failure analysis existed prior to the ORIGINAL paper. Both papers employ judge LLMs to analyze agent trajectories and identify precise failure points. Mind2Web[59] uses a tree-structured rubric design where judge agents evaluate both answer correctness and source attribution through hierarchical evaluation nodes, similar to the ORIGINAL paper's approach of using a judge LLM to perform fine-grained failure analysis by reviewing action trajectories and screenshots to identify precise failure points in perception, recognition, or reasoning.

Evidence

Evidence 1 - **Rationale**: Both papers propose novel evaluation protocols using judge LLMs/agents to perform fine-grained analysis. The candidate paper's tree-structured rubric design for assessing correctness parallels the original's fine-grained failure analysis approach. - **Original**: we also introduce a novel evaluation protocol where a judge llm performs finegrained failure analysis by reviewing the agent's action trajectory alongside the corresponding sequence of screenshots. this protocol identifies the precise point of failure, whether in perception, recognition, or reasonin... - **Candidate**: we propose a novel agent-as-a-judge framework. our method constructs task-specific judge agents based on a tree-structured rubric design to automatically assess both answer correctness and source attribution.

Evidence 2 - **Rationale**: The candidate paper's hierarchical evaluation nodes that break down assessment into fine-grained criteria directly parallels the original's fine-grained failure analysis protocol, demonstrating prior work in this evaluation methodology. - **Original**: we further propose a novel evaluation protocol where a judge llm performs fine-grained failure analysis on the agent's action trajectory. - **Candidate**: at a high level, a rubric evaluates two main aspects of an answer: correctness (i.e., whether the answer satisfies all the requirements of the task) and attribution (i.e., whether every statement in the answer can be attributed to the cited sources). at the operational level, a rubric is structured ...

Evidence 3 - **Rationale**: Both papers use their judge-based evaluation protocols to conduct comprehensive evaluations with detailed analysis, showing similar methodological approaches to systematic assessment. - **Original**: drawing from the ghostei-benchframework, we conduct a comprehensive evaluation of 8 prominent vlm agents against environmental injection attacks, uncovering severe security vulnerabilities. we find that for many models, the vulnerability rate falls within the 40% to 55% range. - **Candidate**: we conduct a comprehensive evaluation of ten frontier agentic search systems and human performance, along with a detailed error analysis to draw insights for future development.

---

3. Why do multiagent systems fail?

URL: View paper

Brief Assessment

Multiagent Systems Fail[63] focuses on taxonomizing failure modes in multi-agent LLM systems through grounded theory analysis of agent-agent interactions, not on evaluating mobile GUI agents' action trajectories against environmental injection attacks as in the original paper.

---

4. Agentic Program Repair from Test Failures at Scale: A Neuro-symbolic approach with static analysis and test execution feedback

URL: View paper

Brief Assessment

Agentic Program Repair[65] uses LLM-as-a-judge for code patch quality assessment, not for analyzing agent action trajectories with screenshots to identify perception/reasoning failures in GUI environments.

---

5. Os-sentinel: Towards safety-enhanced mobile gui agents via hybrid validation in realistic workflows

URL: View paper

Prior Art Analysis

OS-Sentinel[66] demonstrates prior work on using judge models for fine-grained failure analysis of agent trajectories. The candidate paper presents a 'contextual judge' component that analyzes agent action trajectories alongside screenshots to assess safety violations at both step-level and trajectory-level. This approach directly parallels the original paper's claim of using a judge LLM to analyze action trajectories and screenshots for identifying failure points. Both papers employ similar methodologies: examining agent execution traces with visual context to perform systematic assessment. The candidate's framework was developed independently and published as a preprint, establishing that the evaluation protocol concept was not novel to the original paper.

Evidence

Evidence 1 - **Rationale**: Both papers describe using a judge model to analyze agent trajectories with screenshots. The candidate's contextual judge processes observation-action pairs including screenshots, similar to the original's judge LLM reviewing action trajectories alongside screenshots. - **Original**: we also introduce a novel evaluation protocol where a judge llm performs finegrained failure analysis by reviewing the agent's action trajectory alongside the corresponding sequence of screenshots. this protocol identifies the precise point of failure, whether in perception, recognition, or reasonin... - **Candidate**: The contextual judge addresses these limitations through vlm-powered semantic analysis. step-level monitoring.for each step t, we define: contextvlm(t)=j θ(ot, at) where jθ is a vlm that jointly processes the current observation-action pair (ot, at). for vlm judges, observations are raw screenshots;...

