Towards Trustworthy GUI Agents: A Survey

面向可信 GUI 代理的综述

Yucheng Shi [1] , Wenhao Yu [2] , Wenlin Yao [3] , Wenhu Chen [4] , Ninghao Liu [1]

石宇成 [1] , 余文浩 [2] , 姚文林 [3] , 陈文虎 [4] , 刘宁浩 [1]

[1] University of Georgia [2] Tencent AI Seattle Lab

[1] 佐治亚大学 [2] 腾讯 AI 西雅图实验室

[3] Amazon [4] University of Waterloo

[3] 亚马逊 [4] 滑铁卢大学

Abstract

摘要

GUI agents, powered by large foundation models, can interact with digital interfaces, enabling various applications in web automation, mobile navigation, and software testing. However, their increasing autonomy has raised critical concerns about their security, privacy, and safety. This survey examines the trustworthiness of GUI agents in five critical dimensions: security vulnerabilities, reliability in dynamic environments, transparency and explainability, ethical considerations, and evaluation methodologies. We also identify major challenges such as vulnerability to adversarial attacks, cascading failure modes in sequential decision-making, and a lack of realistic evaluation benchmarks. These issues not only hinder real-world deployment but also call for comprehensive mitigation strategies beyond task success. As GUI agents become more widespread, establishing robust safety standards and responsible development practices is essential. This survey provides a foundation for advancing trustworthy GUI agents through systematic understanding and future research.

基于大型基础模型驱动的 GUI 代理能够与数字界面交互，支持网页自动化、移动导航和软件测试等多种应用。然而，其日益增强的自主性也引发了关于安全性、隐私和安全保障的关键问题。本文综述了 GUI 代理可信性的五个关键维度: 安全漏洞、动态环境中的可靠性、透明性与可解释性、伦理考量以及评估方法。我们还指出了主要挑战，如对抗性攻击的脆弱性、序列决策中的级联故障模式以及缺乏现实评估基准。这些问题不仅阻碍了实际部署，也呼吁超越任务成功的全面缓解策略。随着 GUI 代理的广泛应用，建立稳健的安全标准和负责任的开发实践至关重要。本文为通过系统理解和未来研究推动可信 GUI 代理的发展奠定基础。

1 Introduction

1 引言

Large language models (LLMs) and large multimodal models (LMMs) have rapidly evolved from question answering tools to agents capable of interacting with graphical user interfaces (GUIs) through clicks and on-screen parsing (Nguyen et al., 2024a; Wang et al., 2024c; Xie et al., 2024). Deployed on websites, desktops, mobile apps, and diverse software environments, these GUI agents promise wide-ranging applications from automated testing and e-commerce to assistive technologies for users with disabilities (Zhao et al., 2024; Cuadra et al., 2024). Their ability to interpret dynamic interfaces, understand multimodal inputs, and execute precise actions is reshaping how large foundation models assist human operators in routine digital tasks.

大型语言模型 (LLMs) 和大型多模态模型 (LMMs) 已迅速从问答工具发展为能够通过点击和屏幕解析与图形用户界面 (GUI) 交互的代理 (Nguyen 等，2024a；Wang 等，2024c；Xie 等，2024)。这些 GUI 代理部署于网站、桌面、移动应用及多样软件环境，承诺在自动化测试、电子商务及辅助残障用户技术等领域实现广泛应用 (Zhao 等，2024；Cuadra 等，2024)。它们解读动态界面、理解多模态输入并执行精确操作的能力，正在重塑大型基础模型辅助人类完成日常数字任务的方式。

As GUI agents become more capable and begin to play a more significant role in real-world applications, ensuring their trustworthiness has become increasingly critical. Compared to traditional NLP tasks where inputs and outputs are relatively static and limited to textual data, GUI agents can operate in dynamic environments with inputs and outputs in different modalities. Although this flexibility improves utility, it also introduces new risks, which makes security, reliability, and transparency critical for responsible deployment (Arnold and Tilton, 2024; Ma et al., 2024). However, existing research on GUI agents focuses mainly on functional performance metrics, such as task completion rates, while often overlooking essential aspects like security, reliability, and transparency (Arnold and Tilton, 2024; Ma et al., 2024). This oversight poses significant risks, especially in high-stakes environments where these agents operate. Several emerging attacks have exposed these risks: adversarial image perturbations can deceive perception modules (Wu et al., 2025a), malicious webpage elements can manipulate agent behavior (Wu et al., 2024a), and screenshot-based navigation can inadvertently expose sensitive user data (Chen et al., 2024a).

随着 GUI 代理能力提升并在现实应用中扮演更重要角色，确保其可信性变得尤为关键。相比传统自然语言处理任务中输入输出相对静态且限于文本数据，GUI 代理可在动态环境中处理多模态输入输出。尽管这种灵活性提升了实用性，但也带来了新的风险，使安全性、可靠性和透明性成为负责任部署的关键 (Arnold 和 Tilton，2024；Ma 等，2024)。然而，现有关于 GUI 代理的研究主要关注功能性性能指标，如任务完成率，常忽视安全、可靠性和透明性等重要方面 (Arnold 和 Tilton，2024；Ma 等，2024)。这种忽视在高风险环境中尤为危险。多种新兴攻击揭示了这些风险: 对抗性图像扰动可欺骗感知模块 (Wu 等，2025a)，恶意网页元素可操控代理行为 (Wu 等，2024a)，基于截图的导航可能无意中泄露敏感用户数据 (Chen 等，2024a)。

Beyond these threats, the broader social implications are equally pressing. As the brain of GUI agents, the trustworthiness of LLMs and LMMs is crucial because it directly impacts the outcomes of GUI interactions (Liu et al., 2023c; Weidinger et al., 2022; Wu et al., 2024b). In critical domains such as finance and healthcare, trustworthy LLMs and LMMs ensure that decisions made by GUI agents are secure, ethical, transparent, and aligned with human values. Addressing these challenges requires assessing and mitigating risks from different aspects. Here, we categorize the trustworthiness in GUI agents into five key areas:

除上述威胁外，更广泛的社会影响同样紧迫。作为 GUI 代理“大脑”的 LLMs 和 LMMs 的可信性至关重要，因为它直接影响 GUI 交互的结果 (Liu 等，2023c；Weidinger 等，2022；Wu 等，2024b)。在金融和医疗等关键领域，可信的 LLMs 和 LMMs 确保 GUI 代理做出的决策安全、合乎伦理、透明且符合人类价值观。应对这些挑战需从多方面评估和缓解风险。本文将 GUI 代理的可信性划分为五个关键领域:

1. Security: Protecting agents against adversarial manipulation, unauthorized command execution, and data leaks. For instance, WebPI (Wu et al., 2024a) demonstrates how hidden HTML elements can mislead agents into executing unintended actions, posing security risks.

1. 安全性: 保护代理免受对抗性操控、未经授权的命令执行和数据泄露。例如，WebPI(Wu 等, 2024a) 展示了隐藏 HTML 元素如何误导代理执行非预期操作，带来安全风险。

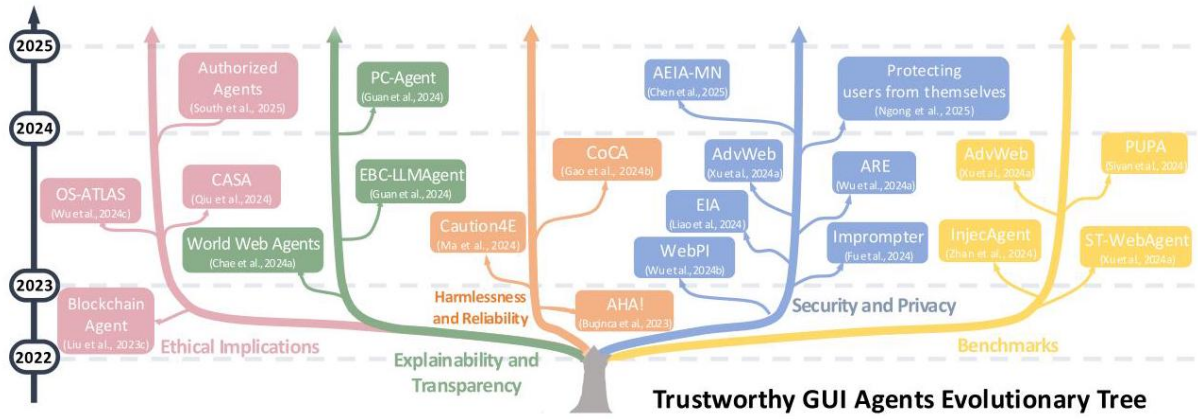

Figure 1: An evolutionary tree of research on trustworthy GUI agents. Each branch represents a research direction, with notable works color-coded by their focus area, demonstrating how the field has evolved toward more comprehensive trustworthiness considerations. This figure is adapted from this repo.

图 1: 可信 GUI 代理研究的演化树。每个分支代表一个研究方向，重点领域的代表性工作以颜色区分，展示该领域如何朝向更全面的可信性考量发展。此图改编自该代码库。

2. Reliability: Ensuring GUI agents function correctly across dynamic interfaces with reliable response. Studies on multimodal agent safety (Liu et al., 2023b) highlight risks where GUI agents may misinterpret visual cues, leading to unsafe or unintended interactions.

2. 可靠性: 确保 GUI 代理在动态界面中稳定运行并做出可靠响应。多模态代理安全性研究 (Liu 等, 2023b) 强调了 GUI 代理可能误解视觉线索，导致不安全或非预期交互的风险。

3. Explainability: Making agent decision-making processes more interpretable and user-friendly. Systems like EBC-LLMAgent (Guan et al., 2024) enhance transparency by learning from user demos to generate clear and interpretable action sequences with corresponding UI mappings and rationales.

3. 可解释性: 使代理的决策过程更加可理解和用户友好。像 EBC-LLMAgent(Guan 等，2024) 这样的系统通过学习用户演示，生成清晰且可解释的动作序列及相应的界面映射和理由，从而提升透明度。

4. Ethical Alignment: Ensuring agents adhere to human values and cultural norms. CASA (Qiu et al., 2024) evaluates agents on social and ethical considerations, emphasizing fairness in decision-making across diverse user populations.

4. 伦理对齐: 确保代理遵循人类价值观和文化规范。CASA(Qiu 等，2024) 评估代理在社会和伦理方面的表现，强调在多样化用户群体中的决策公平性。

5. Evaluation: Developing rigorous testing methods to assess GUI agent behavior under real-world conditions. ST-WebAgentBench (Levy et al., 2024) evaluates policy compliance and risk mitigation strategies for web-based agents.

5. 评估: 开发严格的测试方法以评估 GUI 代理在真实环境下的行为。ST-WebAgentBench(Levy 等，2024) 评估基于网络的代理的策略合规性和风险缓解策略。

Figure 1 illustrates the evolution of research across these dimensions. While previous surveys (Nguyen et al., 2024a; Wang et al., 2024a; Zhang et al., 2024a; Hu et al.; Liu et al., 2025b) primarily focus on the task performance of GUI agents, our work highlights less-explored issues like security, reliability, and transparency, and emerging mitigation strategies for these issues. Our discussion begins with an overview of GUI agent architectures and fundamental capabilities (Section 2). We then examine security and privacy challenges (Section 3) and strategies for enhancing reliability and harmlessness (Section 4). Next, we discuss the importance of explainability and transparency (Section 5) and outline ethical considerations for responsible deployment (Section 6). We conclude by reviewing evaluation methodologies (Section 7) and highlighting future research directions (Section 8). Overall, this survey shifts the focus from task success to holistic trustworthiness, offering researchers and developers insights into the risks, challenges, and solutions for creating secure and responsible GUI agents.

图 1 展示了这些维度的研究演进。尽管以往的综述 (Nguyen 等，2024a；Wang 等，2024a；Zhang 等，2024a；Hu 等；Liu 等，2025b) 主要关注 GUI 代理的任务性能，我们的工作强调了较少探讨的问题，如安全性、可靠性和透明性，以及针对这些问题的新兴缓解策略。我们的讨论从 GUI 代理架构和基本能力概述开始 (第 2 节)，随后考察安全与隐私挑战 (第 3 节) 及提升可靠性和无害性的策略 (第 4 节)。接着讨论可解释性和透明性的重要性 (第 5 节)，并概述负责任部署的伦理考量 (第 6 节)。最后回顾评估方法论 (第 7 节) 并指出未来研究方向 (第 8 节)。总体而言，本综述将关注点从任务成功转向整体可信性，为研究人员和开发者提供了构建安全且负责任 GUI 代理的风险、挑战与解决方案的洞见。

2 Foundation of GUI Agents

2 GUI 代理基础

GUI agents leverage large foundation models to integrate perception, reasoning, planning, and execution, enabling interaction with user interfaces in a human-like manner. This section outlines their key components,

applications, and challenges.

GUI 代理利用大型基础模型整合感知、推理、规划和执行，实现类人方式与用户界面交互。本节概述其关键组件、应用及挑战。

A standard agent pipeline includes multimodal perception, reasoning and planning (task decomposition), and interaction mechanisms (clicks, text entries, and other UI actions) (Wright, 2024; Zhou et al., 2023). For perception, some rely on accessibility APIs, while others parse HTML/DOM structures or process raw screenshots. Hybrid approaches combine these methods for more reliable understanding (Wu et al., 2024c; Nong et al., 2024; Yang et al., 2024; Deng et al., 2023). For interaction, agents perform tasks through replicating human-like interactions, such as clicking or typing (Koh et al., 2024a). Beyond the above perception and interaction, effective task decomposition is a core capability for GUI agents to navigate complex workflows and adapt to dynamic interfaces. A structured planning mechanism allows agents to decompose multi-step tasks and execute actions reliably across diverse environments (Gu et al., 2024a; Zhu et al., 2025; Koh et al., 2024b).

