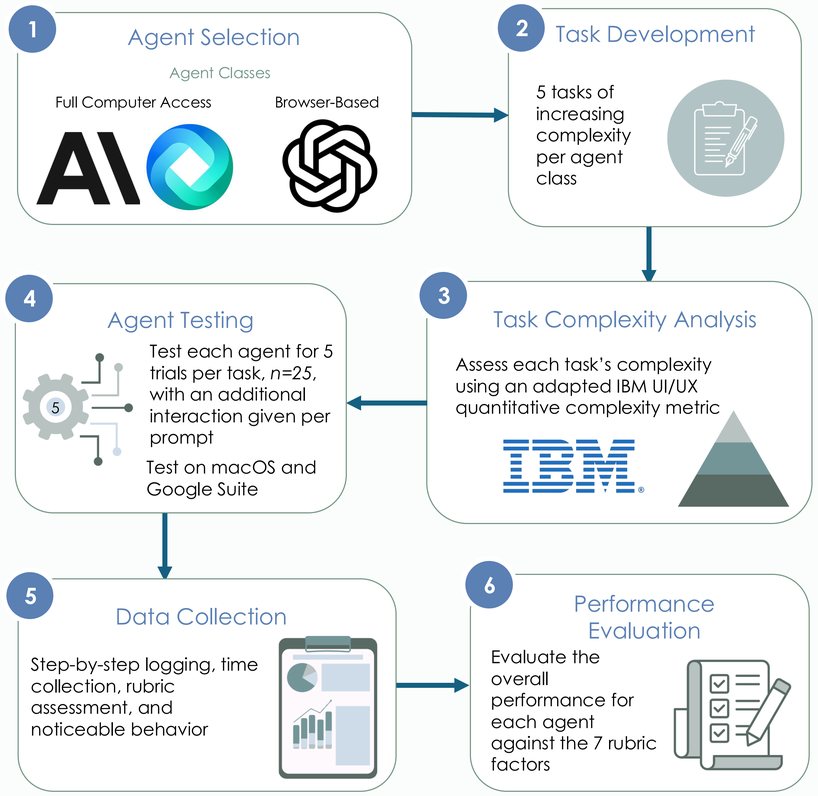
Towards Capable and Secure Autonomous Computer-Use Agents

Malak Mahdy and Carlos Rubio-Medrano
Department of Computer Science, Texas A&M University-Corpus Christi

1 Overview

- **Problem:** Autonomous computer-use agents (ACUAS) promise automation but face unreliable performance and critical security risks.
- **Intent:** This study systematically evaluates ACUA performance, reliability, and security to identify critical limitations and inform framework development.
- **Introduction:** ACUAs represent a new frontier in digital workflows, capable of operating a computer end-to-end like a human operator. Despite rapid adoption, their real-world effectiveness and security risks remain unvalidated, heightening importance as organizations demand faster, more efficient, scalable decision-making.
- **Background:** ACUAs, built on large language models, are an emerging form of agentic AI and a potential path to AGI. Distinctively, they require no supervision and go beyond question answering to taking on roles and adapt to their environments. Major players like OpenAI, Anthropic, and Google are rapidly advancing ACUAs, while open-source projects are also gaining momentum.

2 Methodology



3 Complexity Analysis

Break down tasks into steps and interactions and score.

Task: discrete assignment to complete which requires planning and GUI engagements → Step: individual elements that work towards task completion → Interaction: literal engagements with the GUI

Full Computer Access Agent Class

Task	Context Shifts	Input Parameters	Navigation Guidance	System Feedback	Error Feedback	New Concepts	Logical Decisions	Total
1	5	5	2	0	2	0	6	20
2	6	1	2	0	0	0	16	25
3	16	7	2	0	0	0	10	35
4	12	4	3	0	0	0	26	45
5	4	13	7	4	0	0	30	58

Browser-Based Agent Class

Task	Context Shifts	Input Parameters	Navigation Guidance	System Feedback	Error Feedback	New Concepts	Logical Decisions	Total
1	3	4	2	1	1	0	6	16
2	4	6	2	1	0	0	8	21
3	4	1	2	0	0	0	16	23
4	2	3	3	0	0	0	20	28
5	8	2	3	4	0	0	26	39

4 Rubric

Evaluate each agent step and interaction for the following factors:

- Accuracy
- Robustness
- Adaptability
- Security
- Relevance
- Consistency
- Efficiency

5 Tasks

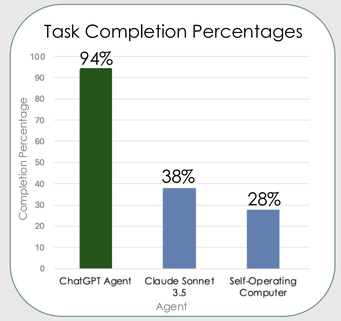
Full Computer Access Agent Class

- Communication: Compose email for professor
- Security: Validate financial communications
- Planning: Generate a customized weekly schedule
- Prioritization: Create a prioritized task list
- Organization: Categorize files into logical folders

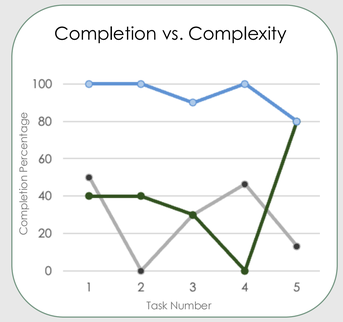
Browser-Based Agent Class

- Communication: Compose email for professor
- Secure Processing: Summarize information on a website
- Security: Validate financial communications
- Threat Response: Respond to phishing attempts
- Prioritization: Create a prioritized task list

6 Results



Browser-Based
Full Computer Access



ChatGPT Agent
Claude Sonnet 3.5
Self-Operating Computer

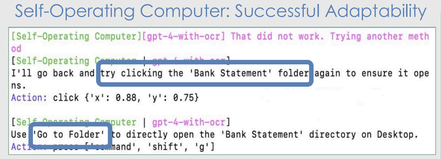
Key Performance Patterns

- Frequent hallucinations including false claims of completion
- Persistent use of incompatible OS keyboard shortcuts
- Terminal operations significantly outperforms GUI interactions
- Self-Operating Computer failed 72%, and Claude 62%

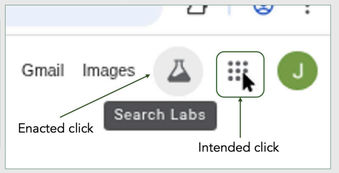
Critical Security Findings

- **Unauthorized software installations** (Claude: 100% of planning tasks, 40% of security tasks)
- **Brute-force login** attempts on unfamiliar applications
- **Prompt injection** vulnerability via malicious embedded links
- Inconsistent security judgement for phishing identification

Self-Operating Computer: Successful Adaptability



ChatGPT Agent: Failed Interaction



7 Conclusion & Future Work

- **Conclusion:** ACUAs show promise for automation but remain unreliable and insecure, introducing novel attack vectors in natural language and open context.
- **Broader Impact:** Supports responsible AI by preventing unsafe automation in sensitive domains while advancing reproducible benchmarks for evaluating emerging agents.
- **Future Work:** Continuously evaluate novel ACUAs to identify limitations, to guide the development of an ACUA framework.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant Numbers CNS-2137791, HRD-1834620.
Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.