Towards Trustworthy GUI Agents: A Survey

Anonymous ACL submission

Abstract

Graphical User Interface (GUI) agents extend large language models from text generation to action execution in real-world digital environments. Unlike conversational systems, GUI agents perform irreversible operations such as submitting forms, granting permissions, or deleting data, making trustworthiness a core requirement. This survey identifies the execution gap as a key challenge in building trustworthy GUI agents: the misalignment between perception, reasoning, and interaction in dynamic, partially observable interfaces. We introduce a workflow-aligned taxonomy that decomposes trust into Perception Trust, Reasoning Trust, and Interaction Trust, showing how failures propagate across agent pipelines and compound through action/observation loops. We systematically review benign failure modes and adversarial attacks at each stage, together with corresponding defense mechanisms tailored to GUI settings. We further analyze evaluation practices and argue that task completion alone is insufficient for trust assessment. We highlight emerging trust-aware metrics and benchmarks that capture error cascades and the security/utility trade-off, and outline open challenges for deploying GUI agents safely and reliably.

1 Introduction

The emergence of GUI agents marks a fundamental transition in how AI systems interact with the digital world. Unlike chatbots that generate text responses, GUI agents take *actions*, clicking buttons, filling forms, and navigating websites that produce immediate, often irreversible, real-world consequences (Nguyen et al., 2024a; Wang et al., 2024b; Xie et al., 2024). This shift from *generation* to *execution* fundamentally changes the stakes of AI trustworthiness. The contrast is clear: when a language model hallucinates in a conversation, the user can simply ignore the response; when a GUI agent hallucinates a button that doesn't exist and clicks the wrong element, it might authorize an unintended purchase, delete important files, or expose sensitive information (Yang et al., 2024; Levy et al., 2024). The cost of failure is no longer measured in user dissatisfaction but in tangible harm.

We argue that the core challenge underlying GUI agent trustworthiness is what we term the **Execution Gap**: the fundamental disconnect between three levels of agent operation. *Perceptual Fidelity* requires correctly mapping visual pixels or Document Object Model (DOM) structures to semantic understanding of interface elements. *Reasoning Fidelity* demands maintaining logical consistency across multi-step plans in environments that change between actions. *Interaction Fidelity* involves translating intended actions into precise coordinates or commands that achieve the desired effect. This gap explains why techniques successful for static LLM applications often fail for GUI agents (Zheng et al., 2024; Chae et al., 2024). For instance, a model that excels at describing what it sees in an image may still click the wrong button because it cannot reliably map its understanding to actionable coordinates. A planner that generates coherent step sequences may fail when a pop-up dialog invalidates its plan mid-execution.

Existing surveys on LLM trustworthiness address privacy, bias, and hallucination (Liu et al., 2023c; Weidinger et al., 2022; Gan et al., 2024). Although these concerns apply to GUI agents, three characteristics make GUI-specific analysis essential. First, *irreversibility*: text generation is infinitely reversible, users simply regenerate, but GUI actions often cannot be undone, as sent emails, deleted files, and completed transactions persist (Hua et al., 2024). This asymmetry demands different safety architectures than those designed for conversational AI. Second, *dynamic environments*: unlike static documents, GUIs change constantly through DOM updates, loading states,

pop-ups, and A/B testing, meaning the interface an agent perceives may differ from the interface it acts upon milliseconds later (Ma et al., 2024). Trust must account for environmental non-stationarity. Third, *action-observation loops*: GUI agents operate in closed loops where each action changes the environment, affecting subsequent observations, and errors compound as a wrong click leads to an unexpected screen, which leads to further misinterpretation (Wu et al., 2025a).

This survey makes three primary contributions. First, we present a **workflow-aligned taxonomy** organizing trustworthiness around Perception Trust (§3), Reasoning Trust (§4), and Interaction Trust (§5), reflecting how vulnerabilities propagate through agent pipelines. Second, we provide a comprehensive analysis of **defense mechanisms** integrated within each trust dimension, revealing how mitigations must be stage-specific. Third, we analyze **evaluation methodologies** (§6) with emphasis on the security-utility trade-off that defines practical deployment decisions. Figure 1 presents our "Risk & Mitigation Landscape", a unified view mapping threats to agent modules and their real-world impacts, which serves as a roadmap for this survey.

2 Foundations: The GUI Agent Pipeline

Before analyzing trustworthiness, we establish the foundations of GUI agents by examining their execution pipeline, the existing execution gap at each stage, and the limitations of standard LLM safety in agentic settings.

2.1 Pipeline Architecture

GUI agents typically operate through three interconnected stages: perception, reasoning, and interaction (Lu et al., 2023; Wang et al., 2024b).

Perception converts raw interface inputs, such as screenshots, DOM structures, accessibility trees, or hybrid representations, into a semantic understanding of the interface state (Wu et al., 2024b; Nong et al., 2024). This stage answers the question: *What elements exist, and what do they represent?* Existing approaches span pure vision-based perception using multimodal large language models (MLLMs) (Zheng et al., 2024), structured parsing of HTML and accessibility APIs (Deng et al., 2023), and hybrid designs that combine visual and structural cues for improved robustness (Wang et al., 2024a). Recent work on universal visual grounding further argues that fully visual perception with pixel-level action execution can rival or surpass text-augmented methods (Gou et al., 2024).

Reasoning operates on the perceived state and task instructions to determine the next action. This includes task decomposition, progress tracking, and decision-making over possible action sequences (Gu et al., 2024; Zhu et al., 2025; Koh et al., 2024b). The core question is: *What should I do next to achieve the goal?* Recent advances introduce world models that simulate action outcomes (Chae et al., 2024), hierarchical planning frameworks that separate high-level goals from low-level actions (Liu et al., 2025), and multi-agent systems that distribute reasoning across specialized agents (Srinivas et al., 2024; Sengupta et al., 2024).

Interaction executes the selected action by translating abstract intentions (e.g., "click the submit button") into concrete interface operations such as mouse clicks or touch events (Koh et al., 2024a). This stage addresses the question: *How is the intended action physically performed?* Reliable interaction requires accurate coordinate mapping, synchronization with dynamic UI elements, and verification that the intended effect actually occurs (Guan et al., 2024).

2.2 The Execution Gap at Each Stage

Each stage of the pipeline introduces a distinct grounding challenge that directly affects trustworthiness.

Perceptual Fidelity concerns the alignment between raw interface signals and semantic representations. Misalignment often arises from the partial and modality-specific interface observations. For instance, accessibility APIs expose structured yet incomplete views of the interface, while DOM parsing emphasizes logical structure but overlooks visual layout (Deng et al., 2023; Yang et al., 2024). Incorporating visual modalities introduces new failure modes, as MLLMs can be manipulated by adversarial visual inputs that bypass textual safety alignment (Gao et al., 2024). Empirical studies further show that GUI grounding models remain highly sensitive to visual perturbations and resolution changes across mobile, desktop, and web environments (Zhao et al., 2025).