标准代理流程包括多模态感知、推理与规划 (任务分解) 及交互机制 (点击、文本输入及其他界面操作)(Wright，2024；Zhou 等，2023)。在感知方面，有些依赖辅助功能 API，有些解析 HTML/DOM 结构或处理原始截图。混合方法结合这些手段以实现更可靠的理解 (Wu 等，2024c；Nong 等，2024；Yang 等，2024；Deng 等，2023)。在交互方面，代理通过模拟人类交互如点击或输入执行任务 (Koh 等，2024a)。除上述感知与交互外，有效的任务分解是 GUI 代理导航复杂工作流和适应动态界面的核心能力。结构化规划机制使代理能够分解多步骤任务，并在多样环境中可靠执行动作 (Gu 等，2024a；Zhu 等，2025；Koh 等，2024b)。

GUI agents can serve diverse applications. In mobile settings, they automate navigation and data entry via hierarchical planning (Nong et al., 2024; Zhu et al., 2025). On the web, they can support tasks such as automated testing, phishing detection, and e-commerce applications (Cao et al., 2024; Wang et al., 2024b; Gu et al., 2024b). In specialized domains like healthcare and education, they assist multimodal reasoning while ensuring privacy and accessibility (Cuadra et al., 2024; Arnold and Tilton, 2024; Srinivas et al., 2024).

GUI 代理可服务于多种应用。在移动场景中，它们通过分层规划实现导航和数据输入自动化 (Nong 等，2024；Zhu 等，2025)。在网络环境中，可支持自动化测试、钓鱼检测和电子商务等任务 (Cao 等，2024；Wang 等，2024b；Gu 等，2024b)。在医疗和教育等专业领域，代理辅助多模态推理，同时保障隐私和无障碍性 (Cuadra 等，2024；Arnold 和 Tilton，2024；Srinivas 等，2024)。

Despite recent advances, key challenges remain. Agents are still vulnerable to adversarial multimodal inputs, which can trigger unpredictable behavior (Gao et al., 2024; Ma et al., 2024). They also struggle to generalize to unfamiliar interfaces, highlighting the need for stronger robustness (Kim et al., 2024b). Additionally, balancing real-time performance with safety remains an ongoing challenge, necessitating more efficient architectures (Shen et al., 2024; Nguyen et al., 2024b).

尽管近期取得进展，关键挑战依然存在。代理仍易受多模态对抗输入影响，可能引发不可预测行为 (Gao 等，2024；Ma 等，2024)。它们在面对陌生界面时泛化能力不足，凸显了增强鲁棒性的需求 (Kim 等，2024b)。此外，实时性能与安全性的平衡仍是持续挑战，需更高效的架构设计 (Shen 等，2024；Nguyen 等，2024b)。

As GUI agents are increasingly deployed in real-world scenarios, addressing challenges related to security,

reliability, explainability, and ethical alignment becomes essential. The following section examines these dimensions in depth, and we include a brief overview of key dimensions for building trustworthy GUI agents in Figure 2.

随着 GUI 代理在现实场景中的广泛部署，解决安全、可靠性、可解释性和伦理对齐相关挑战变得至关重要。下一节将深入探讨这些维度，并在图 2 中简要概述构建可信 GUI 代理的关键维度。

3 Security and Privacy

3 安全与隐私

Security and privacy are central concerns for GUI agents. This section first outlines significant attacks and vulnerabilities that arise when agents interact with graphical interfaces. It then addresses risks surrounding user data, and finally discusses defense strategies and open problems. Table 1 summarizes key threats and defenses.

安全与隐私是 GUI 代理的核心关注点。本节首先概述代理与图形界面交互时出现的主要攻击和漏洞，随后讨论用户数据相关风险，最后探讨防御策略及未解决的问题。表 1 总结了主要威胁与防御措施。

3.1 Attacks and Vulnerabilities

3.1 攻击与漏洞

The interactive nature of GUI agents gives rise to novel exploits, including adversarial image perturbations and malicious prompt injections embedded in webpages (Janowczyk et al., 2024). Attacks such as Imprompter (Fu et al., 2024) manipulate a single product image to hijack an agent's actions with high success rates. Browser-based jailbreak-ing also emerges, whereby refusal-trained LLMs inadvertently execute harmful behaviors in non-chat contexts (Kumar et al., 2024). Some studies reveal that malicious instructions can be hidden in website structures, where compromised webpages consistently mislead the agent (Wu et al., 2024a). AdvWeb demonstrates how black-box adversarial prompts injected into web pages can mislead web agents into executing unintended actions while remaining undetectable to users (Xu et al., 2024a).

图形用户界面 (GUI) 代理的交互特性催生了新型攻击手段，包括嵌入网页中的对抗性图像扰动和恶意提示注入 (Janowczyk 等，2024)。如 Imprompter(Fu 等，2024) 等攻击通过操控单一产品图像，高效劫持代理的行为。基于浏览器的越狱攻击也随之出现，拒绝训练的大型语言模型 (LLM) 在非对话场景中意外执行有害行为 (Kumar 等，2024)。部分研究揭示恶意指令可隐藏于网站结构中，受损网页持续误导代理 (Wu 等，2024a)。AdvWeb 展示了如何将黑箱对抗性提示注入网页，误导网页代理执行非预期操作且用户难以察觉 (Xu 等，2024a)。

On mobile devices, multiple attack paths target both the perception and reasoning modules of multimodal agents, emphasizing the breadth of vulnerabilities (Yang et al., 2024). Similarly, AEIA-MN shows that mobile GUI agents are highly vulnerable to environmental injection attacks, where malicious elements disguised as system features disrupt decision-making with up to 93% success rates (Chen et al., 2025).

在移动设备上，多条攻击路径针对多模态代理的感知和推理模块，凸显了漏洞的广泛性 (Yang 等，2024)。同样，AEIA-MN 表明移动 GUI 代理极易受到环境注入攻击，恶意元素伪装成系统功能，干扰决策，成功率高达 93% (Chen 等，2025)。

Altogether, these exploits demonstrate that seemingly benign elements, such as small visual perturbations or hidden HTML code, can manipulate complex agent pipelines. Because GUI agents operate across modalities and maintain hidden internal state, such threats can slip through and spread across components over time (Wu et al., 2025a). Ensuring security thus requires a holistic view of the GUI agent's entire workflow.

综上所述，这些攻击表明看似无害的元素，如微小的视觉扰动或隐藏的 HTML 代码，能够操控复杂的代理流程。由于 GUI 代理跨模态运行且保持隐藏的内部状态，此类威胁可能悄然渗透并随时间在组件间传播 (Wu 等，2025a)。因此，保障安全需对 GUI 代理的整体工作流程进行全面审视。

3.2 Privacy Risks

3.2 隐私风险

Beyond security exploits, GUI agents raise pressing privacy concerns by accessing sensitive personal or enterprise data through visual and textual interfaces (Chen et al., 2024a; Zhang et al., 2024b). Screenshot-based perception can be particularly sensitive, potentially exposing private details without user awareness. Furthermore, when agents run in the cloud or on remote servers, the risk of data leakage increases (Gan et al., 2024; Xu et al., 2024b). These risks become especially critical when GUI agents interact with regulated domains such as finance and healthcare, where the exposure of Personally Identifiable Information (PII), including names, addresses, and financial details, can have severe consequences.

除了安全攻击，GUI 代理还引发了紧迫的隐私问题，因为它们通过视觉和文本接口访问敏感的个人或企业数据 (Chen 等，2024a；Zhang 等，2024b)。基于截图的感知尤为敏感，可能在用户不知情的情况下暴露私人信息。此外，当代理运行于云端或远程服务器时，数据泄露风险加剧 (Gan 等，2024；Xu 等，2024b)。当 GUI 代理涉及金融和医疗等受监管领域时，个人身份信息 (PII) 如姓名、地址和财务细节的泄露可能带来严重后果。

Recent studies highlight emerging threats, such as Environmental Injection Attacks (EIA), which covertly manipulate web environments to extract PII from GUI agents with up to 70% success rates (Liao et al., 2024). Similarly, web-enabled LLM agents have been found to enhance cyber-attacks, such as automated PII harvesting, impersonation post generation, and spear-phishing (Kim et al., 2024a), which reveals critical gaps in existing security measures. Beyond direct attacks, users may also inadvertently disclose sensitive information during normal interactions with GUI agents, underscoring the need for contextual privacy protections (Ngong et al., 2025).

近期研究强调了新兴威胁，如环境注入攻击 (EIA)，该攻击隐秘操控网络环境，从 GUI 代理中提取个人身份信息，成功率高达 70%(Liao 等，2024)。同样，支持网页的 LLM 代理被发现助长了网络攻击，如自动化个人信息收集、生成冒充内容和鱼叉式钓鱼 (Kim 等，2024a)，暴露了现有安全措施的关键漏洞。除直接攻击外，用户在与 GUI 代理的正常交互中也可能无意泄露敏感信息，凸显了情境隐私保护的必要性 (Ngong 等，2025)。

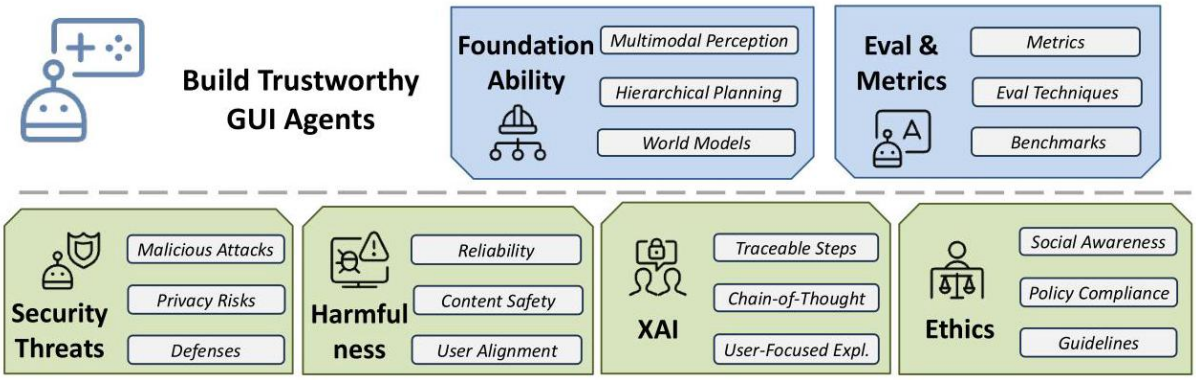

Figure 2: Overview of key dimensions for building trustworthy GUI agents, highlighting foundational abilities, evaluation metrics, security threats, reliability, harmfulness, explainability, transparency, and ethical implications.

图 2: 构建可信 GUI 代理的关键维度概览，突出基础能力、评估指标、安全威胁、可靠性、有害性、可解释性、透明度及伦理影响。

3.3 Defenses and Mitigation Strategies

3.3 防御与缓解策略

Researchers have proposed multiple approaches to mitigate security and privacy risks in GUI agents. Input validation and prompt injection detection aim to filter out unsafe content before it reaches the core model, as demonstrated by Sharma et al. (2024). Other work introduces specialized guardrail agents that intercept and inspect commands generated by primary agents, blocking disallowed actions (Xi-ang et al., 2024). Some frameworks leverage adversarial training or visual analytics systems, like AdversaFlow (Deng et al., 2024), to identify vulnerabilities collaboratively. AutoDroid enhances mobile GUI automation by integrating language models with dynamic UI analysis, enabling scalable, hands-free task execution across arbitrary Android apps without manual effort (Wen et al., 2024). Similarly, G-Safeguard applies graph-based anomaly detection to multi-agent systems, mitigating prompt injection attacks and securing agent collaboration (Wang et al., 2025).

研究者提出多种方法以缓解 GUI 代理的安全与隐私风险。输入验证和提示注入检测旨在过滤不安全内容，防止其进入核心模型，正如 Sharma 等 (2024) 所示。其他工作引入专门的护栏代理，拦截并检查主代理生成的命令，阻止不允许的操作 (Xi-ang 等，2024)。部分框架利用对抗训练或视觉分析系统，如 AdversaFlow(Deng 等，2024)，协同识别漏洞。AutoDroid 通过将语言模型与动态 UI 分析结合，提升移动 GUI 自动化，实现跨任意安卓应用的可扩展免手动任务执行 (Wen 等，2024)。类似地，G-Safeguard 采用基于图的异常检测保护多代理系统，缓解提示注入攻击，保障代理协作安全 (Wang 等，2025)。

To address privacy, solutions like CLEAR (Chen et al., 2024a) analyze user-provided data and privacy policies to highlight potential leakages. Others advocate secure sandboxing, local processing, and advanced authentication to constrain agent permissions (Zhang et al., 2024b; Gu et al., 2024a). PAPILLON proposes a privacy-conscious delegation framework that selectively routes queries between local and proprietary LLMs to minimize sensitive data exposure while maintaining high response quality (Siyan et al., 2024).

针对隐私问题，诸如 CLEAR(Chen 等，2024a) 等方案分析用户提供的数据和隐私政策，突出潜在泄露点。另有研究倡导安全沙箱、本地处理及高级认证以限制代理权限 (Zhang 等，2024b；Gu 等，2024a)。PAPILLON 提出隐私意识委托框架，有选择地在本地与专有大型语言模型间路由查询，最大限度减少敏感数据暴露，同时保持高质量响应 (Siyan 等，2024)。

Recent commercial implementations of GUI agent frameworks also provide valuable insights into multi-layered defense strategies. For example, OpenAI's CUA employs a comprehensive defense-in-depth approach. This includes preventative measures such as website blocklists and refusal training, interactive safeguards like user confirmations for critical actions, and detection systems for real-time moderation and monitoring of suspicious content (OpenAI, 2025). This strategy recognizes that perfect prevention is unattainable and instead focuses on using complementary systems to gradually reduce risk. On the other hand, Anthropic advises limiting computer use to secure environments, such as virtual machines with minimal privileges, to mitigate ongoing vulnerabilities to jailbreaking and prompt injection (Anthropic, 2025).

最近商业化的 GUI 代理框架实现也为多层防御策略提供了宝贵的见解。例如，OpenAI 的 CUA 采用了全面的纵深防御方法。这包括预防措施，如网站黑名单和拒绝训练，交互式保护措施，如关键操作的用户确认，以及用于实时审核和监控可疑内容的检测系统 (OpenAI，2025)。该策略认识到完美的预防是不可能的，而是通过互补系统逐步降低风险。另一方面，Anthropic 建议将计算机使用限制在安全环境中，如具有最小权限的虚拟机，以减轻持续存在的越狱和提示注入漏洞 (Anthropic，2025)。

3.4 Future Directions

3.4 未来方向

Securing GUI agents requires balanced solutions that protect users while maintaining usability. Three promising directions deserve exploration:

保障 GUI 代理的安全需要在保护用户和保持可用性之间取得平衡。有三个值得探索的有前景的方向:

Smarter Defense Tools: We need lightweight, real-time mechanisms to detect hidden attacks like those demonstrated in AdvWeb (Xu et al., 2024a) and Imprompter (Fu et al., 2024). Browser extensions could sanitize webpage elements before agents process them, while mobile applications might verify UI elements against device sensors to detect overlay attacks similar to those in AEIA-MN (Chen et al., 2025). Simple visual filters could automatically blur sensitive information on screens before agents capture screenshots, preventing data leakage without complex infrastructure changes, addressing concerns raised by Chen et al. (2024a). In parallel, recent advances such as ZIP (Shi et al., 2023a) demonstrate how zero-shot image purification techniques can effectively defend against visual backdoor attacks, offering a direction for mitigating image-based threats in GUI agents.