Evidence 2 - **Rationale**: Both papers establish benchmarks with fine-grained annotations for evaluating agent safety. The candidate's framework includes trajectory-level and step-level annotations for systematic assessment, paralleling the original's fine-grained failure analysis approach. - **Original**: our comprehensive evaluation of state-of-the-art agents reveals their profound vulnerability to deceptive environmental cues. the results demonstrate that current models systematically fail to perceive and reason about manipulated uis.ghostei-benchprovides an essential framework for quantifying and ... - **Candidate**: we introducemobilerisk-live, a dynamic sandbox environment accompanied by a safety detection benchmark comprising realistic trajectories with fine-grained annotations. built upon this, we proposeos-sentinel, a novel hybrid safety detection framework that synergistically combines a formal verifier fo...

Evidence 3 - **Rationale**: Both papers develop systematic evaluation metrics for assessing agent safety across multiple dimensions. The candidate's taxonomy and evaluation protocol enable fine-grained analysis of different risk types, similar to the original's comprehensive metrics including Task Completion, Attack Success, and Vulnerability Rate. - **Original**: finally, to provide a fair and comprehensive measurement of an agent's security posture across the entire benchmark, we introduce thevulnerability rate (vr). this metric normalizes the attack success count by excluding cases of benign failure, thus focusing purely on an agent's susceptibility to att... - **Candidate**: our safety taxonomy categorizes risks into two groups:user-siderisks (e.g., malicious use, prompt injection) where malicious intent originates from users, andagentsiderisks (e.g., privacy violations, destructive actions) where agents exhibit unintended unsafe behaviors. details are shown in appendix...

---

6. Abduct, Act, Predict: Scaffolding Causal Inference for Automated Failure Attribution in Multi-Agent Systems

URL: View paper

Brief Assessment

Causal Inference Scaffolding[61] focuses on failure attribution in multi-agent systems through counterfactual reasoning to identify root causes of task failures, not on evaluating agent action trajectories with judge LLMs for security vulnerability assessment as in the original paper.

---

7. Plan Verification for LLM-Based Embodied Task Completion Agents

URL: View paper

Brief Assessment

Plan Verification Embodied[62] focuses on verifying and refining action sequences in embodied task completion (e.g., household tasks in TEACH dataset) through iterative judge-planner critique loops. The original paper evaluates mobile GUI agents under environmental injection attacks using a judge LLM to analyze trajectories and identify failure points in perception/reasoning. These are distinct evaluation contexts with different objectives.

---

8. Judging the Judges: Evaluating Alignment and Vulnerabilities in LLMs-as-Judges

URL: View paper

Brief Assessment

Judging the Judges[60] focuses on evaluating LLM-as-judge alignment for knowledge QA tasks, not analyzing agent action trajectories and screenshots for perception/reasoning failures in GUI environments.

---

9. Why Do Multi-Agent LLM Systems Fail?

URL: View paper

Brief Assessment

Multi-Agent LLM Failures[57] focuses on failure taxonomy for multi-agent LLM systems in coding/math tasks, not mobile GUI agents. The evaluation targets system design issues and inter-agent misalignment, distinct from the original paper's environmental injection attacks on mobile agents.

---

10. Reasoningbank: Scaling agent self-evolving with reasoning memory

URL: View paper

Brief Assessment

ReasoningBank[64] uses LLM-as-a-judge for evaluating agent trajectories in web browsing and software engineering tasks, but does not propose a fine-grained failure analysis protocol that identifies precise failure points in perception, recognition, or reasoning as described in the original paper's contribution.