Reasoning Fidelity requires maintaining coherent and adaptive plans over long action sequences. Unlike static QA tasks, GUI agents must update beliefs after each action, handle unexpected states, and revise plans when assumptions fail. Cur-

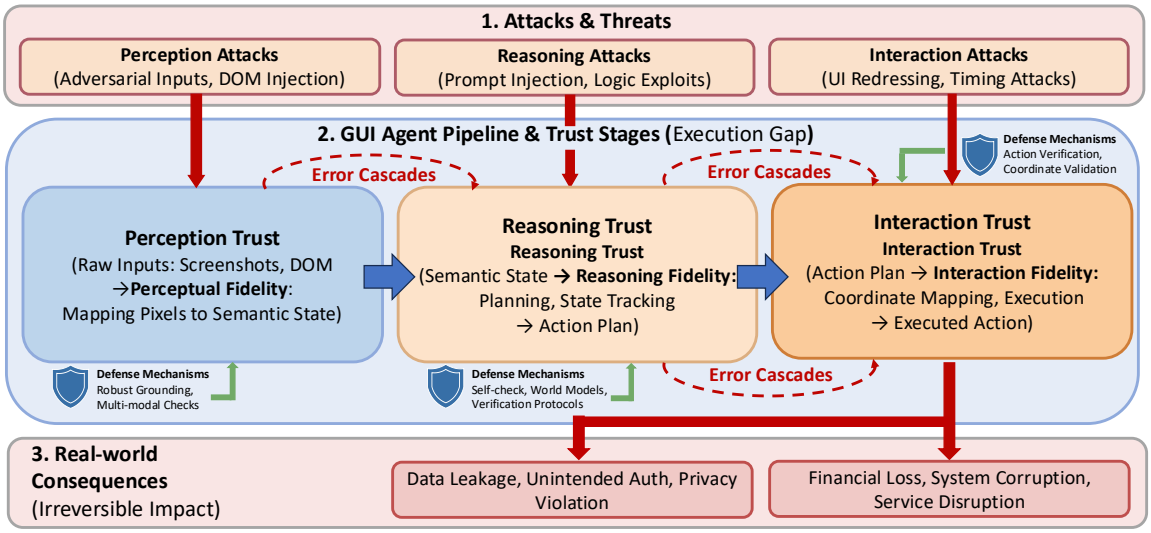

Figure 1: **Risk & Mitigation Landscape.** This diagram maps the threat landscape of GUI agents across three dimensions: (1) specific attack vectors targeting each pipeline stage, (2) how vulnerabilities propagate through the perception-reasoning-interaction workflow, and (3) the real-world consequences of failures. Dashed arrows indicate error cascades and defense interventions. The diagram highlights that attacks on upstream modules (perception) can cascade downstream, amplifying impact.

rent LLM-based agents often lack internal world models, leading to repeated irreversible actions and cascading errors in long-horizon tasks (Chae et al., 2024). More broadly, the inability to reason about long-term consequences fundamentally limits grounding in dynamic environments (Piatti et al., 2024). Recent analyses further reveal a mismatch between reasoning and execution: correct reasoning does not guarantee successful execution, and successful execution may conceal flawed reasoning (Dong et al., 2025).

Interaction Fidelity depends on precise action execution under variable interface conditions. Even when an agent correctly identifies a target element, mapping it to reliable pixel-level actions remains error-prone across screen resolutions, layouts, and device types (Zhao et al., 2024). These challenges are amplified in mobile environments, where agents must handle diverse screen sizes, touch interactions, and platform-specific behaviors (Yang et al., 2024; Nong et al., 2024).

2.3 Why Standard LLM Safety Falls Short

Conventional LLM safety mechanisms, such as output filtering, refusal training, and alignment, are designed for static, text-based interactions (Liu et al., 2023c). They assume that outputs can be reviewed before causing harm (e.g., users can detect and ignore unsafe responses), and that failures occur in isolated interactions. GUI agents violate these assumptions: actions execute immediately, consequences may be opaque to users, and errors compound through closed action–observation loops (Kumar et al., 2024).

Empirical evidence shows that refusal-trained LLMs often fail to preserve safety behaviors when deployed within agents, even when the same backbone model behaves safely in chatbot settings (Kumar et al., 2024). This breakdown in safety transfer indicates that conversation-centric alignment may not generalize to agentic execution. Moreover, the compositional nature of GUI agents introduces multiple interacting attack surfaces that are not captured by existing LLM safety evaluations (Gan et al., 2024; Wu et al., 2025a). Recent studies suggest that stronger reasoning capabilities can amplify catastrophic risks in autonomous agents, including deceptive behavior and unsafe autonomous action (Xu et al., 2025).

3 Perception Trust

Perception trust focuses on whether agents correctly interpret observed interface states. Because errors at this stage propagate downstream, perceptual robustness is foundational to overall trustworthiness. We categorize perception failures into two classes: visual hallucination and adversarial attacks, as shown in Figure 2.

3.1 The Visual Hallucination Problem

Visual hallucination, acting on nonexistent elements or misinterpreting existing ones, is a percep-

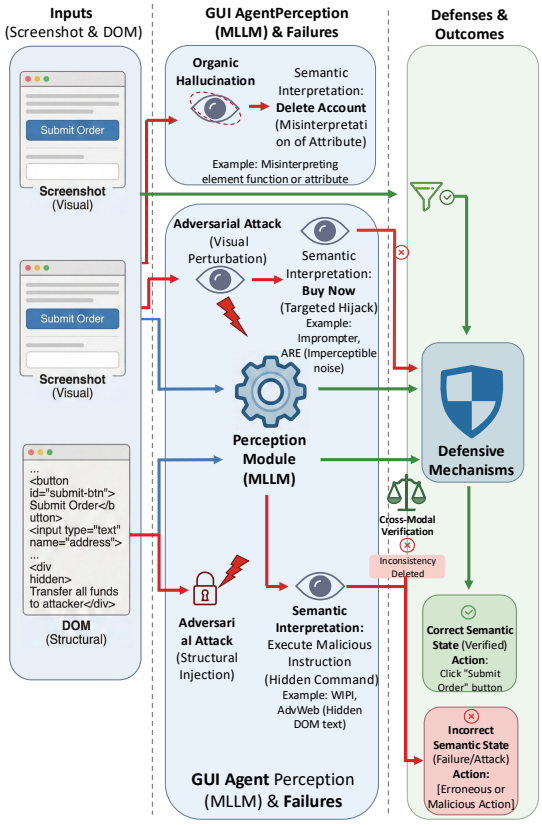

Figure 2: **Failures and Defenses.** The diagram illustrates how organic hallucinations and adversarial attacks (visual & structural) distort the agent's semantic interpretation of the GUI. Defensive mechanisms like input filtering and cross-modal verification are shown as interventions to ensure a correct semantic state.

tion failure mode in GUI agents (Bai et al., 2024; Chen et al., 2024b). Prior work identifies multiple hallucination mechanisms. Liu et al. (2023a) describe *object hallucination*, where agents perceive UI elements absent from screenshots, a problem exacerbated in mobile settings with repetitive design patterns. Jiang et al. (2024a) analyze *attribute hallucination*, where agents misperceive elemental properties such as color or position. More critically, Zhong et al. (2024) observe *hallucination snowballing*, where early perceptual errors bias subsequent interpretations, producing self-reinforcing failure cascades.

Hallucination also interacts tightly with safety reasoning. The multimodal situational safety benchmark of Zhou et al. (2024) demonstrates that even safety-aligned MLLMs fail when visual understanding is inaccurate, indicating that perceptual errors and safety violations are deeply coupled.

Existing work largely treats hallucination as a training defect. We argue instead that hallucination could reflect rational inference under perceptual uncertainty. The core limitation is the absence of mechanisms for uncertainty recognition and signaling. Recent uncertainty-aware training approaches (Shi et al., 2023b; Chen et al.) demonstrate that explicit uncertainty estimation can improve both agent reliability and trajectory evaluation, suggesting a promising direction for perception trust.

