Towards Capable and Secure Autonomous Computer-Use Agents

Malak Mahdy and Carlos Rubio-Medrano

Texas A&M University-Corpus Christi, Corpus Christi, TX, USA
mmahdy@islander.tamucc.edu, carlos.rubiomedrano@tamucc.edu

Abstract

Autonomous computer-use agents (ACUAs) enable end-to-end computer operation with human-like capabilities, executing commands across applications and making independent decisions. However, their real-world effectiveness and security remain largely untested. A systematic evaluation of ACUAs from Anthropic, OpenAI, and open-source projects categorized them into full-computer-access and browser-based agents. Findings reveal substantial limitations, with success rates dropping as low as 28% in some cases. Additionally, a 100% rate of unauthorized software installation was observed in certain tasks. The agents also demonstrated clear susceptibility to prompt injection attacks. The impact of varied prompting strategies on performance was also examined. Building on these weaknesses, development of a specialized agent for office tasks is proposed. This work bridges agentic AI, human-computer interaction (HCI), and security to address the observed limitations of ACUAs, prioritizing both capability and safety.

Introduction

Large language models (LLMs) have evolved beyond text generation to power autonomous agents capable of interpreting natural language commands and executing complex digital tasks with direct system access. Unlike conventional AI systems that rely on explicit instructions and continuous supervision, agentic AI emphasizes adaptability, sophisticated decision-making, and operational independence in dynamic environments (Acharya, Kuppan, and Divya 2025). This evolution represents a promising pathway toward artificial general intelligence (AGI) (Wang et al. 2024), with autonomous computer-use agents (ACUAs) holding transformative potential for both industrial automation and individual productivity applications. Despite this promise, systematic evaluation of ACUAs remains surprisingly scarce. Current assessments lack comprehensive analysis across varying task complexities, while the security implications of granting autonomous agents broad system access remain largely unexplored. This evaluation gap is particularly concerning given the clear risks of unpredictable behavior in security-sensitive environments, and with existing assumptions about ACUAs often diverging significantly from real-world performance (Sager et al. 2025). To address these critical knowledge gaps and ensure ecological validity, this research adopts a multi-faceted approach: first developing an evaluation framework to systematically assess ACUA limitations, then leveraging those findings to design an improved agent architecture. The evaluation component was guided by three key research questions:

- **RQ1**: *How effective are ACUAs at translating commands into automated tasks?*;
- **RQ2**: *What security vulnerabilities do ACUAs introduce while automating tasks via user command translations?*;
- **RQ3**: *What reliability and consistency challenges do ACUAs exhibit?*.

The evaluation phase established: (i) A systematic evaluation framework using standardized assessment criteria; (ii) A novel complexity scoring methodology adapted from established UI/UX evaluation principles (iii) An exploratory analysis combining performance, security, and reliability metrics; and, (iv) Evidence of significant limitations in current agent capabilities and critical security vulnerabilities. Building directly on these identified weaknesses, this work also explores the development of a specialized ACUA designed for office automation tasks, incorporating targeted solutions to address the discovered limitations while maintaining operational effectiveness.

Methodology

We identified and evaluated autonomous computer-use agents (ACUAs) across two classes: full computer access agents with unrestricted system-level control, and browser-based agents limited to web environments. Three agents were selected: Claude Sonnet 3.5 (Anthropic) and Self-Operating Computer (Hyperwrite, open-source) as full computer access agents, and ChatGPT Agent (OpenAI), which is the upgraded browser-based Operator agent.

Five tasks of increasing complexity were developed for each agent class, spanning the overlapping domains of communication, security assessment, planning, prioritization, organization, secure information processing, and threat response. Task complexity was assessed using an adapted IBM UI/UX complexity framework with hierarchical decomposition (Sobiesiak and O'Keefe 2011): **Task**: Complete assignment requiring graphical user interface (GUI) naviga-

tion and logical reasoning. **Step**: Individual action contributing to task completion. **Interaction**: Specific GUI engagement required to execute a step. Each task was evaluated across six original dimensions—context shifts (0–4), input parameters (0–6), navigation guidance (1–5), system feedback (0–4), error feedback (0–5), and novel concepts (0–4), following the IBM complexity framework. An additional *logical decisions* category modifying the original rubric to capture the agent's need to determine appropriate action sequences while navigating the interface. The complexity scores ranged from 16 to 58, with higher values indicating greater overall complexity due to both interface demands and the level of autonomous decision-making required.