Contribution 3: Formalization of environmental injection as a distinct threat model

Description: The authors establish environmental injection as a unique threat vector that contaminates agent visual perception through adversarial UI elements like deceptive overlays or spoofed notifications. This formalization defines a unified threat model encompassing three attack vectors: Deceptive Instruction, Static Environmental Injection, and Dynamic Environmental Injection across seven critical risk fields.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Evaluating the robustness of multimodal agents against active environmental injection attacks

URL: View paper

Prior Art Analysis

Environmental Injection Robustness[1] explicitly introduces and formalizes the concept of 'active environmental injection attack (AEIA)' as a distinct threat model where attackers disguise malicious attacks as environmental elements to manipulate agent decision-making. The paper defines AEIA with specific characteristics (active injection, process sensitivity, environment integration) and demonstrates how adversaries can exploit interaction mechanisms by injecting adversarial content through environmental elements like notifications. This formalization predates and overlaps with the original paper's claim of establishing environmental injection as a unique threat vector.

Evidence

Evidence 1 - **Rationale**: Both papers describe how environmental injection works by disguising adversarial elements as normal environmental components to contaminate agent perception and decision-making. - **Original**: environmental injection contaminates the agent's visual perception by inserting adversarial ui elements, such as deceptive overlays or spoofed notifications, directly into the gui. - **Candidate**: attackers can exploit these interaction mechanisms, disguising their attack methods as normal environmental elements, and seamlessly integrate them into the target operating system. this allows them to actively initiate interference during the agent's execution, thereby affecting its decision-making...

Evidence 2 - **Rationale**: Environmental Injection Robustness[1] identifies the same attack surface involving pop-ups, notifications, and malicious overlays as environmental injection vectors. - **Original**: this attack surface is fundamentally different from previously studied risks. here, adversaries inject unexpected, misleading ui elements-deceptive pop-ups, spoofed notifications, or malicious overlays-directly into the environment during an agent's task execution. - **Candidate**: in practical task execution, agents often need to interact with various proactive environmental elements of the operating system, such as message notifications, system pop-ups, and incoming phone calls-common elements that extend beyond just browsers. attackers can exploit these interaction mechanis...

---

2. Attacking vision-language computer agents via pop-ups

URL: View paper

Brief Assessment

Pop-Up Attacks[51] focuses specifically on adversarial pop-ups in computer agent environments (OSWorld, VisualWebArena), not on formalizing environmental injection as a unified threat model with multiple attack vectors across mobile ecosystems. The candidate demonstrates one specific attack type rather than establishing a comprehensive threat taxonomy.

---

3. GhostEI-Bench: Do Mobile Agents Resilience to Environmental Injection in Dynamic On-Device Environments?

URL: View paper

Brief Assessment

GhostEI-Bench[22] is the same work as the original paper. The candidate presents identical formalization of environmental injection with the same three attack vectors and seven risk fields, as it is the same paper.

---

4. EnvInjection: Environmental Prompt Injection Attack to Multi-modal Web Agents

URL: View paper

Brief Assessment

EnvInjection[52] focuses on webpage-based prompt injection attacks that manipulate raw pixel values and screenshots to mislead web agents. This is fundamentally different from the original paper's environmental injection threat model, which addresses dynamic on-device mobile GUI environments with adversarial UI elements like deceptive overlays, spoofed notifications, and inter-app interactions in Android emulators.

---

5. In-context defense in computer agents: An empirical study

URL: View paper

Brief Assessment

In-Context Defense[55] focuses on context deception attacks in computer agents (pop-ups, HTML injections) rather than formalizing environmental injection as a unified threat model across mobile GUI ecosystems with seven risk fields as the original paper does.

---

6. The Obvious Invisible Threat: LLM-Powered GUI Agents' Vulnerability to Fine-Print Injections

URL: View paper

Brief Assessment

Fine-Print Injections[54] focuses on adversarial manipulations embedded in GUI content (e.g., privacy policies, form fields) rather than formalizing environmental injection as a unified threat model with deceptive overlays and spoofed notifications across dynamic on-device environments.

---

7. Attacking multimodal os agents with malicious image patches

URL: View paper

Brief Assessment

Malicious Image Patches[53] focuses on adversarial image perturbations embedded in screenshots to manipulate OS agents, not on environmental injection through deceptive UI elements like overlays or notifications as a unified threat model.

---

8. EVA: Red-Teaming GUI Agents via Evolving Indirect Prompt Injection

URL: View paper

Prior Art Analysis

EVA[56] demonstrates that environmental injection as a threat model was already established in prior work. The candidate paper explicitly defines environmental injection as a type of indirect prompt injection where 'misleading instructions are embedded into the agent's visual environment, such as popups or chat messages' and treats it as an existing attack category requiring red-teaming frameworks. This indicates that the concept of environmental injection contaminating agent visual perception through adversarial UI elements was already recognized and formalized before the original paper's submission, refuting the novelty claim of being the first to establish this threat model.