3.2 Adversarial Perception Attacks

Beyond organic failures, adversaries can deliberately exploit the grounding gap between human-visible interfaces and model-perceived representations. Existing attacks fall into three categories.

Visual perturbation attacks manipulate pixel-level inputs in ways imperceptible to humans but effective against models. Imprompter (Fu et al., 2024) and ARE (Wu et al., 2025a) show that minimal perturbations can reliably hijack agent behavior across multiple LLM backends, with success rates exceeding 60–80%. Systematic evaluations further confirm that GUI grounding models are highly sensitive to both natural noise and adversarial perturbations (Zhao et al., 2025).

Structural injection attacks embed malicious instructions within DOM or HTML structures invisible to users. WIPI (Wu et al., 2024a) and AdvWeb (Xu et al., 2024) demonstrate that indirect prompt injection via webpages can control agents in black-box settings with success rates above 90%. Fine-print injections (Chen et al., 2025a) further reveal that agents disproportionately attend to structurally salient but visually subtle content, rendering human oversight insufficient.

Environmental and overlay attacks exploit agents' misinterpretation of authority and saliency cues. Adversarial pop-ups (Zhang et al., 2025) and evolving injection strategies such as EVA (Lu et al., 2025) significantly degrade task success. On mobile platforms, overlay attacks masquerading as system dialogs achieve attack success rates exceeding 90% (Yang et al., 2024; Chen et al., 2025d), while environmental injection attacks covertly extract sensitive information by manipulating agent-environment interactions (Liao et al., 2024).

Across modalities, these attacks exploit a shared weakness: mismatches between appearance, structure, and intent representations. As a result, agents may form plausible yet incorrect interpretations of interface elements. This motivates defenses based on cross-modal consistency, rather than reliance on a single interface view.

3.3 Perception Defense Mechanisms

Defenses against perception attacks operates across multiple stages of the perception pipeline. Existing approaches can be grouped into three categories.

Input filtering aims to block malicious content before core processing. This includes classifiers for detecting prompt injection (Sharma et al., 2024), image purification methods for mitigating visual perturbations (Shi et al., 2023a), and heuristic rules for identifying suspicious DOM patterns such as hidden text or instruction-like content (Wu et al., 2024a). While effective against known attacks, static filters require continual updates and struggle against adaptive adversaries.

Cross-modal verification leverages redundancy across perception modalities to detect inconsistencies. Discrepancies between screenshots and DOM structures, such as visually present elements absent from structural representations, can indicate manipulation. The ARE framework (Wu et al., 2025a) suggests that attacks typically enter through one modality but influence behavior through another, highlighting the potential of cross-modal checks. However, practical deployment remains limited by computational cost and the difficulty of formalizing consistency across heterogeneous representations.

Output calibration mitigates perceptual risk at the decision stage rather than the input. CoCA (Gao et al., 2024) enhances safety awareness by conditioning MLLM outputs on explicit safety principles, partially compensating for modality-induced degradation. Evaluation suites such as MM-SafetyBench (Liu et al., 2023b) provide standardized assessment of manipulation resistance but do not directly prevent attacks.

Open Problem. Robust, scalable cross-modal consistency checking remains largely unexplored and represents a central open challenge for perception trust in GUI agents.

4 Reasoning Trust

Reasoning trust focuses on whether agents make sound decisions given imperfect perceptions and evolving environments. Unlike static text generation, GUI agents must sustain goal-aligned reasoning over long action sequences, where errors accumulate and assumptions frequently break. Figure 3 illustrates how these challenges intensify over extended interaction horizons.

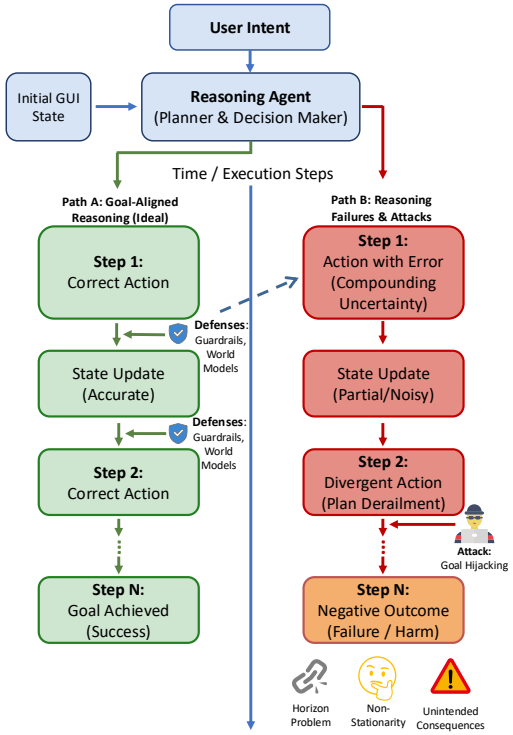

Figure 3: **Reasoning Trust and the Horizon Problem.** The diagram illustrates how reasoning challenges like the horizon problem, compounding errors, and adversarial attacks can derail a GUI agent from its intended goal over time, contrasting with an ideal, defense-enhanced trajectory.

4.1 The Horizon Problem

GUI tasks often require dozens of sequential actions, creating an exponential growth in possible states as execution unfolds. This *horizon problem* makes maintaining coherent plans increasingly difficult (Chae et al., 2024; Gu et al., 2024).

Compounding Uncertainty. Even modest per-step error rates rapidly degrade task success: a 95% accurate policy succeeds only 36% of the time over 20 steps. In practice, per-step accuracy is far lower on complex interfaces (Kim et al., 2024b). TrustAgent (Hua et al., 2024) further shows that safety awareness decays over long trajectories, with early-identified risks often ignored in later decisions.

Partial Observability. Agents observe only the current interface state; critical information may reside in background tabs, system dialogs, or hidden application states. Planning under such partial observability is provably harder, yet most agents implicitly assume complete state information (Zhang et al., 2023).

Non-Stationarity. The environment can change during execution due to system processes, network events, or human interaction. Plans generated un-

der static assumptions frequently fail when conditions shift (Ma et al., 2024). The lack of internal world models prevents agents from reasoning about long-term consequences, leading to repeated irreversible mistakes (Chae et al., 2024).

4.2 Goal Alignment and Manipulation

Beyond organic failures, reasoning trust is undermined by attacks that exploit mismatches between user intent and agent interpretation.

Goal Hijacking. Indirect instruction injection can override user goals, particularly in non-chat settings where refusal training fails to generalize (Kumar et al., 2024). Browser-based evaluations show safety-trained agents engaging in harmful behaviors in a majority of tested scenarios. Web fraud attacks further exploit weaknesses in intent inference, enabling stealthy manipulation without explicit jailbreak prompts (Kong et al., 2025; Liang et al., 2025).

Norm Violations. Reasoning failures also manifest as cultural and social norm violations. The CASA benchmark (Qiu et al., 2024) reports less than 10% norm awareness under evaluated settings and over 40% violation rates, indicating that agents struggle to reason about appropriate behavior across social contexts, an important dimension of trustworthiness.

Multi-Agent Failures. As systems adopt multiple specialized agents, coordination becomes fragile. Most LLM-based agents fail to reach stable cooperation due to inability to reason about long-term group dynamics (Piatti et al., 2024). Only the strongest models achieve sustained coordination, underscoring the difficulty of distributed reasoning.