更智能的防御工具: 我们需要轻量级、实时的机制来检测隐藏攻击，如 AdvWeb(Xu 等，2024a) 和 Imprompter(Fu 等，2024) 中展示的攻击。浏览器扩展可以在代理处理网页元素之前对其进行净化，而移动应用则可能通过设备传感器验证 UI 元素，以检测类似 AEIA-MN(Chen 等，2025) 中的覆盖攻击。简单的视觉滤镜可以在代理截屏前自动模糊屏幕上的敏感信息，防止数据泄露，无需复杂的基础设施改动，回应了 Chen 等 (2024a) 提出的担忧。与此同时，诸如 ZIP(Shi 等，2023a) 等最新进展展示了零样本图像净化技术如何有效防御视觉后门攻击，为缓解 GUI 代理中的基于图像的威胁提供了方向。

Approach/Attack	Key Characteristics
Malicious Attack & Vulnerabilities	
Imprompter (Fu et al., 2024)	Hijacks agent actions through modified product images
Browser-based jailbreaking (Kumar et al., 2024)	Induces harmful behaviors in non-chat environments
WebPI (Wu et al., 2024a)	Embeds malicious commands in webpage structures
AdvWeb (Xu et al., 2024a)	Injects adversarial prompts with high user undetectability
AEIA-MN (Chen et al., 2025)	Disguises malicious elements as system features (93% success)
ARE (Wu et al., 2025a)	Propagates threats across agent module boundaries
Privacy Risk	
Screenshot leakage (Chen et al., 2024a)	Captures sensitive information in interface snapshots
EIA (Liao et al., 2024)	Extracts PII with up to 70% success rate
Agent-enabled cyberattacks (Kim et al., 2024a)	Facilitates PII harvesting and spear-phishing
Cloud-based processing (Gan et al., 2024)	Amplifies exposure risk in distributed architectures
Protecting Users (Ngong et al., 2025)	Reveals sensitive information during normal operations
Defense & Mitigation Approaches	
GuardAgent (Xiang et al., 2024)	Intercepts and blocks disallowed agent actions
AdversaFlow (Deng et al., 2024)	Identifies vulnerabilities through collaborative analysis
AutoDroid (Wen et al., 2024)	Automates tasks by combining LLMs with dynamic UI analysis
CLEAR (Chen et al., 2024a)	Analyzes data against privacy policies
PAPILLON (Siyan et al., 2024)	Minimizes exposure while maintaining response quality
Input validation (Sharma et al., 2024)	Screens unsafe content before model processing

方法/攻击	关键特征
恶意攻击与漏洞	
Imprompter(傅等，2024)	通过修改的产品图片劫持代理行为
基于浏览器的越狱 (库马尔等，2024)	在非聊天环境中诱导有害行为
WebPI(吴等，2024a)	在网页结构中嵌入恶意命令
AdvWeb(徐等，2024a)	注入对用户高度隐蔽的对抗性提示
AEIA-MN(陈等，2025)	将恶意元素伪装为系统功能 (成功率 93%)
ARE(吴等，2025a)	跨代理模块边界传播威胁
隐私风险	
截图泄露 (陈等，2024a)	捕获界面快照中的敏感信息
EIA(廖等，2024)	提取个人身份信息 (PII)，成功率高达 70%
基于代理的网络攻击 (金等，2024a)	促进个人身份信息收集和定向钓鱼
基于云的处理 (甘等，2024)	在分布式架构中放大暴露风险
保护用户 (农等，2025)	在正常操作中泄露敏感信息
防御与缓解方法	
GuardAgent(向等，2024)	拦截并阻止不允许的代理行为
AdversaFlow(邓等，2024)	通过协同分析识别漏洞
AutoDroid(温等，2024)	结合大型语言模型 (LLMs) 与动态界面分析自动化任务
CLEAR(陈等，2024a)	根据隐私政策分析数据
PAPILLON(司燕等，2024)	在保持响应质量的同时最小化暴露
输入验证 (沙玛等，2024)	在模型处理前筛查不安全内容

Table 1: Taxonomy of Security and Privacy Considerations for GUI Agents. The table presents three key dimensions: (1) attacks and vulnerabilities exploiting multimodal interfaces, (2) privacy risk mechanisms that can expose sensitive information, and (3) defense and mitigation approaches aimed at protecting agent operations and user data.

表 1:GUI 代理的安全与隐私考量分类。表中展示了三个关键维度:(1) 利用多模态界面的攻击与漏洞,(2) 可能暴露敏感信息的隐私风险机制，以及 (3) 旨在保护代理操作和用户数据的防御与缓解方法。

User-Controlled Privacy: Drawing inspiration from mobile permissions, future GUI agents should request specific, time-limited access to data (e.g., ”view this webpage for 5 minutes”). Combined with local models that automatically redact personal information and clear activity logs, this approach would give users meaningful control while preserving convenience, as suggested by Zhang et al. (2024b). Users should understand what their agent sees and how their data is used.

用户控制的隐私: 借鉴移动权限的理念，未来的 GUI 代理应请求特定且时限明确的数据访问权限 (例如，“查看此网页 5 分钟”)。结合自动脱敏个人信息和清除活动日志的本地模型，该方法在保障便利性的同时赋予用户实质控制权，正如 Zhang 等人 (2024b) 所建议。用户应清楚代理所见内容及其数据使用方式。

Connected Defense Layers: Since GUI agents process multiple types of information (images, text, system states), defenses should verify consistency across channels (Wu et al., 2025a). For instance, an agent should cross-check text from a button image with the underlying HTML to catch any possible tampering. In high-stakes scenarios like payments or accessing healthcare data, agents should leverage a hardware-based security checking approach

and require explicit user confirmation for any suspicious or sensitive actions, following ideas similar to those in Xiang et al. (2024).

联动防御层: 由于 GUI 代理处理多种信息类型 (图像、文本、系统状态)，防御措施应验证各通道间的一致性 (Wu 等，2025a)。例如，代理应将按钮图像中的文本与底层 HTML 交叉核对，以发现潜在篡改。在支付或访问医疗数据等高风险场景中，代理应采用基于硬件的安全检测方法，并对任何可疑或敏感操作要求用户明确确认，类似于 Xiang 等人 (2024) 的思路。

4 Reliability and Harmlessness

4 可靠性与无害性

This section examines how GUI agents handle visual hallucination, inappropriate content, and alignment with human values. They are core challenges for building both robust and safe GUI agents.

本节探讨 GUI 代理如何应对视觉幻觉、不当内容及与人类价值观的对齐问题。这些是构建稳健且安全 GUI 代理的核心挑战。

4.1 Reliability

4.1 可靠性

Ensuring stable and accurate interaction with visual interfaces is essential for GUI agents, especially in tackling the challenge of hallucination, where agents generate actions or interpretations that do not match the visual content (Bai et al., 2024; Chen et al., 2024d). Such errors can include fabricated UI elements, incorrect readings of interface components, or lapses in visual focus, all of which can lead to unreliable behavior (Liu et al., 2023a; Jiang et al., 2024a; Yu et al., 2024). Furthermore, recent work (Ma et al., 2024) also reveals that even in benign, non-malicious environments, multimodal GUI agents are vulnerable to environmental distractions that undermine their reliability.

确保与视觉界面的稳定且准确交互对 GUI 代理至关重要，尤其是在应对幻觉问题时，即代理生成与视觉内容不符的动作或解读 (Bai 等，2024；Chen 等，2024d)。此类错误包括虚构的界面元素、界面组件的错误读取或视觉焦点的偏差，均可能导致行为不可靠 (Liu 等，2023a；Jiang 等，2024a；Yu 等，2024)。此外，最新研究 (Ma 等，2024) 表明，即使在无恶意的环境中，多模态 GUI 代理也易受环境干扰，影响其可靠性。

To mitigate these issues, several strategies have been developed. Opera introduces an over-trust penalty to prevent reliance on misleading summary tokens (Huang et al., 2024a), while Volcano employs self-feedback for natural language correction (Lee et al., 2023). Contrastive learning techniques help distinguish between hallucinative and non-hallucinative text, enhancing model robustness (Jiang et al., 2024a).

为缓解这些问题，已开发多种策略。Opera 引入过度信任惩罚以防止依赖误导性摘要标记 (Huang 等，2024a)，Volcano 则采用自我反馈进行自然语言纠正 (Lee 等，2023)。对比学习技术有助区分幻觉文本与非幻觉文本，提升模型鲁棒性 (Jiang 等，2024a)。

Real-time detection frameworks like UNIHD validate outputs against visual evidence, and methods such as Residual Visual Decoding address ”hallucination snowballing” by revising outputs with residual visual input (Chen et al., 2024d; Zhong et al., 2024). Specialized datasets like LRV-Instruction and M-HalDetect further aid in reducing hallucination rates by providing targeted training resources (Liu et al., 2023a; Gunjal et al., 2023). Multi-agent systems also offer promising solutions by combining self-correction, external feedback, and agent debate to maintain accurate grounding in complex interactions (Yu et al., 2024).

实时检测框架如 UNIHD 通过视觉证据验证输出，残差视觉解码方法则通过残差视觉输入修正输出，解决“幻觉雪球效应”(Chen 等，2024d；Zhong 等，2024)。专门数据集如 LRV-Instruction 和 M-HalDetect 通过提供针对性训练资源进一步降低幻觉率 (Liu 等，2023a；Gunjal 等，2023)。多代理系统结合自我纠正、外部反馈和代理辩论，为复杂交互中保持准确基础提供了有前景的解决方案 (Yu 等，2024)。

4.2 Content Safety

4.2 内容安全

Because GUI agents can generate and display multimodal content, ensuring safe and appropriate outputs is critical (Gao et al., 2024; Liu et al., 2023b). Image-based manipulations may prompt harmful or toxic responses, undermining the base LLM’s alignment. Tailored calibration approaches, such as CoCA (Gao et al., 2024), attempt to restore the model’s original safety guardrails under multimodal contexts. Recent work by Zou et al. (2024) introduces ”circuit breakers” that can interrupt GUI agents when generating harmful outputs, functioning effectively even against sophisticated adversarial attacks in multimodal settings. Similarly, frameworks like RapGuard (Jiang et al., 2024b) dynamically generate scenario-specific safety prompts to reduce risks in each interaction.

由于 GUI 代理可生成并展示多模态内容，确保输出安全且适当至关重要 (Gao 等，2024；Liu 等，2023b)。基于图像的操控可能引发有害或有毒反应，破坏基础大语言模型 (LLM) 的对齐。定制校准方法如 CoCA(Gao 等，2024) 试图在多模态环境下恢复模型原有的安全防护。Zou 等人 (2024) 提出的“断路器”机制能在生成有害输出时中断 GUI 代理，即使面对复杂的多模态对抗攻击也能有效发挥作用。同样，RapGuard 框架 (Jiang 等，2024b) 动态生成场景特定的安全提示，降低每次交互的风险。

Recent work on self-defense mechanisms offers promising strategies to protect GUI agents from manipulation. Phute et al. (2023) demonstrated that agents can effectively filter their own responses to block harmful content generation while not sacrificing functionality. Moving on, Xie et al. (2023) developed a complementary ”system-mode self-reminder” technique, which wraps user queries in prompts that reinforce safe behavior. This method reduced jailbreak attack success rates from 67% to 19%. For agents handling sensitive tasks, Greenblatt et al. (2023) proposed robust safety protocols like ”trusted editing” and ”untrusted monitoring” that remain effective even when the agent actively tries to bypass them. At the model level, Liu et al. (2024) introduced Selective Knowledge Unlearning (SKU) to remove harmful knowledge from the underlying models powering GUI agents while preserving performance on legitimate tasks.

近期关于自我防御机制的研究为保护 GUI 代理免受操控提供了有前景的策略。Phute 等人 (2023) 证明，代理能够有效过滤自身响应，阻止有害内容生成，同时不牺牲功能性。随后，Xie 等人 (2023) 开发了一种互补的“系统模式自我提醒”技术，通过在用户查询中嵌入强化安全行为的提示语。这种方法将越狱攻击成功率从 67% 降低到 19%。对于处理敏感任务的代理，Greenblatt 等人 (2023) 提出了如“可信编辑”和“不可信监控”等强健的安全协议，即使代理主动尝试绕过也能保持有效。在模型层面，Liu 等人 (2024) 引入了选择性知识遗忘 (Selective Knowledge Unlearning，SKU)，以从驱动 GUI 代理的底层模型中移除有害知识，同时保持对合法任务的性能。

In practice, content safety depends on the agent's ability to reject unsafe requests, avoid exposing sensitive information, and handle ambiguous inputs responsibly. To catch potential harms before deployment, frameworks like AHA! (Anticipating Harms of AI)(Buçinca et al., 2023) support developers in identifying how different AI behaviors might negatively impact various stakeholders. AHA! creates example scenarios that show different ways agent systems can go wrong, based on responses from both crowd workers and language models. Interactive benchmarks such as ST-WebAgentBench highlight how easily agent alignment can fail when faced with real-world websites (Levy et al., 2024). Overall, adding stronger content filtering throughout the pipeline, along with consistent logging of agent actions, can help limit the impact when alignment breaks down.