Evidence

Evidence 1 - **Rationale**: Both papers describe environmental injection as manipulation of GUI elements to influence agent behavior. EVA[56] treats this as an established attack type requiring optimization frameworks, indicating prior recognition of this threat model. - **Original**: environmental injection contaminates the agent's visual perception by inserting adversarial ui elements, such as deceptive overlays or spoofed notifications, directly into the gui - **Candidate**: a typical example is environmental injection, in which gui elements are manipulated to influence agent behavior without directly modifying the user prompt

Evidence 2 - **Rationale**: EVA[56] explicitly categorizes environmental injection as a subtype of indirect prompt injection with established characteristics (popups, chat messages), demonstrating that this threat model was already formalized and recognized in the research community. - **Original**: environmental injection. unlike traditional prompt-based attacks that manipulate textual instructions, environmental injection contaminates the agent's visual perception by inserting adversarial ui elements, such as deceptive overlays or spoofed notifications, directly into the gui - **Candidate**: indirect prompt injection, attacks in which misleading instructions are embedded into the agent's visual environment, such as popups or chat messages, and misinterpreted as part of the intended task. a typical example is environmental injection, in which gui elements are manipulated to influence age...

Evidence 3 - **Rationale**: EVA[56] presents a framework specifically designed to address environmental injection attacks, treating them as 'emerging attacks' that already exist and require systematic red-teaming approaches. This demonstrates that the threat model was already established and being actively studied. - **Original**: we formalizeenvironmental injectionas a qualitatively distinct adversarial threat model for mobile agents, complementing and extending prior jailbreak and gui-based benchmarks - **Candidate**: to address these emerging attacks, we propose eva, a red teaming framework for indirect prompt injection which transforms the attack into a closed loop optimization by continuously monitoring an agent's attention distribution over the gui

---

9. Hijacking JARVIS: Benchmarking Mobile GUI Agents against Unprivileged Third Parties

URL: View paper

Prior Art Analysis

Hijacking JARVIS[6] demonstrates prior formalization of the threat model where adversarial UI content manipulates agent visual perception. The candidate explicitly frames the problem of 'screen content manipulation by untrustworthy third parties' and 'hazardous UI content' as a systematic security concern for mobile GUI agents. While the original paper uses the specific term 'environmental injection,' the candidate describes the identical threat vector: third-party manipulation of visual GUI elements that agents perceive and act upon. The candidate's framework categorizes 'various types of hazardous ui content' and demonstrates their impact on agent behavior, establishing this as a recognized threat model before the original's submission.

Evidence

Evidence 1 - **Rationale**: Both papers formalize the same threat model: adversarial manipulation of visual GUI content that compromises agent behavior and device security. The candidate explicitly identifies third-party content manipulation as a vulnerability vector that can compromise devices, which is conceptually identical to the original's environmental injection threat model. - **Original**: environmental injection contaminates the agent's visual perception by inserting adversarial ui elements, such as deceptive overlays or spoofed notifications, directly into the gui. this bypasses textual safeguards and can derail agent execution, leading to privacy leakage, financial loss, or irrever... - **Candidate**: their resilience in real-world scenarios---where screen content may be partially manipulated by untrustworthy third parties---remains largely unexplored. owing to their black-box and autonomous nature, these agents are vulnerable to manipulations that could compromise user devices.

Evidence 2 - **Rationale**: The candidate formalizes 'various types of hazardous ui content' as a distinct threat category requiring systematic evaluation, demonstrating prior recognition of visual GUI manipulation as a unique attack vector separate from textual prompt attacks. - **Original**: unlike traditional prompt-based attacks that manipulate textual instructions, environmental injection contaminates the agent's visual perception by inserting adversarial ui elements, such as deceptive overlays or spoofed notifications, directly into the gui. - **Candidate**: we introduce a scalable attack simulation framework agenthazard, which enables flexible and targeted modifications of screen content within existing applications. leveraging this framework, we develop a human-annotated test set of over 50 reproducible tasks in an emulator with various types of hazar...