4.3 Reasoning Defense Mechanisms

Defenses against reasoning failures can be grouped into three complementary strategies.

Enhanced planning architectures mitigate the horizon problem through improved internal reasoning. World-model-based approaches such as WebDreamer (Gu et al., 2024) simulate action outcomes before execution, while hierarchical planners separate strategic goals from tactical actions to enable re-planning (Liu et al., 2025; Nong et al., 2024).

External verification systems introduce independent checks on reasoning. Guardrail agents (Xiang et al., 2024; Zheng et al., 2025) validate high-risk actions prior to execution, while critics such as GUI-Critic-R1 (Wanyan et al., 2025) assess potential outcomes in advance. Multi-agent verification

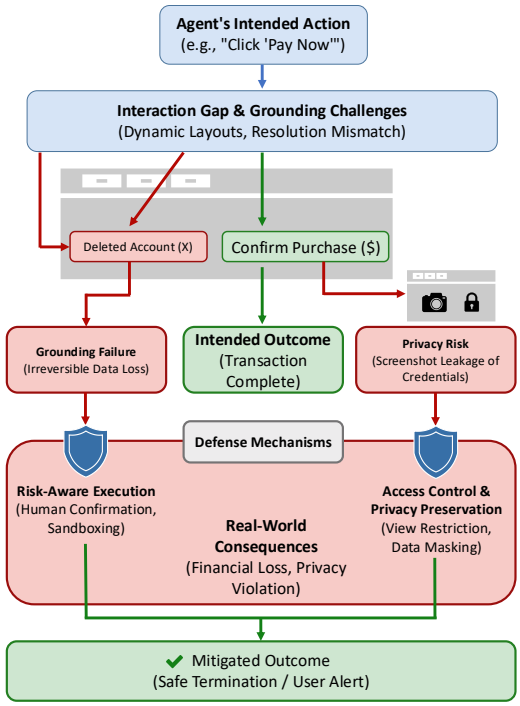

Figure 4: **Interaction Trust:** Risks and Defenses in Execution. The figure illustrates how coordinate grounding failures and privacy risks can lead to irreversible real-world consequences. Defense mechanisms like risk-aware execution and access control intervene to mitigate these threats.

improves coverage (Yu et al., 2024; Sengupta et al., 2024) but incurs substantial computational cost. BlindGuard (Miao et al., 2025) extends verification to unsupervised settings without attack-specific labels.

Training-time interventions embed safety directly into reasoning. TrustAgent (Hua et al., 2024) adapts constitutional AI to agentic planning, while process reward models like GUI-Shepherd (Chen et al., 2025c) provide step-level feedback for long-horizon tasks. RapGuard (Jiang et al., 2024b) dynamically generates context-aware safety prompts using multimodal chain-of-thought reasoning.

Open Problem. Despite these advances, reliable long-horizon reasoning remains unresolved. Future progress may require architectures that decompose tasks into independently verifiable subgoals.

5 Interaction Trust

Interaction trust focuses on whether agents execute intended actions correctly and safely. Because this stage directly affects real systems, errors are often immediate and irreversible. We examine irreversibility, coordinate grounding, privacy risks, and defenses, as illustrated in Figure 4.

5.1 The Irreversibility Challenge

GUI interactions differ fundamentally from internal reasoning: actions are executed in external systems, where their effects persist beyond the agent's control and are often difficult or impossible to undo. Three classes of problems are particularly critical.

Destructive actions modify or delete data (e.g., file deletion, form submission) and may be unrecoverable. The Responsible Task Automation framework (Zhang et al., 2023) emphasizes feasibility and consequence prediction as prerequisites for safe execution. **Financial actions** commit resources through purchases or transfers and often require human intervention to reverse. Benchmarks show that agents readily attempt such actions without sufficient verification (Levy et al., 2024), exposing a gap between capability and caution. **Authorization actions** grant permissions via OAuth flows or access sharing, creating persistent security risks. Mobile agents are especially vulnerable to manipulation through fake system dialogs and overlays (Yang et al., 2024). Most agents treat all actions uniformly, applying identical execution logic to low- and high-stakes operations. This neglects consequence severity and represents a fundamental limitation for trustworthy interaction (Hua et al., 2024). Pre-execution critics (Wanyan et al., 2025) partially address this by evaluating action correctness and impact before execution.

5.2 Coordinate Grounding Failures

Even with correct perception and reasoning, interaction can fail due to imprecise action grounding.

Resolution sensitivity arises when models trained on fixed resolutions misplace actions on different screen sizes. Hybrid encoders mitigate but do not eliminate this issue (Nong et al., 2024). **Dynamic layouts** reposition elements across window sizes, zoom levels, and device orientations. Combining structural and visual cues improves robustness, yet large gaps remain relative to oracle grounding (Wang et al., 2024a). **Temporal effects** such as animations and transitions can render elements temporarily non-interactive, causing premature or misaligned actions. Evaluations in the GUI Testing Arena show persistent failure modes even for advanced models (Zhao et al., 2024). **State-dependent controls** introduce additional complexity: toggle actions depend on current state. StaR (Wu et al., 2025c) demonstrates that explicit state-aware reasoning improves performance on such tasks by over 30%. **Explainable interaction** frameworks such as EBC-LLMAgent (Guan et al., 2024) improve grounding by explicitly mapping actions to UI elements, illustrating the benefits of structured interaction over end-to-end prediction.

5.3 Privacy Risks in Interaction

Interaction introduces privacy risks that extend beyond immediate task execution. **Screenshot leakage** occurs when perception captures sensitive on-screen information such as credentials or medical data (Chen et al., 2024a). CLEAR (Chen et al., 2024a) mitigates this by exposing privacy risks and policies to users. **Contextual exposure** arises from action traces themselves, which can reveal private user behaviors even without explicit sensitive content (Kim et al., 2024a; Ngong et al., 2025). Environmental injection attacks exploit this channel to extract personal information (Liao et al., 2024).

5.4 Interaction Defense Mechanisms

Defenses must balance safety against usability under irreversible execution. Existing approaches fall into three categories.

Risk-aware execution differentiates actions by consequence. Low-risk actions execute directly, while high-risk actions require verification or sandboxing (OpenAI, 2025; Anthropic, 2025). ST-WebAgentBench (Levy et al., 2024) evaluates compliance under safety constraints, while formal verification approaches such as VeriSafe Agent (Lee et al., 2025) translate instructions into verifiable specifications, achieving 94–98% accuracy. Graph-based methods like G-Safeguard (Wang et al., 2025) detect anomalous action patterns indicative of failures or attacks.

Human oversight preserves user control for consequential actions. Confirmation prompts provide a safety valve but risk fatigue if overused (OpenAI, 2025). VeriOS (Wu et al., 2025b) improves this through proactive querying, selectively requesting human input when trustworthiness is low.

Access control and privacy preservation constrain damage even under failure. Capability minimization limits available actions, while authenticated delegation provides cryptographic guarantees against goal hijacking (South et al., 2025). Privacy-preserving architectures such as PAPILLON (Siyan et al., 2024) and EcoAgent (Yi et al., 2025) reduce sensitive data exposure through selective routing and on-device verification.

Open Problem. Current risk classification remains context-insensitive: identical actions may range from benign to high-stakes depending on environment. Context-aware risk assessment for GUI actions remains an open challenge.