在实际应用中，内容安全依赖于代理拒绝不安全请求、避免泄露敏感信息以及负责任地处理模糊输入的能力。为了在部署前捕捉潜在危害，诸如 AHA!(Anticipating Harms of AI)(Buçinca 等人，2023) 等框架支持开发者识别不同 AI 行为可能对各利益相关者产生的负面影响。AHA! 基于众包工作者和语言模型的响应，创建展示代理系统可能出错的示例场景。交互式基准测试如 ST-WebAgentBench 突出显示了代理对真实网站时对齐失败的易发性 (Levy 等人，2024)。总体而言，在整个流程中加强内容过滤，并持续记录代理行为，有助于在对齐失败时限制其影响。

4.3 Alignment with Human Values

4.3 与人类价值观的对齐

Aligning GUI agents with human values means weighing individual user goals, like efficiency and personalization, against broader concerns such as fairness and inclusivity. One approach is to define these principles up front, as in Hua et al. (2024), where agent constitutions are used to embed safety guidelines during the planning process. Other frameworks, such as ResponsibleTA (Zhang et al., 2023), structure collaboration among multiple agent components to verify each step's feasibility and security. FREYR introduces a modular approach to tool integration in LLMs, improving adaptability to user needs without requiring extensive model fine-tuning (Gallotta et al., 2025).

使 GUI 代理与人类价值观对齐意味着在考虑个体用户目标 (如效率和个性化) 与更广泛关注点 (如公平性和包容性) 之间权衡。一种方法是预先定义这些原则，如 Hua 等人 (2024) 所示，通过代理宪章在规划过程中嵌入安全指导方针。其他框架如 ResponsibleTA(Zhang 等人，2023) 构建多代理组件间的协作结构，以验证每一步的可行性和安全性。FREYR 引入了 LLM 中工具集成的模块化方法，提高了对用户需求的适应性，而无需大量模型微调 (Gallotta 等人，2025)。

Moral reasoning and cultural sensitivity are increasingly relevant, particularly when agents operate in diverse contexts or handle sensitive content (Qiu et al., 2024; Piatti et al., 2024). To achieve deeper alignment, agents

may need to model nuanced social norms or policy constraints. Visual-Critic demonstrates how LMMs can assess visual content quality from a human perspective, a critical capability for ensuring user-aligned perception in GUI interactions (Huang et al., 2024b).

道德推理和文化敏感性日益重要，尤其当代理在多样化环境中运行或处理敏感内容时 (Qiu 等人，2024；Piatti 等人，2024)。为了实现更深层次的对齐，代理可能需要建模细致的社会规范或政策约束。Visual-Critic 展示了大型多模态模型 (LMM) 如何从人类视角评估视觉内容质量，这对于确保 GUI 交互中用户对感知的对齐至关重要 (Huang 等人，2024b)。

Commercial GUI agent implementations provide further insights into practical alignment strategies. OpenAI's CUA, for example, implements "watch mode" on sensitive websites (e.g., email) to ensure that critical operations are supervised by users, and it declines higher-risk tasks, such as banking transactions or sensitive decision-making, thereby enforcing clear capability boundaries as a safety mechanism (OpenAI, 2025). Similarly, Anthropic restricts its computer use beta feature from creating accounts or generating content on social platforms to prevent human impersonation (Anthropic, 2025).

商业 GUI 代理的实现为实际对齐策略提供了更多见解。例如，OpenAI 的 CUA 在敏感网站 (如电子邮件) 上实施“观察模式”，确保关键操作由用户监督，并拒绝高风险任务，如银行交易或敏感决策，从而作为安全机制强制执行明确的能力边界 (OpenAI，2025)。类似地，Anthropic 限制其计算机使用测试版功能创建账户或在社交平台生成内容，以防止冒充他人 (Anthropic，2025)。

Another critical challenge is user intent understanding. As noted by Kim et al. (2024b), GUI agents still struggle to accurately infer user goals across diverse applications, achieving only poor accuracy on unseen websites. Designing models that generalize effectively across varying tasks is crucial, particularly for handling contextual variations in user interactions and predicting user behavior in complex interfaces (Stefanidi et al., 2022; Gao et al., 2023). Recent research on Role-Playing Language Agents (RPLAs) highlights how LLMs can simulate personas and dynamically adapt to user preferences, offering a pathway to more personalized and context-aware GUI interactions (Chen et al., 2024b).

另一个关键挑战是理解用户意图。正如 Kim 等人 (2024b) 指出，GUI 代理在不同应用中仍难以准确推断用户目标，在未见过的网站上准确率较低。设计能够跨多样任务有效泛化的模型至关重要，特别是在处理用户交互的上下文变化和预测复杂界面中的用户行为方面 (Stefanidi 等人，2022；Gao 等人，2023)。近期关于角色扮演语言代理 (Role-Playing Language Agents，RPLAs) 的研究强调了大型语言模型 (LLM) 如何模拟角色并动态适应用户偏好，为更个性化和上下文感知的 GUI 交互提供了路径 (Chen 等人，2024b)。

4.4 Future Directions

4.4 未来方向

To enhance GUI agent reliability and safety while maintaining practical implementations, research should focus on these promising directions:

为了提升 GUI 代理的可靠性和安全性，同时保持实际应用，研究应聚焦于以下有前景的方向:

Real-Time Hallucination Prevention: Building on work by Chen et al. (2024d) and Zhong et al. (2024), future systems need lightweight verification mechanisms that catch inconsistencies before they cause errors. Browser extensions could cross-verify agent actions against actual webpage structures, flagging discrepancies immediately. Interactive correction interfaces would allow users to adjust an agent's visual attention during errors, creating valuable feedback loops to improve perception models. Additionally, environmental awareness systems could detect real-world distractions that might compromise reliability, addressing concerns raised by Ma et al. (2024).

实时幻觉预防: 基于 Chen 等人 (2024d) 和 Zhong 等人 (2024) 的工作，未来系统需要轻量级验证机制，在错误发生前捕捉不一致。浏览器扩展可以将代理行为与实际网页结构交叉验证，立即标记差异。交互式纠正界面允许用户在错误时调整代理的视觉关注点，形成宝贵的反馈循环以改进感知模型。此外，环境感知系统能够检测可能影响可靠性的现实干扰，回应了 Ma 等人 (2024) 提出的关注点。

Adaptive Safety Architecture: Rather than applying uniform safety measures, agents should dynamically adjust protection levels based on context. When financial or medical interfaces are detected, content filters could automatically tighten, similar to the "watch mode" implemented in commercial systems (OpenAI, 2025). Modular safety components, like specialized verifiers for payment dialogs, could be plugged in as needed, extending the "circuit breaker" concept introduced by Zou et al. (2024). For critical operations, requiring physical confirmation (e.g., a device authentication) could provide an additional security layer inspired by multi-agent verification (Yu et al., 2024).

自适应安全架构: 代理应根据上下文动态调整保护级别，而非采用统一的安全措施。当检测到金融或医疗界面时，内容过滤器可以自动收紧，类似于商业系统中实施的“观察模式”(OpenAI, 2025)。模块化安全组件，如针对支付对话的专用验证器，可按需插入，扩展了 Zou 等人 (2024) 提出的“断路器”概念。对于关键操作，要求物理确认 (例如设备认证) 可提供额外的安全层，灵感来自多代理验证 (Yu 等人, 2024)。

Learning from Failures: Perhaps most promising is the systematic improvement of agents through failure analysis. Community-driven reporting of rare errors could create diverse testing datasets beyond what developers anticipate. Automated post-failure analysis reports would help identify perception or reasoning gaps, extending approaches like those in Gunjal et al. (2023). By prioritizing fixes based on real-world impact rather than theoretical concerns, development resources could target the most critical reliability issues first.

从失败中学习: 最有前景的或许是通过失败分析系统性地改进代理。社区驱动的罕见错误报告可创建超出开发者预期的多样化测试数据集。自动化的失败后分析报告有助于识别感知或推理缺陷，扩展了 Gunjal 等人 (2023) 的方法。通过优先解决基于现实影响的问题而非理论顾虑，开发资源可优先针对最关键的可靠性问题。

5 Explainability and Transparency

5 可解释性与透明度

Explainability and transparency foster trust in GUI agents by helping users understand how the system perceives, reasons, and acts. This section discusses mechanisms for providing explanations, transparent decision-making, and user-centric presentation.

可解释性与透明度通过帮助用户理解系统如何感知、推理和行动，促进对 GUI 代理的信任。本节讨论提供解释、透明决策和以用户为中心的展示机制。

5.1 Techniques for Explaining Agent Behavior

5.1 解释代理行为的技术

Many methods focus on decomposing the agent's decision pipeline to surface intermediate steps. Explainable Behavior Cloning (EBC-LLMAgent) (Guan et al., 2024) captures demonstrations, generates executable code, and maps the code to UI elements. By documenting these transformations, the agent can clarify how it arrived at a particular action. Similarly, hierarchical designs that separate high-level planning from low-level execution offer more interpretable structures (Liu et al., 2025a; Zhu et al., 2025; Agashe et al., 2024).

许多方法侧重于分解代理的决策流程以展示中间步骤。可解释行为克隆 (Explainable Behavior Cloning, EBC-LLMAgent)(Guan 等人, 2024) 捕捉示范，生成可执行代码，并将代码映射到 UI 元素。通过记录这些转换，代理能阐明其采取特定行动的过程。同样，将高层规划与低层执行分离的分层设计提供了更易解释的结构 (Liu 等人, 2025a; Zhu 等人, 2025; Agashe 等人, 2024)。

Other efforts highlight introspection through multi-agent or chain-of-thought strategies (Nguyen et al., 2024c; Wang and Liu, 2024). Here, LLM-based agents iteratively reflect on previous reasoning steps, generating self-explanations or corrections. These reflective traces not only boost performance but also produce human-readable rationales. However, ensuring that explanations remain truthful rather than post-hoc justifications is still an open research challenge.

其他工作强调通过多代理或链式思维策略进行内省 (Nguyen 等人, 2024c; Wang 和 Liu, 2024)。此处，基于大型语言模型的代理迭代反思先前推理步骤，生成自我解释或修正。这些反思轨迹不仅提升性能，还产生人类可读的理由。然而，确保解释是真实的而非事后合理化仍是一个开放的研究挑战。

5.2 Transparency in Decision-Making

5.2 决策透明度

Transparency is especially crucial for high-stakes domains like finance or healthcare, where agents may access sensitive data or perform costly actions. Systems such as XMODE (Nooralahzadeh et al., 2024) rely on multimodal decomposition, combining textual and visual analytics to highlight evidence supporting each decision. Providing users with comprehensible summaries, such as color-coded or textual rationales, can help users trace the logic of agents (Houssel et al., 2024; Arnold and Tilton, 2024).

透明度对金融或医疗等高风险领域尤为重要，这些领域代理可能访问敏感数据或执行高成本操作。系统如 XMODE(Nooralahzadeh 等人, 2024) 依赖多模态分解，结合文本和视觉分析突出支持每个决策的证据。向用户提供易懂的摘要，如颜色编码或文本理由，有助于用户追踪代理的逻辑 (Houssel 等人, 2024; Arnold 和 Tilton, 2024)。

Equally vital is the agent's ability to justify or revise actions when confronted with unexpected outcomes. World models can enhance transparency by simulating multiple paths and explaining why certain actions seem preferable (Chae et al., 2024; Gu et al., 2024a). While this can help build user trust, it also comes with added computational cost; therefore, designs need to strike a balance to keep interactions responsive.

同样重要的是代理在面对意外结果时，能够为行动辩护或修正。世界模型通过模拟多条路径并解释为何某些行动更优，增强透明度 (Chae 等人, 2024; Gu 等人, 2024a)。虽然这有助于建立用户信任，但也增加了计算成本；因此设计需权衡以保持交互响应性。

5.3 User-Centric Explanations

5.3 以用户为中心的解释

User-centered design focuses on tailoring explanations based on a person's context, preferences, and familiarity with a given domain (Xu et al., 2024b). For example, a health data entry system designed for older adults might adjust how it highlights input errors or suggests alternatives (Cuadra et al., 2024). In contrast, enterprise software might generate justifications that align with domain-specific workflows and terminology (Srinivas et al., 2024).

以用户为中心的设计侧重于根据个人的上下文、偏好和领域熟悉度定制解释 (Xu 等人, 2024b)。例如，为老年人设计的健康数据录入系统可能调整错误提示或建议替代方案的方式 (Cuadra 等人, 2024)。相比之下，企业软件可能生成符合特定领域工作流程和术语的理由 (Srinivas 等人, 2024)。

Beyond presentation, recent research explores how GUI agents can generate inherently interpretable outputs grounded in user-understandable concepts. For example, in vision-based tasks, synthesizing explanations with human-verifiable visual features has been shown to improve model reasoning and transparency, which could potentially enable GUI agents to justify actions in complex visual environments (Shi et al., 2025). On the language side, interpreting LLM representations using mutual information and sparse activations allows for controllable, semantically meaningful explanations, offering a promising direction for more steerable and trustworthy GUI behaviors (Wu et al., 2025b). These techniques bridge model internals with users, helping GUI agents adapt explanation styles while maintaining transparency and alignment.

除了展示，近期研究探索 GUI 代理如何生成基于用户可理解概念的内在可解释输出。例如，在基于视觉的任务中，合成人类可验证的视觉特征解释已被证明能提升模型推理和透明度，可能使 GUI 代理在复杂视觉环境中为行动提供理由 (Shi 等人, 2025)。在语言方面，利用互信息和稀疏激活解释大型语言模型表示，实现可控且语义丰富的解释，为更可引导且可信的 GUI 行为提供了有前景的方向 (Wu 等人, 2025b)。这些技术架起模型内部与用户之间的桥梁，帮助 GUI 代理在保持透明度和一致性的同时调整解释风格。

5.4 Future Directions

5.4 未来方向

Enhancing explainability and transparency in GUI agents requires practical solutions that balance technical depth with user accessibility. Two promising directions emerge:

提升 GUI 代理的可解释性与透明度需要兼顾技术深度与用户可访问性的实用方案。两条有前景的方向浮现:

Interactive Explanation Tools: Real-time visualization of agent reasoning could transform how users understand automated processes. Browser extensions could display decision chains (e.g., "identify search box → enter query → select result") using interactive flowcharts that visualize the agent's current focus, building on techniques from EBC-LLMAgent (Guan et al., 2024). On mobile devices, lightweight on-device models could highlight interface elements being analyzed without cloud processing delays, extending approaches from hierarchical agent designs (Liu et al., 2025a; Zhu et al., 2025). When operations fail, automated error playback could compare intended versus actual results, incorporating reflective techniques from XA-gent (Nguyen et al., 2024c) to make troubleshooting intuitive.