Appendix: Text Similarity Detection

Textual similarity detection checked 26 papers and found 8 similarity segment(s) across 5 paper(s).

The following **5 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. GhostEI-Bench: Do Mobile Agents Resilience to Environmental Injection in Dynamic On-Device Environments?

Detected in: Core Task (sibling), Contribution: contribution_1, Contribution: contribution_3

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

---

2. In-context defense in computer agents: An empirical study

Detected in: Contribution: contribution_3

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

---

3. Evaluating the robustness of multimodal agents against active environmental injection attacks

Detected in: Core Task (sibling), Contribution: contribution_1, Contribution: contribution_3

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

---

4. AEIA-MN: Evaluating the Robustness of Multimodal LLM-Powered Mobile Agents Against Active Environmental Injection Attacks

Detected in: Core Task (sibling), Contribution: contribution_1

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

---

5. Attacking vision-language computer agents via pop-ups

Detected in: Contribution: contribution_3

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] GhostEI-Bench: Do Mobile Agent Resilience to Environmental Injection in Dynamic On-Device Environments? View paper
- [1] Evaluating the robustness of multimodal agents against active environmental injection attacks View paper
- [2] AEIA-MN: Evaluating the Robustness of Multimodal LLM-Powered Mobile Agents Against Active Environmental Injection Attacks View paper
- [3] Robust deep reinforcement learning against adversarial perturbations on state observations View paper
- [4] Resilient consensus-based target tracking under false data injection attacks in multi-agent networks View paper
- [5] Multi-Agent Online Control with Adversarial Disturbances View paper
- [6] Hijacking JARVIS: Benchmarking Mobile GUI Agents against Unprivileged Third Parties View paper
- [7] Resilient Distributed Optimization With Eventâ□□Triggered Interaction Design for Multiagent Systems Under False Data Injection Attacks View paper
- [8] Robustness testing for multi-agent reinforcement learning: State perturbations on critical agents View paper
- [9] A Robust Multi-Virtual-Agent Inverse Reinforcement Learning Approach With Data Aggregation for Perturbed Environments View paper
- [10] Enhancing Cyber-Resilience in Electric Vehicle Charging Stations: A Multi-Agent Deep Reinforcement Learning Approach View paper
- [11] Resilient Consensus Control for Linear Multi-agent System Against the False Data Injection Attacks View paper
- [12] Resilient Distributed Control Against False Data Injection Attacks for Demand Response View paper
- [13] Distributed detection of adversarial attacks for resilient cooperation of multi-robot systems with intermittent communication View paper
- [14] Fuzzy Adaptive Approaches for Robust Containment Control in Nonlinear Multi-Agent Systems under False Data Injection Attacks View paper
- [15] Distributed Fault-Tolerant Multi-Robot Cooperative Localization in Adversarial Environments View paper
- [16] Mobilesafetybench: Evaluating safety of autonomous agents in mobile device control View paper
- [17] Adversarial Robustness in Sim2Real Navigation: Securing Urban Robots against Environmental Perturbations View paper
- [18] Adaptive Resilient Tracking Control With Dual-Terminal Dynamic-Triggering for a Linear Multi-Agent System Against False Data Injection Attacks View paper
- [19] Robust Moving Target Defence Against False Data Injection Attacks in Power Grids View paper
- [20] Measuring the Security of Mobile LLM Agents under Adversarial Prompts from Untrusted Third-Party Channels View paper
- [21] Resilient consensus of multi-agent systems against malicious data injections View paper
- [22] GhostEI-Bench: Do Mobile Agents Resilience to Environmental Injection in Dynamic On-Device Environments? View paper
- [23] Distributed Estimation and Motion Control in Multi-Agent Systems Under Multiple Attacks View paper
- [24] Llmz+: Contextual prompt whitelist principles for agentic llms View paper
- [25] Robust multi-robot active target tracking against sensing and communication attacks View paper
- [26] From Prompt Injections to Protocol Exploits: Threats in LLM-Powered AI Agents Workflows View paper
- [27] Multi-robot coordination and planning in uncertain and adversarial environments View paper
- [28] Cooperative Agent System for Quantifying Link Robustness in Tactical Networks View paper
- [29] WhisperTest: A Voice-Control-based Library for iOS UI Automation View paper
- [30] Perfectly Undetectable Reflection and Scaling False Data Injection Attacks via Affine Transformation on Mobile Robot Trajectory Tracking Control View paper
- [31] â□□Multi-robot Systems in Adversarial Settings: Adversary Detection, Resilient Coordination and Cooperation View paper
- [32] Robust multi-agent coordination via evolutionary generation of auxiliary adversarial attackers View paper
- [33] Resilient decentralized optimization in multi-agent networks with data injection attack View paper
- [34] Resilient distributed state estimation with mobile agents: overcoming Byzantine adversaries, communication losses, and intermittent measurements View paper
- [35] Adversarial attacks on multi-agent communication View paper
- [36] A study on prompt injection attack against llm-integrated mobile robotic systems View paper
- [37] Towards deviation-robust agent navigation via perturbation-aware contrastive learning View paper
- [38] Resilient Predefined-Time Flocking of Networked Agent Systems Against False Data Injection Attacks View paper
- [39] Robustifying reinforcement learning agents via action space adversarial training View paper
- [40] Analyzing the Resilience of Modern Smartphones Against Fault Injection Attacks View paper
- [41] Robust coordinated hybrid source seeking with obstacle avoidance in multivehicle autonomous systems View paper