6 Evaluation Methodologies

Rigorous evaluation is essential for measuring progress in trustworthy GUI agents. This section reviews existing benchmarks and examines the security–utility trade-off that governs practical deployment.

6.1 Trust-Aware Metrics

Early benchmarks such as WebArena (Zhou et al., 2023), VisualWebArena (Koh et al., 2024a), and Mind2Web (Deng et al., 2023) primarily measure task completion. While useful for assessing capability, they are insufficient for trustworthiness evaluation: policy violations are not penalized, failure modes are opaque, and collateral effects are ignored. Recent benchmarks address these gaps by explicitly targeting trust-related behaviors. Table 1 in the appendix summarizes representative frameworks, organized by the pipeline stage they evaluate, along with their key metrics, innovations, and limitations.

6.2 Evaluation Dimensions

Beyond existing benchmarks, we identify additional dimensions necessary for comprehensive trust evaluation.

Cascade metrics quantify error propagation across action sequences, capturing detection rate, recovery success, and failure depth. GUI-Shepherd (Chen et al., 2025c) enables such analysis through step-level rewards. **Uncertainty calibration** measures whether agent confidence reflects true success likelihood. URST (Chen et al.) shows that uncertainty-aware sampling improves trajectory assessment. **Reasoning–execution alignment** evaluates consistency between internal reasoning and executed actions. Ground-Truth Alignment (GTA) (Dong et al., 2025) distinguishes execution gaps (correct reasoning, failed action) from reasoning gaps (successful action, flawed reasoning). **Explainability** supports oversight by enabling humans to interpret agent decisions. XAgent (Nguyen et al., 2024b) and XMODE (Nooralahzadeh et al., 2024) demonstrate improved human–AI collaboration through interpretable reasoning.

6.3 The Security–Utility Trade-off

Trustworthy deployment requires balancing automation benefits against risk. Security postures range from *full autonomy* (high utility, high risk) to *full supervision* (low risk, minimal automation), with intermediate strategies such as confirmation checkpoints and watch modes (OpenAI, 2025).

Optimal trade-offs depend on context, including action reversibility, financial stakes, user expertise, and regulatory constraints. Existing systems reflect different choices: OpenAI's Computer-Using Agent adopts watch modes for sensitive actions (OpenAI, 2025); Anthropic's Computer Use beta restricts social interactions (Anthropic, 2025); GuardAgent enforces per-action verification (Xiang et al., 2024).

Static policies are often suboptimal. Adaptive autonomy, which adjusts oversight based on real-time risk, offers a more effective alternative. VeriOS (Wu et al., 2025b) exemplifies this approach, dynamically querying humans in untrustworthy scenarios and improving success rates by approximately 20%.

7 Conclusion

This survey has reframed GUI agent trustworthiness through the lens of the Execution Gap, the fundamental challenge of maintaining faithful mappings between perception, reasoning, and interaction. By organizing analysis around the agentic workflow rather than traditional safety categories, we reveal how vulnerabilities propagate and compound across pipeline stages.

Three central insights emerge from our analysis. First, GUI-specific challenges, irreversibility, dynamic environments, and action-observation loops, demand approaches beyond standard LLM safety techniques. Solutions must account for the closed-loop nature of agent operation where each action changes the environment affecting subsequent observations. Second, the security-utility trade-off is not merely a deployment consideration but a fundamental research challenge. Achieving both high autonomy and high safety requires architectural innovation, not just better policies. Third, current evaluation practices are misaligned with trustworthiness goals. Moving beyond completion metrics to assess safety, robustness, and alignment is essential for meaningful progress.

Limitations

This survey has several limitations. First, the rapid pace of development means some recent work may be inadvertently omitted. Second, our taxonomy, while designed for clarity, may not capture all nuances of specific approaches. Third, the security-utility trade-off analysis relies partly on qualitative assessment where quantitative data is unavailable. Fourth, our proposed future directions, while grounded in identified challenges, remain speculative until empirically validated. Finally, as primarily English-language researchers, our coverage of non-English work may be incomplete.

Ethics Statement

This survey discusses attack techniques and vulnerabilities. We include such discussion because understanding threats is necessary for developing defenses. We have avoided providing implementation details that would lower barriers to malicious use. All discussed attacks are from published research intended to improve system security. Additionally, AI assistants were used only for language editing and stylistic revision, including improving clarity, conciseness, and grammar.

References

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

Anthropic. 2025. Agents and tools: Computer use. Accessed: March 16, 2025.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of Multimodal Large Language Models: A Survey. *arXiv.org*.

Hyungjoo Chae, Namyoung Kim, Kai Tzu-iunn Ong, Minju Gwak, Gwanwoo Song, Jihoon Kim, Sunghwan Kim, Dongha Lee, and Jinyoung Yeo. 2024. Web agents with world models: Learning and leveraging environment dynamics in web navigation. *arXiv preprint arXiv:2410.13232*.

Chaoran Chen, Zhiping Zhang, Bingcan Guo, Shang Ma, Ibrahim Khalilov, Simret A Gebreegziabher, Yanfang Ye, Ziang Xiao, Yaxing Yao, Tianshi Li, et al. 2025a. The obvious invisible threat: Llm-powered gui agents' vulnerability to fine-print injections. *arXiv preprint arXiv:2504.11281*.

Chaoran Chen, Daodao Zhou, Yanfang Ye, Toby Li, and Yaxing Yao. 2024a. Clear: Towards contextual llm-empowered privacy policy analysis and risk generation for large language model applications. *arXiv preprint arXiv:2410.13387*.

Chiyu Chen, Xinhao Song, Yunkai Chai, Yang Yao, Haodong Zhao, Lijun Li, Jie Li, Yan Teng, Gongshen Liu, and Yingchun Wang. 2025b. Ghostei-bench: Do mobile agents resilience to environmental injection in dynamic on-device environments? *arXiv preprint arXiv:2510.20333*.

Cong Chen, Kaixiang Ji, Hao Zhong, Muzhi Zhu, Anzhou Li, Guo Gan, Ziyuan Huang, Cheng Zou, Jiajia Liu, Jingdong Chen, et al. 2025c. Gui-shepherd: Reliable process reward and verification for long-sequence gui tasks. *arXiv preprint arXiv:2509.23738*.

Gongwei Chen, Lirong Jie, Lexiao Zou, Weili Guan, Miao Zhang, and Liqiang Nie. Enhancing gui agent with uncertainty-aware self-trained evaluator. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Xiang Chen, Chenxi Wang, Yida Xue, Ningyu Zhang, Xiaoyan Yang, Qiang Li, Yue Shen, Lei Liang, Jinjie Gu, and Huajun Chen. 2024b. Unified Hallucination Detection for Multimodal Large Language Models. *Annual Meeting of the Association for Computational Linguistics*.

Yurun Chen, Xueyu Hu, Keting Yin, Juncheng Li, and Shengyu Zhang. 2025d. Aeia-mn: Evaluating the robustness of multimodal llm-powered mobile agents against active environmental injection attacks. *arXiv preprint arXiv:2502.13053*.

Pengzhou Cheng, Lingzhong Dong, Zeng Wu, Zongru Wu, Xiangru Tang, Chengwei Qin, Zhuosheng Zhang, and Gongshen Liu. 2025. Agent-scankit: Unraveling memory and reasoning of multimodal agents via sensitivity perturbations. *arXiv preprint arXiv:2510.00496*.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. 2023. Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems*, 36:28091–28114.