交互式解释工具: 实时可视化代理推理过程有望彻底改变用户对自动化流程的理解。浏览器扩展可以使用交互式流程图展示决策链 (例如，“识别搜索框 → 输入查询 → 选择结果”)，通过可视化代理当前关注点，基于 EBC-LLMAgent(Guan 等，2024) 的技术。在移动设备上，轻量级的本地模型能够突出显示正在分析的界面元素，避免云端处理延迟，扩展了分层代理设计的方法 (Liu 等，2025a；Zhu 等，2025)。当操作失败时，自动错误回放可以比较预期与实际结果，结合 XA-gent(Nguyen 等，2024c) 中的反思技术，使故障排查更加直观。

Context-Adaptive Explanations: Different users require different types of transparency. Future systems should provide role-based explanations, offering technical details for developers while generating simplified summaries for general users, extending the user-centric approaches seen in (Cuadra et al., 2024). Cultural context filters could automatically adjust explanation styles and privacy considerations based on regional norms, addressing localization challenges. Accessibility-focused explanation channels (such as voice explanations for visually impaired users) would ensure transparency benefits reach diverse populations, aligning with inclusive design principles (Xu et al., 2024b).

情境自适应解释: 不同用户需要不同类型的透明度。未来系统应提供基于角色的解释，为开发者提供技术细节，同时为普通用户生成简化摘要，扩展了 (Cuadra 等，2024) 中以用户为中心的方法。文化背景过滤器可根据地区规范自动调整解释风格和隐私考量，解决本地化挑战。面向无障碍的解释渠道 (如为视障用户提供语音解释) 将确保透明度惠及多样化人群，符合包容性设计原则 (Xu 等，2024b)。

6 Ethical Implications

6 伦理影响

Developing GUI agents responsibly entails going beyond technical design to incorporate ethical principles, cultural sensitivity, and policy considerations. This section highlights core guidelines, discusses the need for cultural and social awareness, and addresses regulatory and policy implications.

负责任地开发 GUI 代理不仅涉及技术设计，还需融入伦理原则、文化敏感性和政策考量。本节重点介绍核心指导方针，讨论文化与社会意识的必要性，并涉及监管与政策影响。

6.1 Cultural and Social Awareness

6.1 文化与社会意识

Agents that serve diverse user groups need to account for cultural context and social norms (Qiu et al., 2024). For example, platforms in e-commerce or online discussions often contain content that carries culturally specific meaning, which requires context-aware handling. Benchmarks like CASA measure assess how well agents navigate these cross-cultural settings without overstepping boundaries (Qiu et al., 2024). Similarly, frameworks that embed moral reasoning (Piatti et al., 2024) encourage cooperative behaviors aligned with universalized ethical principles. Recent work argues that cultural NLP often lacks a unified theoretical foundation, emphasizing the need for localization-focused approaches rather than relying on static cultural templates (Zhou et al., 2025).

服务多元用户群的代理需考虑文化背景和社会规范 (Qiu 等，2024)。例如，电商或在线讨论平台常包含具有特定文化意义的内容，需进行情境感知处理。CASA 等基准测试评估代理在跨文化环境中导航的能力，确保不越界 (Qiu 等，2024)。类似地，嵌入道德推理的框架 (Piatti 等，2024) 鼓励符合普遍伦理原则的合作行为。近期研究指出，文化自然语言处理 (NLP) 常缺乏统一理论基础，强调需采用以本地化为中心的方法，而非依赖静态文化模板 (Zhou 等，2025)。

Benchmark	Focus Area	Key Metrics
Security & Privacy Evaluation		
InjecAgent (Zhan et al., 2024)	Tool-integrated agent vulnerability	Attack success rate across 17 user tools
BrowserART (Kumar et al., 2024)	Browser agent jailbreaking	Harmful behavior attempt rate
AdvWeb (Xu et al., 2024a)	Black-box adversarial web attacks	Stealth effectiveness, Success rate
EIA (Liao et al., 2024)	Web agent privacy risks	PII extraction rate, Attack detection
PUPA (Siyan et al., 2024)	Privacy-preserving evaluation	PII exposure, Response quality
ARE (Wu et al., 2025a)	Adversarial robustness	Flow of adversarial information
Harmfulness & Reliability Assessment		
Agent-SafetyBench (Zhang et al., 2024c)	Comprehensive agent safety	Safety scores across 8 risk categories
AgentHarm (Andriushchenko et al., 2024)	Harmfulness assessment	Refusal rate, Task completion
MobileSafetyBench (Lee et al., 2024)	Mobile device control safety	Risk management, Injection resistance
ST-WebAgentBench (Levy et al., 2024)	Web safety and trustworthiness	Completion Under Policy, Risk Ratio
GTArena (Zhao et al., 2024)	Automated GUI testing	Test intention, Defect detection
MM-SafetyBench (Liu et al., 2023b)	Image-based manipulations	Visual attack resilience
Human & Cultural Alignment		
MSSBench (Zhou et al., 2024)	Multimodal situational safety	Safety reasoning, Visual understanding
CASA (Qiu et al., 2024)	Cultural and social awareness	Awareness coverage, Violation rate

基准测试	关注领域	关键指标
安全与隐私评估		
InjecAgent (Zhan 等, 2024)	工具集成代理漏洞	17 种用户工具的攻击成功率
BrowserART (Kumar 等, 2024)	浏览器代理越狱	有害行为尝试率
AdvWeb (Xu 等, 2024a)	黑盒对抗性网页攻击	隐蔽性效果，成功率
EIA (Liao 等, 2024)	网页代理隐私风险	个人身份信息 (PII) 提取率，攻击检测
PUPA (Siyan 等, 2024)	隐私保护评估	个人身份信息 (PII) 泄露，响应质量
ARE (Wu 等, 2025a)	对抗鲁棒性	对抗信息流
有害性与可靠性评估		
Agent-SafetyBench (Zhang 等, 2024c)	综合代理安全性	8 类风险的安全评分
AgentHarm (Andriushchenko 等, 2024)	有害性评估	拒绝率，任务完成度
MobileSafetyBench (Lee 等, 2024)	移动设备控制安全	风险管理，注入抵抗力
ST-WebAgentBench (Levy 等, 2024)	网页安全与可信度	策略下完成率，风险比
GTArena (Zhao 等, 2024)	自动化图形用户界面 (GUI) 测试	测试意图，缺陷检测
MM-SafetyBench (Liu 等, 2023b)	基于图像的操控	视觉攻击韧性
人类与文化对齐		
MSSBench (Zhou 等, 2024)	多模态情境安全	安全推理，视觉理解
CASA (Qiu 等, 2024)	文化与社会意识	意识覆盖率，违规率

Table 2: Taxonomy of GUI Agent Evaluation Frameworks. The benchmarks are categorized into three dimensions: (1) security and privacy evaluation, focusing on vulnerability assessment and attack resistance; (2) harmfulness and reliability assessment, measuring agent compliance with safety protocols and failure modes; and (3) human and cultural alignment, evaluating agents' ability to handle visual manipulations and conform to social norms.

表 2:GUI 代理评估框架分类。基准测试分为三个维度:(1) 安全与隐私评估，侧重于漏洞评估和攻击抵抗能力；(2) 有害性与可靠性评估，衡量代理对安全协议的遵守情况及故障模式；(3) 人类与文化适应性，评估代理处理视觉操控和遵守社会规范的能力。

At the same time, agents also need to address accessibility, meeting the needs of older adults or individuals with sensory impairments (Cuadra et al., 2024). Designing flexible interaction paths, whether through speech, visual cues, or textual descriptions, will allow broader inclusivity. As technology advances, bridging cultural gaps and ensuring accessibility will likely require more elaborate training data and dedicated modules.

同时，代理还需关注无障碍设计，满足老年人或感官障碍者的需求 (Cuadra 等，2024)。设计灵活的交互路径，无论是通过语音、视觉提示还是文本描述，都能实现更广泛的包容性。随着技术进步，弥合文化差异并确保无障碍访问，可能需要更复杂的训练数据和专门模块。

6.2 Policy Implications

6.2 政策影响

Because GUI agents can execute complex actions with real-world consequences, policy and regulatory considerations are paramount (Gan et al., 2024; Chen et al., 2024a). In regulated sectors such as healthcare or finance, compliance with data protection requirements becomes mandatory. Meanwhile, governments and institutions face difficulties in overseeing technologies that are rapidly evolving and often proprietary. Decentralized governance

frameworks, such as those leveraging blockchain, have been proposed to enhance transparency, accountability, and decision rights in foundation-model-based AI systems (Liu et al., 2023d).

由于 GUI 代理能够执行具有现实后果的复杂操作，政策和监管考量至关重要 (Gan 等，2024；Chen 等，2024a)。在医疗或金融等受监管领域，遵守数据保护要求成为强制性。与此同时，政府和机构面临监管快速发展且常为专有技术的挑战。基于区块链等去中心化治理框架被提议用于提升基础模型 (foundation-model)AI 系统的透明度、问责性和决策权 (Liu 等，2023d)。

Several initiatives encourage open-sourcing benchmarks and best practices (Levy et al., 2024), fostering community-driven standards for agent safety. The collaboration between industry, academia, and policymakers could also help clarify rules around data use and accountability. In the long term, building in responsible practices, through clear guidelines, strong evaluations, and cross-sector oversight, can better align GUI agents with societal values while supporting innovation.

多个倡议鼓励开源基准和最佳实践 (Levy 等，2024)，促进社区驱动的代理安全标准。产业、学术界与政策制定者的合作也有助于明确数据使用和问责规则。长期来看，通过明确指南、严格评估和跨部门监督，内嵌负责任实践可更好地使 GUI 代理符合社会价值，同时支持创新。

6.3 Guidelines and Principles

6.3 指导原则

Some efforts have focused on formalizing design principles through pattern-based architectures that ensure security, accountability, and fairness across the agent's lifecycle (Lu et al., 2023; Wu et al., 2024c). Modular systems make it easier to trace how different components handle data and interact, improving transparency and alignment with user inputs (Zhang et al., 2023; Hua et al., 2024). On the security side, newer authentication schemes aim to tighten control over delegation, making it harder for agents to take unauthorized actions while keeping the chain of responsibility clear (South et al., 2025).

部分工作致力于通过基于模式的架构形式化设计原则，确保代理生命周期内的安全性、问责性和公平性 (Lu 等，2023；Wu 等，2024c)。模块化系统便于追踪不同组件如何处理数据和交互，提升透明度并与用户输入保持一致 (Zhang 等，2023；Hua 等，2024)。在安全方面，更新的认证方案旨在加强委托控制，防止代理执行未授权操作，同时保持责任链清晰 (South 等，2025)。

In real-world settings, developers must consider both the power and risks of autonomy. When agents handle critical tasks, such as financial transactions or medical record management, clear guidelines for fallback procedures and user oversight should be essential (Wright, 2024). Recent studies also highlight ethical concerns beyond security, such as how interactions with agents may inadvertently shape user beliefs, with evidence showing that LLM-powered conversational agents can significantly amplify false memories in sensitive contexts like witness interviews (Chan et al., 2024).

在实际环境中，开发者必须权衡自治的能力与风险。当代理处理关键任务，如金融交易或医疗记录管理时，明确的回退程序和用户监督指南至关重要 (Wright，2024)。最新研究还强调安全之外的伦理问题，例如代理交互可能无意中影响用户信念，有证据表明基于大型语言模型 (LLM) 的对话代理在敏感场景如证人访谈中显著放大虚假记忆 (Chan 等，2024)。

7 Evaluation Frameworks and Benchmarks

7 评估框架与基准

Evaluating GUI agents requires solid frameworks to assess reliability and trust. This section covers current metrics, practical evaluation methods, and trustworthiness-specific benchmarks used to test performance and behavior. Table 2 summarizes key evaluation frameworks across different dimensions.

评估 GUI 代理需要稳健的框架以衡量其可靠性和信任度。本节涵盖当前指标、实用评估方法及专门测试性能和行为的信任度基准。表 2 总结了不同维度的关键评估框架。

7.1 Metrics for Assessing Trustworthiness

7.1 信任度评估指标

Evaluation often begins with task completion: whether the agent navigates, inputs data, or detects anomalies accurately (Koh et al., 2024a; Chen et al., 2024c). However, success rate alone cannot capture trustworthiness. ST-WebAgentBench, for example, evaluates how well agents follow explicit policy rules, flagging any violations as signs of unsafe behavior (Levy et al., 2024). To detect problems earlier, intermediate metrics like URL or form field matching are also used to pinpoint where agents make mistakes (Zhou et al., 2023; Shi et al., 2017; Yao et al., 2022).

评估通常从任务完成度开始: 代理是否准确导航、输入数据或检测异常 (Koh 等，2024a；Chen 等，2024c)。但成功率不足以全面反映信任度。例如，ST-WebAgentBench 评估代理遵守明确政策规则的程度，任何违规均视为不安全行为 (Levy 等，2024)。为提前发现问题，还使用中间指标如 URL 或表单字段匹配，定位代理出错环节 (Zhou 等，2023；Shi 等，2017；Yao 等，2022)。

Researchers also propose metrics for robustness under adversarial conditions (Wu et al., 2025a), cultural or social awareness (Qiu et al., 2024), and situational safety (Zhou et al., 2024; Liu et al., 2023b). These approaches emphasize that a reliable GUI agent must not only achieve the user's intended outcome but also demonstrate safe and consistent behavior throughout the process.

研究者还提出了针对对抗条件下的鲁棒性 (Wu 等，2025a)、文化或社会意识 (Qiu 等，2024) 及情境安全 (Zhou 等，2024；Liu 等，2023b) 的指标。这些方法强调，可靠的 GUI 代理不仅要实现用户预期结果，还需在整个过程中表现出安全且一致的行为。

7.2 Comprehensive Evaluation Techniques

7.2 综合评估技术

Comprehensive frameworks often adopt a modular approach. ChEF (Comprehensive Evaluation Framework) systematically tests scenario variation, instruction diversity, inference strategies, and flexible metrics (Shi et al., 2023b). GTArena partitions automated GUI testing into intent generation, test execution, and defect detection for mobile apps (Zhao et al., 2024). By capturing multiple facets, including correctness, error handling, and safety, such evaluations reveal more profound insights into agent behavior.