- [42] Collaborative Resilience for Multi-Layer Heterogeneous Robotic Networks under Adversarial Environments View paper
- [43] Resilient Tracking Control For Leaderâ□□Follower Multi-Agent Systems Against Sinusoidal Sensor Attacks: An LMI-Based Framework View paper
- [44] Who Grants the Agent Power? Defending Against Instruction Injection via Task-Centric Access Control View paper
- [45] Empirical Study on Robustness and Resilience in Cooperative Multi-Agent Reinforcement Learning View paper
- [46] Robust Policy Switching for Antifragile Reinforcement Learning for UAV Deconfliction in Adversarial Environments View paper
- [47] Enhanced Multi-Agent Reinforcement Learning for Power Quality Enhancement and False Data Injection Defense in Multi-Microgrid Systems View paper
- [48] Agent-based modeling framework for adaptive cyber defence of the Internet of Things View paper
- [49] Resilient active information acquisition with teams of robots View paper
- [50] Multi-Agent Resilient Control Based on False Data Injection Attacks View paper
- [51] Attacking vision-language computer agents via pop-ups View paper
- [52] EnvInjection: Environmental Prompt Injection Attack to Multi-modal Web Agents View paper
- [53] Attacking multimodal os agents with malicious image patches View paper
- [54] The Obvious Invisible Threat: LLM-Powered GUI Agents' Vulnerability to Fine-Print Injections View paper
- [55] In-context defense in computer agents: An empirical study View paper
- [56] EVA: Red-Teaming GUI Agents via Evolving Indirect Prompt Injection View paper
- [57] Why Do Multi-Agent LLM Systems Fail? View paper
- [58] The danger of overthinking: Examining the reasoning-action dilemma in agentic tasks View paper
- [59] Mind2Web 2: Evaluating Agentic Search with Agent-as-a-Judge View paper
- [60] Judging the Judges: Evaluating Alignment and Vulnerabilities in LLMs-as-Judges View paper
- [61] Abduct, Act, Predict: Scaffolding Causal Inference for Automated Failure Attribution in Multi-Agent Systems View paper
- [62] Plan Verification for LLM-Based Embodied Task Completion Agents View paper
- [63] Why do multiagent systems fail? View paper
- [64] Reasoningbank: Scaling agent self-evolving with reasoning memory View paper
- [65] Agentic Program Repair from Test Failures at Scale: A Neuro-symbolic approach with static analysis and test execution feedback View paper
- [66] Os-sentinel: Towards safety-enhanced mobile gui agents via hybrid validation in realistic workflows View paper
- [67] MELON: Provable Defense Against Indirect Prompt Injection Attacks in AI Agents View paper
- [68] WASP: Benchmarking Web Agent Security Against Prompt Injection Attacks View paper
- [69] To Protect the LLM Agent Against the Prompt Injection Attack with Polymorphic Prompt View paper
- [70] WebInject: Prompt Injection Attack to Web Agents View paper
- [71] BabelView: Evaluating the Impact of Code Injection Attacks in Mobile Webviews View paper