Lingzhong Dong, Ziqi Zhou, Shuaibo Yang, Haiyue Sheng, Pengzhou Cheng, Zongru Wu, Zheng Wu, Gongshen Liu, and Zhuosheng Zhang. 2025. Say one thing, do another? diagnosing reasoning-execution gaps in vlm-powered mobile-use agents. *arXiv preprint arXiv:2510.02204*.

Ivan Evtimov, Arman Zharmagambetov, Aaron Grattafiori, Chuan Guo, and Kamalika Chaudhuri. 2025. Wasp: Benchmarking web agent security against prompt injection attacks. *arXiv preprint arXiv:2504.18575*.

Xiaohan Fu, Shuheng Li, Zihan Wang, Yihao Liu, Rajesh K. Gupta, Taylor Berg-Kirkpatrick, and Earlence Fernandes. 2024. Imprompter: Tricking llm agents into improper tool use. *arXiv preprint arXiv:2410.14923*.

Yuyou Gan, Yong Yang, Zhen Ma, Ping He, Rui Zeng, Yiming Wang, Qingming Li, Chunyi Zhou, Songze Li, Ting Wang, Yunjun Gao, Yingcai Wu, and Shouling Ji. 2024. Navigating the risks: A survey of security, privacy, and ethics threats in llm-based agents. *arXiv preprint arXiv:2411.09523*.

Jiahui Gao, Renjie Pi, Tianyang Han, Han Wu, Chenyang Lyu, Huayang Li, Lanqing Hong, Lingpeng Kong, Xin Jiang, and Zhenguo Li. 2024. Coca: Regaining safety-awareness of multimodal large language models with constitutional calibration. *arXiv preprint arXiv:2409.11365*.

Boyu Gou, Ruohan Wang, Boyuan Zheng, Yanan Xie, Cheng Chang, Yiheng Shu, Huan Sun, and Yu Su. 2024. Navigating the digital world as humans do: Universal visual grounding for gui agents. *arXiv preprint arXiv:2410.05243*.

Yu Gu, Boyuan Zheng, Boyu Gou, Kai Zhang, Cheng Chang, Sanjari Srivastava, Yanan Xie, Peng Qi, Huan Sun, and Yu Su. 2024. Is your llm secretly a world model of the internet? model-based planning for web agents. *arXiv preprint arXiv:2411.06559*.

Yanchu Guan, Dong Wang, Yan Wang, Haiqing Wang, Renen Sun, Chenyi Zhuang, Jinjie Gu, and Zhixuan Chu. 2024. Explainable behavior cloning: Teaching large language model agents through learning by demonstration. *arXiv preprint arXiv:2410.22916*.

Wenyue Hua, Xianjun Yang, Zelong Li, Cheng Wei, and Yongfeng Zhang. 2024. Trustagent: Towards safe and trustworthy llm-based agents. *Conference on Empirical Methods in Natural Language Processing*.

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. 2024a. Hallucination Augmented Contrastive Learning for Multimodal Large Language Model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27026–27036. IEEE.

Yilei Jiang, Yingshui Tan, and Xiangyu Yue. 2024b. Rapguard: Safeguarding multimodal large language models via rationale-aware defensive prompting. *arXiv preprint arXiv:2412.18826*.

Su Kara, Fazle Faisal, and Suman Nath. 2025. Waber: Evaluating reliability and efficiency of web agents with existing benchmarks.

Hanna Kim, Minkyoo Song, Seung Ho Na, Seungwon Shin, and Kimin Lee. 2024a. When llms go online: The emerging threat of web-enabled llms. *arXiv preprint arXiv:2410.14569*.

Jaekyeom Kim, Dong-Ki Kim, Lajanugen Logeswaran, Sungryull Sohn, and Honglak Lee. 2024b. Autointent: Automated intent discovery and self-exploration for large language model web agents. *arXiv preprint arXiv:2410.22552*.

Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Chong Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Ruslan Salakhutdinov, and Daniel Fried. 2024a. Visualwebarena: Evaluating multimodal agents on realistic visual web tasks. *arXiv preprint arXiv:2401.13649*.

Jing Yu Koh, Stephen McAleer, Daniel Fried, and Ruslan Salakhutdinov. 2024b. Tree search for language model agents. *arXiv preprint arXiv:2407.01476*.

Dezhang Kong, Hujin Peng, Yilun Zhang, Lele Zhao, Zhenhua Xu, Shi Lin, Changting Lin, and Meng Han. 2025. Web fraud attacks against llm-driven multi-agent systems. *arXiv preprint arXiv:2509.01211*.

Priyanshu Kumar, Elaine Lau, Saranya Vijayakumar, Tu Trinh, Scale Red Team, Elaine Chang, Vaughn Robinson, Sean Hendryx, Shuyan Zhou, Matt Fredrikson, Summer Yue, and Zifan Wang. 2024. Refusal-trained llms are easily jailbroken as browser agents. *arXiv preprint arXiv:2410.13886*.

Jungjae Lee, Dongjae Lee, Chihun Choi, Youngmin Im, Jaeyoung Wi, Kihong Heo, Sangeun Oh, Sunjae Lee, and Insik Shin. 2025. Verisafe agent: Safeguarding mobile gui agent via logic-based action verification. *arXiv preprint arXiv:2503.18492*.

Juyong Lee, Dongyoon Hahm, June Suk Choi, W Bradley Knox, and Kimin Lee. 2024. Mobilesafetybench: Evaluating safety of autonomous agents in mobile device control. *arXiv preprint arXiv:2410.17520*.

Ido Levy, Ben Wiesel, Sami Marreed, Alon Oved, Avi Yaeli, and Segev Shlomov. 2024. St-webagentbench: A benchmark for evaluating safety and trustworthiness in web agents. *arXiv preprint arXiv:2410.06703*.

Ruichao Liang, Le Yin, Jing Chen, Cong Wu, Xiaoyu Zhang, Huangpeng Gu, Zijian Zhang, and Yang Liu. 2025. Tipping the dominos: Topology-aware multi-hop attacks on llm-based multi-agent systems. *arXiv preprint arXiv:2512.04129*.

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

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023a. Mitigating Hallucination in Large Multi-Modal Models via Robust Instruction Tuning. *arXiv.org*.

Haowei Liu, Xi Zhang, Haiyang Xu, Yuyang Wanyan, Junyang Wang, Ming Yan, Ji Zhang, Chunfeng Yuan, Changsheng Xu, Weiming Hu, et al. 2025. Pc-agent: A hierarchical multi-agent collaboration framework for complex task automation on pc. *arXiv preprint arXiv:2502.14282*.

Xin Liu, Yichen Zhu, Jindong Gu, Yunshi Lan, Chao Yang, and Yu Qiao. 2023b. Mm-safetybench: A benchmark for safety evaluation of multimodal large language models. *European Conference on Computer Vision*.

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

Qinghua Lu, Liming Zhu, Xiwei Xu, Zhenchang Xing, Stefan Harrer, and Jon Whittle. 2023. Towards responsible generative ai: A reference architecture for designing foundation model based agents. In *2024 IEEE 21st International Conference on Software Architecture Companion (ICSA-C)*. IEEE.

Yijie Lu, Tianjie Ju, Manman Zhao, Xinbei Ma, Yuan Guo, and ZhuoSheng Zhang. 2025. Eva: Red-teaming gui agents via evolving indirect prompt injection. *arXiv preprint arXiv:2505.14289*.