综合框架通常采用模块化方法。ChEF(综合评估框架) 系统测试场景变化、指令多样性、推理策略及灵活指标 (Shi 等，2023b)。GTArena 将移动应用自动化 GUI 测试划分为意图生成、测试执行和缺陷检测 (Zhao 等，2024)。通过捕捉正确性、错误处理和安全性等多方面，这类评估揭示了代理行为的更深层次洞见。

Distinctions between closed-world and open-world tests matter for ecological validity. Closed-world environments, like curated sets of web pages, enable controlled experimentation but lack real-world unpredictability. Open-world evaluations allow dynamic changes and unknown interfaces (Chen et al., 2024c; He et al., 2024), forcing agents to adapt. Balancing reproducibility and realism remains an ongoing challenge.

封闭世界测试与开放世界测试的区别对于生态有效性至关重要。封闭世界环境，如精心策划的网页集合，便于控制实验，但缺乏现实世界的不可预测性。开放世界评估允许动态变化和未知接口 (Chen et al., 2024c；He et al., 2024)，迫使智能体适应。如何平衡可重复性与现实性仍是一个持续的挑战。

7.3 Case Studies and Benchmarks

7.3 案例研究与基准测试

Numerous case studies develop domain-specific or specialized benchmarks. Mind2Web measures task completion on live websites, revealing difficulties in grounding instructions (Zheng et al., 2024). MSSBench focuses on ”multimodal situational safety,” where half of the query-image pairs require context-sensitive reasoning (Zhou et al., 2024). Similarly, VETL (Wang et al., 2024b) and WebCanvas (Koh et al., 2024a) test web GUI interactions and bug detection.

大量案例研究开发了特定领域或专业化的基准测试。Mind2Web 衡量在实时网站上的任务完成情况，揭示了指令落地的难点 (Zheng et al., 2024)。MSSBench 聚焦于“多模态情境安全”(multimodal situational safety)，其中一半的查询-图像对需要上下文敏感的推理 (Zhou et al., 2024)。类似地，VETL(Wang et al., 2024b) 和 WebCanvas(Koh et al., 2024a) 测试网页 GUI 交互和漏洞检测。

These benchmarks illustrate that trustworthy evaluation is inherently multifaceted. Future work could unify the disparate tasks, data sources, and metrics into more holistic frameworks, enabling meaningful comparisons of agents’ safety, robustness, and usability. Such efforts will be critical for driving standardization and progress in this rapidly evolving domain.

这些基准测试表明，可信评估本质上是多维度的。未来工作可将不同任务、数据源和指标统一到更全面的框架中，从而实现对智能体安全性、鲁棒性和可用性的有意义比较。这类努力对于推动该快速发展领域的标准化和进步至关重要。

8 Conclusion

8 结论

This survey has examined trustworthiness in GUI agents across five critical dimensions: security vulnerabilities, reliability, explainability, ethical alignment, and evaluation methodologies. Our analysis reveals significant challenges at the intersection of these dimensions, where multimodal interactions create novel attack surfaces and failure modes that traditional approaches cannot adequately address. While research has primarily focused on functional performance, the integrated nature of GUI agents demands holistic approaches to trustworthiness that span their entire operational pipeline.

本综述考察了 GUI 智能体可信性的五个关键维度: 安全漏洞、可靠性、可解释性、伦理一致性和评估方法。我们的分析揭示了这些维度交汇处的重大挑战，多模态交互产生了传统方法难以充分应对的新型攻击面和失效模式。尽管研究主要关注功能性能，GUI 智能体的集成特性要求跨越其整个操作流程的整体可信性方法。

Looking forward, advancing trustworthy GUI agents will require: (1) robust multimodal defense mechanisms that protect against adversarial manipulations, (2) adaptive safety frameworks that balance autonomy with protection, and (3) user-centered transparency systems that make agent reasoning accessible without compromising security. Evaluation benchmarks should assess both capability and safety. With the right safeguards and cross-field collaboration, GUI agents can be made effective, secure, and aligned with human values.

展望未来，推进可信 GUI 智能体需要:(1) 防御对抗性操控的强健多模态防护机制，(2) 在自主性与保护之间平衡的自适应安全框架，以及 (3) 以用户为中心的透明系统，使智能体推理可访问且不损害安全。评估基准应同时考察能力与安全性。通过适当的保障措施和跨领域协作，GUI 智能体能够实现高效、安全并符合人类价值观。

References

参考文献

Saaket Agashe, Jiuzhou Han, Shuyu Gan, Jiachen Yang, Ang Li, and Xin Eric Wang. 2024. Agent s: An open agentic framework that uses computers like a human. arXiv preprint arXiv:2410.08164.

Saaket Agashe, Jiuzhou Han, Shuyu Gan, Jiachen Yang, Ang Li, and Xin Eric Wang. 2024. Agent s: An open agentic framework that uses computers like a human. arXiv preprint arXiv:2410.08164.

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.

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.

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

Taylor B. Arnold and Lauren Tilton. 2024. Explainable search and discovery of visual cultural heritage collections with multimodal large language models. Workshop on Computational Humanities Research.

Taylor B. Arnold and Lauren Tilton. 2024. Explainable search and discovery of visual cultural heritage collections with multimodal large language models. Workshop on Computational Humanities Research.

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.

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.

Zana Buçinca, Chau Minh Pham, Maurice Jakesch, Marco Tulio Ribeiro, Alexandra Olteanu, and Saleema Amershi. 2023. Aha!: Facilitating AI Impact Assessment by Generating Examples of Harms. arXiv.org.

Zana Buçinca, Chau Minh Pham, Maurice Jakesch, Marco Tulio Ribeiro, Alexandra Olteanu, and Saleema Amershi. 2023. Aha!: Facilitating AI Impact Assessment by Generating Examples of Harms. arXiv.org.

Tri Cao, Chengyu Huang, Yuexin Li, Huilin Wang, Amy He, Nay Oo, and Bryan Hooi. 2024. Phishagent: A robust multimodal agent for phishing webpage detection. arXiv preprint arXiv:2408.10738.

Tri Cao, Chengyu Huang, Yuexin Li, Huilin Wang, Amy He, Nay Oo, and Bryan Hooi. 2024. Phishagent: A robust multimodal agent for phishing webpage detection. arXiv preprint arXiv:2408.10738.

Hyungjoo Chae, Namyoung Kim, Kai Tzu-iunn Ong, Minju Gwak, Gwanwoo Song, Jihoon Kim, Sungh-wan 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.

Hyungjoo Chae, Namyoung Kim, Kai Tzu-iunn Ong, Minju Gwak, Gwanwoo Song, Jihoon Kim, Sungh-wan 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.

Samantha Chan, Pat Pataranutaporn, Aditya Suri, Wazeer Zulfikar, Pattie Maes, and Elizabeth F Loftus. 2024. Conversational ai powered by large language models amplifies false memories in witness interviews. arXiv preprint arXiv:2408.04681.

Samantha Chan, Pat Pataranutaporn, Aditya Suri, Wazeer Zulfikar, Pattie Maes, 和 Elizabeth F Loftus. 2024. 由大型语言模型驱动的对话式人工智能在证人访谈中放大虚假记忆。arXiv 预印本 arXiv:2408.04681。

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.

Chaoran Chen, Daodao Zhou, Yanfang Ye, Toby Li, 和 Yaxing Yao. 2024a. Clear: 面向大型语言模型应用的上下文隐私政策分析与风险生成。arXiv 预印本 arXiv:2410.13387。

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, et al. 2024b. From persona to personalization: A survey on role-playing language agents. arXiv preprint arXiv:2404.18231.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu 等. 2024b. 从角色扮演到个性化: 角色扮演语言代理的综述。arXiv 预印本 arXiv:2404.18231。

Qi Chen, Dileepa Pitawela, Chongyang Zhao, Gengze Zhou, Hsiang-Ting Chen, and Qi Wu. 2024c. We-bvln: Vision-and-language navigation on websites. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pages 1165-1173.

Qi Chen, Dileepa Pitawela, Chongyang Zhao, Gengze Zhou, Hsiang-Ting Chen, 和 Qi Wu. 2024c. We-bvln: 基于网站的视觉与语言导航。载于 AAAI 人工智能会议论文集，第 38 卷，页 1165-1173。

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

Xiang Chen, Chenxi Wang, Yida Xue, Ningyu Zhang, Xiaoyan Yang, Qiang Li, Yue Shen, Lei Liang, Jinjie Gu, 和 Huajun Chen. 2024d. 多模态大型语言模型的统一幻觉检测。计算语言学协会年会论文集。

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

Yurun Chen, Xueyu Hu, Keting Yin, Juncheng Li, 和 Shengyu Zhang. 2025. Aeia-mn: 评估多模态大型语言模型驱动的移动代理对主动环境注入攻击的鲁棒性。arXiv 预印本 arXiv:2502.13053。

Andrea Cuadra, Justine Breuch, Samantha Estrada, David Ihim, Isabelle Hung, Derek Askaryar, Marwan Hassanien, K. Fessele, and J. Landay. 2024. Digital forms for all: A holistic multimodal large language model agent for health data entry. Proceedings of the ACM on Interactive Mobile Wearable and Ubiquitous Technologies.

Andrea Cuadra, Justine Breuch, Samantha Estrada, David Ihim, Isabelle Hung, Derek Askaryar, Marwan Hassanien, K. Fessele, 和 J. Landay. 2024. 面向所有人的数字表单: 用于健康数据录入的整体多模态大型语言模型代理。ACM 交互式移动可穿戴与普适技术会议论文集。

Dazhen Deng, Chuhan Zhang, Hongxing Fan, Zhen-fei Yin, Lu Sheng, Yu Qiao, and Jing Shao. 2024. Adversaflow: Visual red teaming for large language models with multi-level adversarial flow. IEEE Transactions on Visualization and Computer Graphics.

Dazhen Deng, Chuhan Zhang, Hongxing Fan, Zhen-fei Yin, Lu Sheng, Yu Qiao, 和 Jing Shao. 2024. Adver-saflow: 针对大型语言模型的多层次对抗流视觉红队方法。IEEE 可视化与计算机图形学汇刊。

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.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, 和 Yu Su. 2023. Mind2web: 迈向通用网络代理。神经信息处理系统进展，36 卷:28091-28114。

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

Xiaohan Fu, Shuheng Li, Zihan Wang, Yihao Liu, Rajesh K. Gupta, Taylor Berg-Kirkpatrick, 和 Earlence Fernandes. 2024. Imprompter: 诱导大型语言模型代理不当使用工具。arXiv 预印本 arXiv:2410.14923。

Roberto Gallotta, Antonios Liapis, and Georgios N Yannakakis. 2025. Freyr: A framework for recognizing and executing your requests. arXiv preprint arXiv:2501.12423.

Roberto Gallotta, Antonios Liapis, 和 Georgios N Yannakakis. 2025. Freyr: 识别并执行您的请求的框架。arXiv 预印本 arXiv:2501.12423。

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 Shoul-ing Ji. 2024. Navigating the risks: A survey of security, privacy, and ethics threats in llm-based agents. arXiv preprint arXiv:2411.09523.

Yuyou Gan, Yong Yang, Zhen Ma, Ping He, Rui Zeng, Yiming Wang, Qingming Li, Chunyi Zhou, Songze Li, Ting Wang, Yunjun Gao, Yingcai Wu, 和 Shouling Ji. 2024. 风险导航: 基于大型语言模型代理的安全、隐私与伦理威胁综述。arXiv 预印本 arXiv:2411.09523。

Difei Gao, Lei Ji, Zechen Bai, Mingyu Ouyang, Peiran Li, Dongxing Mao, Qinchen Wu, Weichen Zhang, Peiyi Wang, Xiangwu Guo, et al. 2023. Assistgui: Task-oriented desktop graphical user interface automation. arXiv preprint arXiv:2312.13108.

Difei Gao, Lei Ji, Zechen Bai, Mingyu Ouyang, Peiran Li, Dongxing Mao, Qinchen Wu, Weichen Zhang, Peiyi Wang, Xiangwu Guo 等. 2023. Assistgui: 面向任务的桌面图形用户界面自动化。arXiv 预印本 arXiv:2312.13108。

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

Jiahui Gao, Renjie Pi, Tianyang Han, Han Wu, Chenyang Lyu, Huayang Li, Lanqing Hong, Ling-peng Kong, Xin Jiang, 和 Zhenguo Li. 2024. Coca: 通过宪法校准恢复多模态大型语言模型的安全意识。arXiv 预印本 arXiv:2409.11365。

Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. 2023. Ai Control: Improving Safety

Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan 和 Fabien Roger. 2023. AI 控制: 提升安全性

Despite Intentional Subversion. International Conference on Machine Learning.

尽管存在故意颠覆。国际机器学习会议。

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

Yu Gu, Boyuan Zheng, Boyu Gou, Kai Zhang, Cheng Chang, Sanjari Srivastava, Yanan Xie, Peng Qi, Huan Sun 和 Yu Su. 2024a. 你的大型语言模型 (LLM) 是否暗中成为互联网的世界模型? 基于模型的网页代理规划。arXiv 预印本 arXiv:2411.06559。

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

Yu Gu, Boyuan Zheng, Boyu Gou, Kai Zhang, Cheng Chang, Sanjari Srivastava, Yanan Xie, Peng Qi, Huan Sun 和 Yu Su. 2024b. 你的大型语言模型 (LLM) 是否暗中成为互联网的世界模型? 基于模型的网页代理规划。arXiv 预印本 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.

Yanchu Guan, Dong Wang, Yan Wang, Haiqing Wang, Renen Sun, Chenyi Zhuang, Jinjie Gu 和 Zhixuan Chu. 2024. 可解释行为克隆: 通过示范学习教导大型语言模型代理。arXiv 预印本 arXiv:2410.22916。

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2023. Detecting and Preventing Hallucinations in Large Vision Language Models. AAAI Conference on Artificial Intelligence.

Anisha Gunjal, Jihan Yin 和 Erhan Bas. 2023. 检测与防止大型视觉语言模型中的幻觉。AAAI 人工智能会议。

Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Yong Dai, Hongming Zhang, Zhenzhong Lan, and Dong Yu. 2024. Webvoyager: Building an end-to-end web agent with large multimodal models. arXiv preprint arXiv:2401.13919.

Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Yong Dai, Hongming Zhang, Zhenzhong Lan 和 Dong Yu. 2024. Webvoyager: 利用大型多模态模型构建端到端网页代理。arXiv 预印本 arXiv:2401.13919。

Paul RB Houssel, Priyanka Singh, Siamak Layeghy, and Marius Portmann. 2024. Towards explainable network intrusion detection using large language models. arXiv preprint arXiv:2408.04342.

Paul RB Houssel, Priyanka Singh, Siamak Layeghy 和 Marius Portmann. 2024. 迈向可解释的网络入侵检测，基于大型语言模型。arXiv 预印本 arXiv:2408.04342。

Xueyu Hu, Tao Xiong, Biao Yi, Zishu Wei, Ruixuan Xiao, Yurun Chen, Jiasheng Ye, Meiling Tao, Xi-angxin Zhou, Ziyu Zhao, et al. Os agents: A survey on mllm-based agents for computer, phone and browser use.

Xueyu Hu, Tao Xiong, Biao Yi, Zishu Wei, Ruixuan Xiao, Yurun Chen, Jiasheng Ye, Meiling Tao, Xi-angxin Zhou, Ziyu Zhao 等. 操作系统代理: 基于多模态大型语言模型 (MLLM) 的计算机、手机和浏览器代理综述。

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.

Wenyue Hua, Xianjun Yang, Zelong Li, Cheng Wei 和 Yongfeng Zhang. 2024. Trustagent: 迈向安全可信赖的大型语言模型代理。自然语言处理实证方法会议。

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024a. Opera: Alleviating Hallucination in Multi-Modal Large Language Models via Over-Trust Penalty and Retrospection-Allocation. In 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 13418-13427. IEEE.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang 和 Nenghai Yu. 2024a. OPERA: 通过过度信任惩罚和回顾分配缓解多模态大型语言模型中的幻觉。2024 年 IEEE/CVF 计算机视觉与模式识别会议 (CVPR)，第 13418-13427 页。IEEE。

Zhipeng Huang, Zhizheng Zhang, Yiting Lu, Zheng-Jun Zha, Zhibo Chen, and Baining Guo. 2024b. Visualcritic: Making lmms perceive visual quality like humans. arXiv preprint arXiv:2403.12806.

Zhipeng Huang, Zhizheng Zhang, Yiting Lu, Zheng-Jun Zha, Zhibo Chen 和 Baining Guo. 2024b. Visualcritic: 让大型多模态模型像人类一样感知视觉质量。arXiv 预印本 arXiv:2403.12806。

Pete Janowczyk, Linda Laurier, Ave Giulietta, Arlo Octavia, and Meade Cleti. 2024. Seeing is deceiving: Exploitation of visual pathways in multi-modal language models. arXiv preprint arXiv:2411.05056.

Pete Janowczyk, Linda Laurier, Ave Giulietta, Arlo Octavia 和 Meade Cleti. 2024. 眼见未必为实: 多模态语言模型中视觉通路的利用。arXiv 预印本 arXiv:2411.05056。

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.

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang 和 Shikun Zhang. 2024a. 幻觉增强对比学习用于多模态大型语言模型。2024 年 IEEE/CVF 计算机视觉与模式识别会议 (CVPR)，第 27026-27036 页。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.

Yilei Jiang, Yingshui Tan 和 Xiangyu Yue. 2024b. Rapguard: 通过理据感知防御提示保护多模态大型语言模型。arXiv 预印本 arXiv:2412.18826。

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.

Hanna Kim, Minkyoo Song, Seung Ho Na, Seungwon Shin 和 Kimin Lee. 2024a. 当大型语言模型上线: 网络启用大型语言模型的新兴威胁。arXiv 预印本 arXiv:2410.14569。

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

Jaekyeom Kim, Dong-Ki Kim, Lajanugen Logeswaran, Sungryull Sohn, 和 Honglak Lee. 2024b. Auto-intent: 大型语言模型网络代理的自动意图发现与自我探索。arXiv 预印本 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, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Chong Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Ruslan Salakhutdinov, 和 Daniel Fried. 2024a. Visualwebarena: 在真实视觉网络任务中评估多模态代理。arXiv 预印本 arXiv:2401.13649。

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

Jing Yu Koh, Stephen McAleer, Daniel Fried, 和 Ruslan Salakhutdinov. 2024b. 语言模型代理的树搜索。arXiv 预印本 arXiv:2407.01476。

Priyanshu Kumar, Elaine Lau, Saranya Vijayaku-mar, 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.

Priyanshu Kumar, Elaine Lau, Saranya Vijayakumar, Tu Trinh, Scale Red Team, Elaine Chang, Vaughn Robinson, Sean Hendryx, Shuyan Zhou, Matt Fredrikson, Summer Yue, 和 Zifan Wang. 2024. 拒绝训练的大型语言模型作为浏览器代理容易被破解。arXiv 预印本 arXiv:2410.13886。

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

Juyong Lee, Dongyoon Hahm, June Suk Choi, W Bradley Knox, 和 Kimin Lee. 2024. Mobilesafetybench: 评估移动设备控制中自主代理的安全性。arXiv 预印本 arXiv:2410.17520。

Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Min-joon Seo. 2023. Volcano: mitigating multimodal hallucination through self-feedback guided revision. arXiv preprint arXiv:2311.07362.

Seongyun Lee, Sue Hyun Park, Yongrae Jo, 和 Minjoon Seo. 2023. Volcano: 通过自我反馈引导修正缓解多模态幻觉。arXiv 预印本 arXiv:2311.07362。

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.

Ido Levy, Ben Wiesel, Sami Marreed, Alon Oved, Avi Yaeli, 和 Segev Shlomov. 2024. St-webagentbench: 用于评估网络代理安全性和可信度的基准。arXiv 预印本 arXiv:2410.06703。

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.

Zeyi Liao, Lingbo Mo, Chejian Xu, Mintong Kang, Jiawei Zhang, Chaowei Xiao, Yuan Tian, Bo Li, 和 Huan Sun. 2024. Eia: 针对通用网络代理的环境注入攻击导致隐私泄露。arXiv 预印本 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.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, 和 Lijuan Wang. 2023a. 通过稳健指令调优缓解大型多模态模型中的幻觉。arXiv.org。

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

Haowei Liu, Xi Zhang, Haiyang Xu, Yuyang Wanyan, Junyang Wang, Ming Yan, Ji Zhang, Chunfeng Yuan, Changsheng Xu, Weiming Hu, 等. 2025a. Pc-agent: 用于 PC 复杂任务自动化的分层多代理协作框架。arXiv 预印本 arXiv:2502.14282。

William Liu, Liang Liu, Yaxuan Guo, Han Xiao, Weifeng Lin, Yuxiang Chai, Shuai Ren, Xiaoyu Liang, Linghao Li, Wenhao Wang, et al. 2025b. Llm-powered gui agents in phone automation: Surveying progress and prospects.

William Liu, Liang Liu, Yaxuan Guo, Han Xiao, Weifeng Lin, Yuxiang Chai, Shuai Ren, Xiaoyu Liang, Linghao Li, Wenhao Wang, 等. 2025b. 基于大型语言模型的手机自动化 GUI 代理: 进展与前景综述。

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.

Xin Liu, Yichen Zhu, Jindong Gu, Yunshi Lan, Chao Yang, 和 Yu Qiao. 2023b. Mm-safetybench: 多模态大型语言模型安全性评估基准。欧洲计算机视觉会议。

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.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, 和 Hang Li. 2023c. 可信大型语言模型: 评估大型语言模型对齐的综述与指南。arXiv 预印本 arXiv:2308.05374。

Yue Liu, Qinghua Lu, Liming Zhu, and Hye-Young Paik. 2023d. Decentralised governance-driven architecture for designing foundation model based systems: Exploring the role of blockchain in responsible ai. arXiv preprint arXiv:2308.05962.

Yue Liu, Qinghua Lu, Liming Zhu, 和 Hye-Young Paik. 2023d. 基于去中心化治理的基础模型系统设计架构: 探索区块链在负责任人工智能中的作用。arXiv 预印本 arXiv:2308.05962。

Zheyuan Liu, Guangyao Dou, Zhaoxuan Tan, Yijun Tian, and Meng Jiang. 2024. Towards Safer Large Language Models through Machine Unlearning. Annual Meeting of the Association for Computational Linguistics.

Zheyuan Liu, Guangyao Dou, Zhaoxuan Tan, Yijun Tian, 和 Meng Jiang. 2024. 通过机器遗忘迈向更安全的大型语言模型。计算语言学协会年会。

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.

陆清华，朱黎明，徐希伟，邢振昌，斯特凡·哈勒，乔恩·惠特尔。2023。迈向负责任的生成式人工智能: 基于基础模型的智能体设计参考架构。载于 2024 年 IEEE 第 21 届国际软件架构会议伴随会议 (ICSA-C)。IEEE。

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.

马新北，王怡婷，姚瑶，袁同欣，张安斯顿，张卓晟，赵海。2024。环境警示: 多模态智能体易受环境干扰。arXiv 预印本 arXiv:2408.02544。

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.

伊沃琳·农，斯瓦南德·卡德，王浩，穆鲁格桑·基尔蒂拉姆，贾斯汀·D·韦斯兹，阿米特·杜兰达尔，卡尔蒂凯扬·纳特桑·拉马穆尔蒂。2025。保护用户免于自身风险: 保障与对话智能体交互中的上下文隐私。arXiv 预印本 arXiv:2502.18509。

Dang Nguyen, Jian Chen, Yu Wang, Gang Wu, Namy-ong 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 Dernon-court. 2024a. GUI Agents: A Survey. arXiv preprint. ArXiv:2412.13501 [cs].

阮当，陈健，王宇，吴刚，朴南永，胡正勉，吕汉佳，吴俊达，瑞安·阿庞特，夏宇，李新彤，石晶，陈宏杰，赖越达，谢周航，金成哲，张瑞怡，余彤，坦吉姆·梅拉布，内斯林·K·艾哈迈德，普尼特·马图尔，尹承贤，姚琳娜，布拉尼斯拉夫·克韦顿，阮天佑，裴忠，周天一，瑞安·A·罗西，弗兰克·德尔农-库尔。2024a。GUI 智能体: 综述。arXiv 预印本。ArXiv:2412.13501 [计算机科学]。

Dang Nguyen, Viet Dac Lai, Seunghyun Yoon, Ryan A Rossi, Handong Zhao, Ruiyi Zhang, Puneet Mathur, Nedim Lipka, Yu Wang, Trung Bui, et al. 2024b. Dynasaur: Large language agents beyond predefined actions. arXiv preprint arXiv:2411.01747.

阮当，赖越达，尹承贤，瑞安·A·罗西，赵汉东，张瑞怡，普尼特·马图尔，内迪姆·利普卡，王宇，裴忠，等。2024b。Dynasaur: 超越预定义动作的大型语言智能体。arXiv 预印本 arXiv:2411.01747。

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

阮文巴赫，约尔格·施勒特勒，克里斯汀·塞弗特。2024c。Xagent: 利用大型语言模型力量的对话式可解释人工智能智能体。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.

农松勤，朱佳丽，吴锐，金炯超，单硕，黄秀天，徐文浩。2024。Mo-bileflow: 面向移动 GUI 智能体的多模态大型语言模型。arXiv 预印本 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.

法哈德·努拉拉扎德，张毅，乔纳森·弗斯特，库尔特·斯托金格。2024。通过大型语言模型智能体实现自然语言中的可解释多模态数据探索。arXiv 预印本 arXiv:2412.18428。

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

OpenAI。2025。计算机使用智能体。访问日期:2025 年 3 月 16 日。

Mansi Phute, Alec Helbling, Matthew Hull, Sheng Yun Peng, Sebastian Szyller, Cory Cornelius, and Duen Horng Chau. 2023. Llm Self Defense: By Self Examination, LLMs Know They Are Being Tricked. Tiny Papers @ ICLR.

曼西·普特，亚历克·赫布林，马修·赫尔，彭胜云，塞巴斯蒂安·斯齐勒，科里·科尼利厄斯，周敦宏。2023。大型语言模型自我防御: 通过自我审视，LLM 知道自己正在被欺骗。ICLR Tiny Papers。

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.

乔治奥·皮亚蒂，金志晶，马克斯·克莱曼-韦纳，伯恩哈德·舍尔科普夫，姆林玛雅·萨昌，拉达·米哈尔恰。2024。合作还是崩溃: 大型语言模型智能体社会中可持续合作的出现。神经信息处理系统进展，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.

邱浩毅，A. R. 法布里，迪维扬什·阿加瓦尔，黄孔祥，谭莎拉，彭楠云，吴建胜。2024。评估大型语言模型网络智能体的文化与社会意识。arXiv 预印本 arXiv:2410.23252。

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).

雷沙布·K·夏尔马，维纳亚克·古普塔，丹·格罗斯曼。2024。通过用户提供的规范防御基于图像的提示攻击语言模型。2024 年 IEEE 安全与隐私研讨会 (SPW)。

Huawen Shen, Chang Liu, Gengluo Li, Xinlong Wang, Yu Zhou, Can Ma, and Xiangyang Ji. 2024. Falcon-ui: Understanding gui before following user instructions. arXiv preprint arXiv:2412.09362.

沈华文，刘畅，李更洛，王新龙，周宇，马灿，季向阳。2024。Falcon-ui: 在执行用户指令前理解 GUI。arXiv 预印本 arXiv:2412.09362。

Tianlin Shi, Andrej Karpathy, Linxi Fan, Jonathan Hernandez, and Percy Liang. 2017. World of bits: An open-domain platform for web-based agents. In Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research, pages 3135-3144. PMLR.

石天林，Andrej Karpathy，范林曦，Jonathan Hernandez，和梁珀西。2017 年。World of bits: 一个面向开放域的基于网络的智能体平台。载于第 34 届国际机器学习大会论文集，机器学习研究论文集第 70 卷，页 3135-3144。PMLR。

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.

石宇成，杜梦楠，吴轩升，关子涵，孙晋，和刘宁浩。2023a。通过零样本图像净化实现黑盒后门防御。神经信息处理系统进展，36:57336-57366。

Yucheng Shi, Quanzheng Li, Jin Sun, Xiang Li, and Ninghao Liu. 2025. Enhancing cognition and explainability of multimodal foundation models with self-synthesized data. In The Thirteenth International Conference on Learning Representations.