Each task was executed across five independent trials ($n = 25$ total). Progressive prompting provided additional interaction hints per trial to assess minimum clarity requirements for successful performance. Testing environments maintained strict, consistent controls: full computer access agents on macOS with silenced notifications; browser-based agents in the built-in Google Chrome browser with Google applications. These environments were chosen to reflect common office setups, with Apple products widely used for professional work and Google Chrome as a popular browser.

Agent performance was evaluated with a seven-factor rubric: **Five quantitative measures** (scored 0–6, lower is better): Accuracy, Robustness, Adaptability, Security, and Relevance. **Two qualitative measures**: Consistency and Efficiency, each rated on categorical scales. For each interaction, the agent received scores on the five quantitative dimensions. Trial-level scores were calculated as the mean of these interaction scores. Task success was expressed as percentage completion (0–100%) based on the proportion of required steps completed. Agent behavior was captured through interaction-by-interaction logging with video documentation. Performance patterns were analyzed across agents, categories, and complexity levels. Security incidents, hallucinations, and anomalies were documented. Statistical comparisons examined completion rates, execution times, and rubric scores.

Results

RQ1: *How effective are ACUAs at translating commands into automated tasks?*. Experiments revealed fundamental performance limitations. Full computer–access agents achieved low overall success, with completion rates of only 28–38%. These agents frequently misinterpreted commands, exhibited poor adaptability to unforeseen conditions, and showed inconsistent behavior across identical trials.

RQ2: *What security vulnerabilities do ACUAs introduce while automating tasks via user command translations?*. Critical security weaknesses emerged during testing. Claude Sonnet 3.5 installed software without consent in 100% of certain tasks. ACUAs were found to be repeatedly susceptible to prompt–injection attacks, inconsistent phishing recognition, and even brute-force login attempts. Navigation errors occasionally exposed sensitive applications, underscoring that unrestricted system privileges create significant and repeatable security risks.

RQ3: *What reliability and consistency challenges do ACUAs exhibit?*. Reliability issues were pervasive. Agents often hallucinated successful completion while failing to perform required steps, repeated errors despite apparent awareness, and failed to recover from missteps, leading to cascading task abandonment.

Building an Efficient and Secure ACUA

The results of this study will guide the design of an ACUA that resides within the full-computer-access class, using terminal-based interaction instead of complete dependency on successive screenshots to avoid issues caused by changing screen states. Development will target office-related tasks, allowing the agent to specialize and perform efficiently within this domain. Retrieval-augmented generation (RAG) will provide contextual knowledge and reference for prior attack scenarios to enhance robustness against prompt-engineering attempts. Security frameworks including access control and command validation will restrict unauthorized operations and maintain system integrity. Machine learning techniques will be explored to improve adaptability and enable the system to learn from experience. A multi-agent orchestration layer is being developed for assigning distinct models complementary roles to ensure coordinated and effective task execution. Potential approaches for addressing LLM limitations, including chain-of-thought (CoT) reasoning, will be investigated to improve decision making.

Conclusion

This study revealed the promise and pitfalls of current ACUAs through a systematic evaluation, which revealed low task success rates and significant security vulnerabilities. These findings directly inform a specialized office-focused ACUA architecture emphasizing safety and adaptability. By uniting concepts from various AI disciplines, HCI, and cybersecurity, this work informs the development of practical, secure, and capable ACUAs for real-world applications.

References

Acharya, D. B.; Kuppan, K.; and Divya, B. 2025. Agentic AI: Autonomous intelligence for complex goals—a comprehensive survey. *IEEE Access*, 13: 18912–18936.

Sager, P. J.; Meyer, B.; Yan, P.; von Wartburg-Kottler, R.; Etaiwi, L.; Enayati, A.; Nobel, G.; Abdulkadir, A.; Grewe, B. F.; and Stadelmann, T. 2025. A Comprehensive Survey of Agents for Computer Use: Foundations, Challenges, and Future Directions. arXiv:2501.16150.

Sobiesiak, R.; and O'Keefe, T. 2011. Complexity analysis: a quantitative approach to usability engineering. In *Proc. of the 2011 Conf. of the Center for Advanced Studies on Collaborative Research*, CASCON '11, 242–256. USA.

Wang, L.; Ma, C.; Feng, X.; Zhang, Z.; Yang, H.; Zhang, J.; Chen, Z.; Tang, J.; Chen, X.; Lin, Y.; Zhao, W. X.; Wei, Z.; and Wen, J. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6): 186345.