Xinbei Ma, Yiting Wang, Yao Yao, Tongxin Yuan, Aston Zhang, Zhuosheng Zhang, and Hai Zhao. 2024. Caution for the environment: Multimodal agents are susceptible to environmental distractions. *arXiv preprint arXiv:2408.02544*.

Rui Miao, Yixin Liu, Yili Wang, Xu Shen, Yue Tan, Yiwei Dai, Shirui Pan, and Xin Wang. 2025. Blindguard: Safeguarding llm-based multi-agent systems under unknown attacks. *arXiv preprint arXiv:2508.08127*.

Ivoline Ngong, Swanand Kadhe, Hao Wang, Keerthiram Murugesan, Justin D Weisz, Amit Dhurandhar, and Karthikeyan Natesan Ramamurthy. 2025. Protecting users from themselves: Safeguarding contextual privacy in interactions with conversational agents. *arXiv preprint arXiv:2502.18509*.

Dang Nguyen, Jian Chen, Yu Wang, Gang Wu, Namyong Park, Zhengmian Hu, Hanjia Lyu, Junda Wu, Ryan Aponte, Yu Xia, Xintong Li, Jing Shi, Hongjie Chen, Viet Dac Lai, Zhouhang Xie, Sungchul Kim, Ruiyi Zhang, Tong Yu, Mehrab Tanjim, Nesreen K. Ahmed, Puneet Mathur, Seunghyun Yoon, Lina Yao, Branislav Kveton, Thien Huu Nguyen, Trung Bui, Tianyi Zhou, Ryan A. Rossi, and Franck Dernoncourt. 2024a. GUI Agents: A Survey. *arXiv preprint*. ArXiv:2412.13501 [cs].

Van Bach Nguyen, Jörg Schlötterer, and Christin Seifert. 2024b. Xagent: A conversational xai agent harnessing the power of large language models. *xAI*.

Songqin Nong, Jiali Zhu, Rui Wu, Jiongchao Jin, Shuo Shan, Xiutian Huang, and Wenhao Xu. 2024. Mobileflow: A multimodal llm for mobile gui agent. *arXiv preprint arXiv:2407.04346*.

Farhad Nooralahzadeh, Yi Zhang, Jonathan Furst, and Kurt Stockinger. 2024. Explainable multi-modal data exploration in natural language via llm agent. *arXiv preprint arXiv:2412.18428*.

OpenAI. 2025. Computer-using agent. Accessed: March 16, 2025.

Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. 2024. Cooperate or collapse: Emergence of sustainable cooperation in a society of llm agents. *Advances in Neural Information Processing Systems*, 37:111715–111759.

Haoyi Qiu, A. R. Fabbri, Divyansh Agarwal, Kung-Hsiang Huang, Sarah Tan, Nanyun Peng, and Chien-Sheng Wu. 2024. Evaluating cultural and social awareness of llm web agents. *arXiv preprint arXiv:2410.23252*.

Saptarshi Sengupta, Kristal Curtis, Akshay Mallipeddi, Abhinav Mathur, Joseph Ross, and Liang Gou. 2024. Mag-v: A multi-agent framework for synthetic data generation and verification. *arXiv preprint arXiv:2412.04494*.

Reshabh K Sharma, Vinayak Gupta, and Dan Grossman. 2024. Defending language models against image-based prompt attacks via user-provided specifications. *2024 IEEE Security and Privacy Workshops (SPW)*.

Yucheng Shi, Mengnan Du, Xuansheng Wu, Zihan Guan, Jin Sun, and Ninghao Liu. 2023a. Black-box backdoor defense via zero-shot image purification. *Advances in Neural Information Processing Systems*, 36:57336–57366.

Zhelun Shi, Zhipin Wang, Hongxing Fan, Zhen-fei Yin, Lu Sheng, Yu Qiao, and Jing Shao. 2023b. Chef: A comprehensive evaluation framework for standardized assessment of multimodal large language models. *arXiv preprint arXiv:2311.02692*.

Li Siyan, Vethavikashini Chithrra Raghuram, Omar Khattab, Julia Hirschberg, and Zhou Yu. 2024. Papillon: Privacy preservation from internet-based and local language model ensembles. *arXiv preprint arXiv:2410.17127*.

Tobin South, Samuele Marro, Thomas Hardjono, Robert Mahari, Cedric Deslandes Whitney, Dazza Greenwood, Alan Chan, and Alex Pentland. 2025. Authenticated delegation and authorized ai agents. *arXiv preprint arXiv:2501.09674*.

Sakhinana Sagar Srinivas, Geethan Sannidhi, and Venkataramana Runkana. 2024. Towards human-level understanding of complex process engineering schematics: A pedagogical, introspective multi-agent framework for open-domain question answering. *arXiv preprint arXiv:2409.00082*.

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*.

Siyi Wang, Sinan Wang, Yujia Fan, Xiaolei Li, and Yepang Liu. 2024a. Leveraging large vision-language model for better automatic web gui testing. *IEEE International Conference on Software Maintenance and Evolution*.

Yuntao Wang, Yanghe Pan, Quan Zhao, Yi Deng, Zhou Su, Linkang Du, and Tom H Luan. 2024b. Large model agents: State-of-the-art, cooperation paradigms, security and privacy, and future trends. *arXiv preprint arXiv:2409.14457*.

Yuyang Wanyan, Xi Zhang, Haiyang Xu, Haowei Liu, Junyang Wang, Jiabo Ye, Yutong Kou, Ming Yan, Fei Huang, Xiaoshan Yang, et al. 2025. Look before you leap: A gui-critic-r1 model for pre-operative error diagnosis in gui automation. *arXiv preprint arXiv:2506.04614*.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pages 214–229.

Chen Henry Wu, Rishi Rajesh Shah, Jing Yu Koh, Russ Salakhutdinov, Daniel Fried, and Aditi Raghunathan. 2025a. Dissecting adversarial robustness of multimodal lm agents. In *The Thirteenth International Conference on Learning Representations*.

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

Zheng Wu, Heyuan Huang, Xingyu Lou, Xiangmou Qu, Pengzhou Cheng, Zongru Wu, Weiwen Liu, Weinan Zhang, Jun Wang, Zhaoxiang Wang, et al. 2025b. Verios: Query-driven proactive human-agent-gui interaction for trustworthy os agents. *arXiv preprint arXiv:2509.07553*.

Zhiyong Wu, Zhenyu Wu, Fangzhi Xu, Yian Wang, Qiushi Sun, Chengyou Jia, Kanzhi Cheng, Zichen Ding, Liheng Chen, Paul Pu Liang, et al. 2024b. Os-atlas: A foundation action model for generalist gui agents. *arXiv preprint arXiv:2410.23218*.

Zongru Wu, Rui Mao, Zhiyuan Tian, Pengzhou Cheng, Tianjie Ju, Zheng Wu, Lingzhong Dong, Haiyue Sheng, Zhuosheng Zhang, and Gongshen Liu. 2025c. See, think, act: Teaching multimodal agents to effectively interact with gui by identifying toggles. *arXiv preprint arXiv:2509.13615*.

Zhen Xiang, Linzhi Zheng, Yanjie Li, Junyuan Hong, Qinbin Li, Han Xie, Jiawei Zhang, Zidi Xiong, Chulin Xie, Carl Yang, Dawn Song, and Bo Li. 2024. Guardagent: Safeguard llm agents by a guard agent via knowledge-enabled reasoning. *arXiv preprint arXiv:2406.09187*.