石宇成，李全正，孙晋，李翔，和刘宁浩。2025。利用自合成数据增强多模态基础模型的认知能力和可解释性。载于第十三届国际表征学习会议。

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.

石哲伦，王志品，范鸿星，尹振飞，盛璐，乔宇，和邵靖。2023b。Chef: 一个用于多模态大型语言模型标准化评估的综合框架。arXiv 预印本 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.

李思妍，Vethavikashini Chithrra Raghuram，Omar Khattab，Julia Hirschberg，和周瑜。2024。Papillon: 基于互联网和本地语言模型集成的隐私保护。arXiv 预印本 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.

Tobin South, Samuele Marro, Thomas Hardjono, Robert Mahari, Cedric Deslandes Whitney, Dazza Greenwood, Alan Chan, 和 Alex Pentland。2025。认证委托与授权 AI 代理。arXiv 预印本 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.

Sakhinana Sagar Srinivas, Geethan Sannidhi, 和 Venkataramana Runkana. 2024. 面向人类水平理解复杂工艺工程示意图: 一种用于开放域问答的教学性、自省性多智能体框架。arXiv 预印本 arXiv:2409.00082.

Zinovia Stefanidi, George Margetis, Stavroula Ntoa, and George Papagiannakis. 2022. Real-time adaptation of context-aware intelligent user interfaces, for enhanced situational awareness. IEEE Access, 10:23367-23393.

Zinovia Stefanidi, George Margetis, Stavroula Ntoa, 和 George Papagiannakis. 2022. 面向增强情境感知的智能用户界面实时自适应。IEEE Access, 10:23367-23393.

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.

Shilong Wang, Guibin Zhang, Miao Yu, Guancheng Wan, Fanci Meng, Chongye Guo, Kun Wang, 和 Yang Wang. 2025. G-safeguard: 基于拓扑引导的安全透镜及对基于大语言模型 (LLM) 的多智能体系统的防护。arXiv 预印本 arXiv:2502.11127.

Shuai Wang, Weiwen Liu, Jingxuan Chen, Yuqi Zhou, Weinan Gan, Xingshan Zeng, Yuhan Che, Shuai Yu, Xinlong Hao, Kun Shao, et al. 2024a. Gui agents with foundation models: A comprehensive survey. arXiv preprint arXiv:2411.04890.

Shuai Wang, Weiwen Liu, Jingxuan Chen, Yuqi Zhou, Weinan Gan, Xingshan Zeng, Yuhan Che, Shuai Yu, Xinlong Hao, Kun Shao, 等. 2024a. 基于基础模型的 GUI 智能体: 一项综合性综述。arXiv 预印本 arXiv:2411.04890.

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

Siyi Wang, Sinan Wang, Yujia Fan, Xiaolei Li, 和 Yepang Liu. 2024b. 利用大型视觉-语言模型提升自动化网页 GUI 测试效果。IEEE 软件维护与演进国际会议。

Xiaoqiang Wang and Bang Liu. 2024. Oscar: Operating system control via state-aware reasoning and re-planning. arXiv preprint arXiv:2410.18963.

Xiaoqiang Wang 和 Bang Liu. 2024. Oscar: 通过状态感知推理与重新规划实现操作系统控制。arXiv 预印本 arXiv:2410.18963.

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

Yuntao Wang, Yanghe Pan, Quan Zhao, Yi Deng, Zhou Su, Linkang Du, 和 Tom H Luan. 2024c. 大型模型智能体: 最新进展、协作范式、安全与隐私及未来趋势。arXiv 预印本 arXiv:2409.14457.

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.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, 等. 2022. 语言模型风险分类法。载于 2022 年 ACM 公平性、问责制与透明度会议论文集，页 214-229。

Hao Wen, Yuanchun Li, Guohong Liu, Shanhui Zhao, Tao Yu, Toby Jia-Jun Li, Shiqi Jiang, Yunhao Liu, Yaqin Zhang, and Yunxin Liu. 2024. Autodroid: Llm-powered task automation in android. In Proceedings of the 30th Annual International Conference on Mobile Computing and Networking, pages 543-557.

Hao Wen, Yuanchun Li, Guohong Liu, Shanhui Zhao, Tao Yu, Toby Jia-Jun Li, Shiqi Jiang, Yunhao Liu, Yaqin Zhang, 和 Yunxin Liu. 2024. Autodroid: 基于大语言模型的安卓任务自动化。载于第 30 届国际移动计算与网络会议论文集，页 543-557。

Jesse Wright. 2024. Here's charlie! realising the semantic web vision of agents in the age of llms. International Workshop on the Semantic Web.

Jesse Wright. 2024. 这是 Charlie！在大语言模型时代实现语义网智能体愿景。语义网国际研讨会。

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

Chen Henry Wu, Rishi Rajesh Shah, Jing Yu Koh, Russ Salakhutdinov, Daniel Fried, 和 Aditi Raghunathan. 2025a. 多模态图像智能体对抗鲁棒性的剖析。载于第十三届国际学习表征会议。

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

Fangzhou Wu, Shutong Wu, Yulong Cao, 和 Chaowei Xiao. 2024a. Wipi: 一种针对基于大语言模型驱动的网页智能体的新型网络威胁。arXiv 预印本 arXiv:2402.16965.

Xuansheng Wu, Jiayi Yuan, Wenlin Yao, Xiaoming Zhai, and Ninghao Liu. 2025b. Interpreting and steering llms with mutual information-based explanations on sparse autoencoders. arXiv preprint arXiv:2502.15576.

Xuansheng Wu, Jiayi Yuan, Wenlin Yao, Xiaoming Zhai, 和 Ninghao Liu. 2025b. 基于互信息解释的稀疏自编码器用于大语言模型的解释与引导。arXiv 预印本 arXiv:2502.15576.

Xuansheng Wu, Haiyan Zhao, Yaochen Zhu, Yucheng Shi, Fan Yang, Tianming Liu, Xiaoming Zhai, Wenlin Yao, Jundong Li, Mengnan Du, et al. 2024b. Usable xai: 10 strategies towards exploiting explainability in the llm era. arXiv preprint arXiv:2403.08946.

Xuansheng Wu, Haiyan Zhao, Yaochen Zhu, Yucheng Shi, Fan Yang, Tianming Liu, Xiaoming Zhai, Wenlin Yao, Jundong Li, Mengnan Du, 等. 2024b. 可用的可解释人工智能: 面向大语言模型时代的 10 种利用解释性的策略。arXiv 预印本 arXiv:2403.08946.

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

Zhiyong Wu, Zhenyu Wu, Fangzhi Xu, Yian Wang, Qiushi Sun, Chengyou Jia, Kanzhi Cheng, Zichen Ding, Liheng Chen, Paul Pu Liang, 等. 2024c. Os-atlas: 面向通用 GUI 智能体的基础动作模型。arXiv 预印本 arXiv:2410.23218.

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.

Zhen Xiang, Linzhi Zheng, Yanjie Li, Junyuan Hong, Qinbin Li, Han Xie, Jiawei Zhang, Zidi Xiong, Chulin Xie, Carl Yang, Dawn Song, 和 Bo Li. 2024. Guardagent: 通过知识驱动推理的守护代理保障大语言模型代理安全。arXiv 预印本 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.

Junlin Xie, Zhihong Chen, Ruifei Zhang, Xiang Wan, 和 Guanbin Li. 2024. 大型多模态代理: 综述。arXiv 预印本 arXiv:2402.15116。

Yueqi Xie, Jingwei Yi, Jiawei Shao, Justin Curl, Lingjuan Lyu, Qifeng Chen, Xing Xie, and Fangzhao Wu.

2023. Defending ChatGPT against jailbreak attack via self-reminders. Nature Machine Intelligence, 5(12):1486-1496.

Yueqi Xie, Jingwei Yi, Jiawei Shao, Justin Curl, Lingjuan Lyu, Qifeng Chen, Xing Xie, 和 Fangzhao Wu. 2023. 通过自我提醒防御 ChatGPT 越狱攻击。自然机器智能 (Nature Machine Intelligence), 5(12):1486-1496。

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

Chejian Xu, Mintong Kang, Jiawei Zhang, Zeyi Liao, Lingbo Mo, Mengqi Yuan, Huan Sun, 和 Bo Li. 2024a. Advweb: 对基于视觉语言模型的网络代理进行可控黑盒攻击。arXiv 预印本 arXiv:2410.17401。

Yuanyuan Xu, Weiting Gao, Yining Wang, Xinyang Shan, and Yin-Shan Lin. 2024b. Enhancing user experience and trust in advanced llm-based conversational agents. Computing and Artificial Intelligence, 2(2).

Yuanyuan Xu, Weiting Gao, Yining Wang, Xinyang Shan, 和 Yin-Shan Lin. 2024b. 提升基于先进大语言模型的对话代理的用户体验与信任。计算与人工智能 (Computing and Artificial Intelligence), 2(2)。

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.

Yulong Yang, Xinshan Yang, Shuaidong Li, Chenhao Lin, Zhengyu Zhao, Chao Shen, 和 Tianwei Zhang. 2024. 移动设备多模态代理的安全矩阵: 系统性研究与概念验证。arXiv 预印本 arXiv:2407.09295。

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2022. Webshop: Towards scalable real-world web interaction with grounded language agents. Advances in Neural Information Processing Systems, 35:20744-20757.

Shunyu Yao, Howard Chen, John Yang, 和 Karthik Narasimhan. 2022. Webshop: 面向可扩展现实网络交互的基于基础语言代理。神经信息处理系统进展 (Advances in Neural Information Processing Systems), 35:20744-20757。

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.

Chung-En (Johnny) Yu, Brian Jalaian, 和 Nathaniel D. Bastian. 2024. 通过多代理系统事后缓解大型视觉语言模型的幻觉问题。AAAI 研讨会论文集 (Proceedings of the AAAI Symposium Series), 4(1):110-113。

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

Qiusi Zhan, Zhixiang Liang, Zifan Ying, 和 Daniel Kang. 2024. Injecagent: 工具集成大语言模型代理中间接提示注入的基准测试。arXiv 预印本 arXiv:2403.02691。

Chaoyun Zhang, Shilin He, Jiaxu Qian, Bowen Li, Liqun Li, Si Qin, Yu Kang, Minghua Ma, Qingwei Lin, Saravan Rajmohan, et al. 2024a. Large language model-brained gui agents: A survey. arXiv preprint arXiv:2411.18279.

Chaoyun Zhang, Shilin He, Jiaxu Qian, Bowen Li, Liqun Li, Si Qin, Yu Kang, Minghua Ma, Qingwei Lin, Saravan Rajmohan, 等. 2024a. 大语言模型驱动的图形用户界面代理: 综述。arXiv 预印本 arXiv:2411.18279。

Xinyu Zhang, Huiyu Xu, Zhongjie Ba, Zhibo Wang, Yuan Hong, Jian Liu, Zhan Qin, and Kui Ren. 2024b. Privacyasst: Safeguarding user privacy in tool-using large language model agents. IEEE Transactions on Dependable and Secure Computing.

Xinyu Zhang, Huiyu Xu, Zhongjie Ba, Zhibo Wang, Yuan Hong, Jian Liu, Zhan Qin, 和 Kui Ren. 2024b. Privacyasst: 保障使用工具的大语言模型代理中的用户隐私。IEEE 可靠与安全计算汇刊 (IEEE Transactions on Dependable and Secure Computing)。

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

Zhexin Zhang, Shiyao Cui, Yida Lu, Jingzhuo Zhou, Junxiao Yang, Hongning Wang, 和 Minlie Huang. 2024c. Agent-safetybench: 评估大语言模型代理的安全性。arXiv 预印本 arXiv:2412.14470。

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

Zhizheng Zhang, Xiaoyi Zhang, Wenxuan Xie, 和 Yan Lu. 2023. 负责任的任务自动化: 赋能大语言模型成为负责任的任务自动执行者。arXiv 预印本 arXiv:2306.01242。

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.

Kangjia Zhao, Jiahui Song, Leigang Sha, HaoZhan Shen, Zhi Chen, Tiancheng Zhao, Xiubo Liang, 和 Jianwei Yin. 2024. Gui 测试竞技场: 推进自主 GUI 测试代理的统一基准。arXiv 预印本 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, Boyu Gou, Jihyung Kil, Huan Sun, 和 Yu Su. 2024. GPT-4V(ision) 是通用的网络代理, 前提是有落地支持。国际机器学习大会 (International Conference on Machine Learning)。

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.

钟伟宏，冯晓成，赵亮，李启明，黄磊，顾宇轩，马伟涛，徐元，秦兵。2024 年。研究与缓解大型视觉语言模型中的多模态幻觉雪球效应。计算语言学协会年会。

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

周凯青，刘成志，赵轩东，安德森·孔帕拉斯，宋 Dawn，王新 Eric。2024 年。多模态情境安全。arXiv 预印本 arXiv:2410.06172。

Naitian Zhou, David Bamman, and Isaac L Bleaman. 2025. Culture is not trivia: Sociocultural theory for cultural nlp. arXiv preprint arXiv:2502.12057.

周乃天，大卫·班曼，艾萨克·L·布利曼。2025 年。文化非琐事: 面向文化自然语言处理的社会文化理论。arXiv 预印本 arXiv:2502.12057。

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

周淑妍, 徐弗兰克, 朱浩, 周旭辉, 罗伯特·洛, 阿比谢克·斯里达尔, 程贤义, 欧天岳, 约纳坦·比斯克, 丹尼尔·弗里德等。2023 年。We-barena: 构建自主代理的真实网络环境。arXiv 预印本 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.

朱子辰，唐浩，李燕思，刘定业，徐洪深，兰坤尧，张丹阳，姜一轩，周浩，王晨润，张思拓，孙良泰，王一笑，孙宇恒，陈璐，余凯。2025 年。Moba: 多面向记忆增强的自适应规划，用于高效移动任务自动化。预印本，arXiv:2410.13757。

Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, Rowan Wang, Zico Kolter, Matt Fredrikson, and Dan Hendrycks. 2024. Improving Alignment and Robustness with Circuit Breakers. arXiv.org.

邹安迪，潘龙，王贾斯汀，杜瑞克·杜纳斯，林马克斯韦尔，马克西姆·安德鲁申科，王罗文，齐科·科尔特，马特·弗雷德里克森，丹·亨德里克斯。2024 年。利用断路器提升对齐性与鲁棒性。arXiv.org。