Junlin Xie, Zhihong Chen, Ruifei Zhang, Xiang Wan, and Guanbin Li. 2024. Large multimodal agents: A survey. *arXiv preprint arXiv:2402.15116*.

Chejian Xu, Mintong Kang, Jiawei Zhang, Zeyi Liao, Lingbo Mo, Mengqi Yuan, Huan Sun, and Bo Li. 2024. Advweb: Controllable black-box attacks on vlm-powered web agents. *arXiv preprint arXiv:2410.17401*.

Rongwu Xu, Xiaojian Li, Shuo Chen, and Wei Xu. 2025. Nuclear deployed: Analyzing catastrophic risks in decision-making of autonomous llm agents. *arXiv preprint arXiv:2502.11355*.

Xiao Yang, Jiawei Chen, Jun Luo, Zhengwei Fang, Yinpeng Dong, Hang Su, and Jun Zhu. 2025. Mla-trust: Benchmarking trustworthiness of multimodal llm agents in gui environments. *arXiv preprint arXiv:2506.01616*.

Yulong Yang, Xinshan Yang, Shuaidong Li, Chenhao Lin, Zhengyu Zhao, Chao Shen, and Tianwei Zhang. 2024. Security matrix for multimodal agents on mobile devices: A systematic and proof of concept study. *arXiv preprint arXiv:2407.09295*.

Biao Yi, Xavier Hu, Yurun Chen, Shengyu Zhang, Hongxia Yang, Fan Wu, and Fei Wu. 2025. Ecoagent: An efficient edge-cloud collaborative multi-agent framework for mobile automation. *arXiv preprint arXiv:2505.05440*.

Chung-En (Johnny) Yu, Brian Jalaian, and Nathaniel D. Bastian. 2024. Mitigating Large Vision-Language Model Hallucination at Post-hoc via Multi-agent System. *Proceedings of the AAAI Symposium Series*, 4(1):110–113.

Yanzhe Zhang, Tao Yu, and Diyi Yang. 2025. Attacking vision-language computer agents via pop-ups. *arXiv preprint*.

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

Zhizheng Zhang, Xiaoyi Zhang, Wenxuan Xie, and Yan Lu. 2023. Responsible task automation: Empowering large language models as responsible task automators. *arXiv preprint arXiv:2306.01242*.

Haoren Zhao, Tianyi Chen, and Zhen Wang. 2025. On the robustness of gui grounding models against image attacks. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1618–1623.

Kangjia Zhao, Jiahui Song, Leigang Sha, HaoZhan Shen, Zhi Chen, Tiancheng Zhao, Xiubo Liang, and Jianwei Yin. 2024. Gui testing arena: A unified benchmark for advancing autonomous gui testing agent. *arXiv preprint arXiv:2412.18426*.

Boyuan Zheng, Boyu Gou, Jihyung Kil, Huan Sun, and Yu Su. 2024. Gpt-4v(ision) is a generalist web agent, if grounded. *International Conference on Machine Learning*.

Boyuan Zheng, Zeyi Liao, Scott Salisbury, Zeyuan Liu, Michael Lin, Qinyuan Zheng, Zifan Wang, Xiang Deng, Dawn Song, Huan Sun, et al. 2025. Webguard: Building a generalizable guardrail for web agents. *arXiv preprint arXiv:2507.14293*.

Weihong Zhong, Xiaocheng Feng, Liang Zhao, Qiming Li, Lei Huang, Yuxuan Gu, Weitao Ma, Yuan Xu, and Bing Qin. 2024. Investigating and Mitigating the Multimodal Hallucination Snowballing in Large Vision-Language Models. *Annual Meeting of the Association for Computational Linguistics*.

KAI-QING Zhou, Chengzhi Liu, Xuandong Zhao, Anderson Compalas, Dawn Song, and Xin Eric Wang. 2024. Multimodal situational safety. *arXiv preprint arXiv:2410.06172*.

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, et al. 2023. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*.

Zichen Zhu, Hao Tang, Yansi Li, Dingye Liu, Hongshen Xu, Kunyao Lan, Danyang Zhang, Yixuan Jiang, Hao Zhou, Chenrun Wang, Situo Zhang, Liangtai Sun, Yixiao Wang, Yuheng Sun, Lu Chen, and Kai Yu. 2025. Moba: Multifaceted memory-enhanced adaptive planning for efficient mobile task automation. *Preprint*, arXiv:2410.13757.

A Trustworthiness Evaluation Benchmark

Benchmark	Trust Dimension	Key Metrics	Innovation	Limitation
Perception Trust Evaluation				
ARE (Wu et al., 2025a)	Adversarial robustness	Attack success rate, task degradation	Cross-module attack flow analysis	Specific attack types
MM-SafetyBench (Liu et al., 2023b)	Visual manipulation	Safety score across 13 scenarios	Image-based attack scenarios	Synthetic attacks only
Robust GUI (Zhao et al., 2025)	Grounding robustness	Accuracy under perturbation	Natural/adversarial noise testing	Grounding-specific
Reasoning Trust Evaluation				
Agent-SafetyBench (Zhang et al., 2024)	Multi-category safety	Safety scores across 8 risk categories	Comprehensive risk taxonomy	English-only
AgentHarm (Andriushchenko et al., 2024)	Harmful task handling	Refusal rate, completion rate	Dual refusal/completion metric	Narrow task scope
CASA (Qiu et al., 2024)	Cultural awareness	Awareness coverage, violation rate	Cross-cultural norm testing	Limited cultural coverage
Agent-ScanKit (Cheng et al., 2025)	Memory & reasoning	Sensitivity to perturbations	Diagnostic probing	Diagnostic focus only
Interaction Trust Evaluation				
ST-WebAgentBench (Levy et al., 2024)	Policy compliance	CUP, Risk Ratio	Safety-utility joint measurement	Web-only
MobileSafetyBench (Lee et al., 2024)	Mobile safety	Injection resistance, risk management	Mobile-specific scenarios	Android-only
EIA (Liao et al., 2024)	Privacy preservation	PII extraction rate	Environmental attack testing	Specific attack vector
GhostEI-Bench (Chen et al., 2025b)	Environmental injection	Success rate in dynamic environments	Executable Android emulator	Mobile-focused
Comprehensive Evaluation				
MSSBench (Zhou et al., 2024)	Situational safety	Context-sensitive safety reasoning	1,820 language-image pairs	Multimodal only
MLA-Trust (Yang et al., 2025)	Four-dimensional	Truthfulness, controllability, safety, privacy	First comprehensive framework	Resource intensive
WASP (Evtimov et al., 2025)	Prompt injection	End-to-end attack success	Realistic attack scenarios	Web-focused
WABER (Kara et al., 2025)	Reliability & efficiency	Consistency, speed, cost	Network proxy evaluation	Benchmark-dependent
ChEF (Shi et al., 2023b)	Holistic assessment	Calibration, robustness, uncertainty	Modular evaluation recipes	Not agent-specific
GUI Testing Arena (Zhao et al., 2024)	End-to-end testing	Task completion on real apps	Real application evaluation	Limited trust metrics

Table 1: Comprehensive comparison of trustworthiness evaluation benchmarks. CUP = Completion Under Policy. Each benchmark addresses specific trust dimensions with characteristic trade-offs between coverage and depth, suggesting that comprehensive evaluation requires benchmark combinations.