Novelty Assessment Report

Paper: OpenApps: Simulating Environment Variations to Measure UI Agent Reliability
PDF URL: https://openreview.net/pdf?id=cj1MAx7lKs
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-01

Abstract

Reliability is key to realizing the promise of autonomous UI-agents, multimodal agents that directly interact with the apps humans use, as users must be able to trust an agent to complete a given task. Current evaluations rely on fixed environments---often clones of existing apps--- which are limited in that they can only shed light on whether or how often an agent can complete a task within a specific environment. When deployed however, agents are likely to encounter variations in app design and content that can affect an agent's ability to complete a task. To address this blind spot of measuring agent reliability across app variations, we develop OpenApps, a light-weight open-source ecosystem with six apps (messenger, calendar, maps, etc.) that are configurable in appearance and content. OpenApps requires just a single CPU to run, enabling easy generation and deployment of thousands of versions of each app. Specifically, we run more than 10,000 independent evaluations to study reliability across seven leading multimodal agents. We find that while standard reliability within a fixed app is relatively stable, reliability can vary drastically when measured across app variations. Task success rates for many agents can fluctuate by more than 50\% across app variations. For example, Kimi-VL-3B's average success across all tasks fluctuates from 63\% to just 4\% across app versions. We also find agent behaviors such as looping or hallucinating actions can differ drastically depending on the environment configuration. These initial findings highlight the importance of measuring reliability along this new dimension of app variations.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **measuring UI agent reliability across environment variations**

A total of **30 papers** were analyzed and organized into a taxonomy with **15 categories**.

Taxonomy Overview

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

- **Agent Robustness to Interface and Environmental Variations**
- **Test Automation Robustness and Reliability**
- **Agent Architectures and Task Automation Systems**
- **Domain-Specific Agent Benchmarks and Applications**
- **Formal Methods and Model-Based Approaches for UI Reliability**
- **Simulation and Modeling for Agent Evaluation**

Complete Taxonomy Tree

- measuring UI agent reliability across environment variations Survey Taxonomy
- Agent Robustness to Interface and Environmental Variations
  - Robustness to GUI Anomalies and Real-World Interface Variations ★ (4 papers)
  - [0] OpenApps: Simulating Environment Variations to Measure UI Agent Reliability (Anon et al., 2026) View paper
  - [2] Gui agents: A survey (Dang Nguyen, 2025) View paper
  - [3] Towards trustworthy gui agents: A survey (Shi Yucheng, 2025) View paper
  - [6] GUI-Robust: A Comprehensive Dataset for Testing GUI Agent Robustness in Real-World Anomalies (Yang Jing-qi, 2025) View paper
  - Security and Adversarial Robustness in Dynamic Environments (1 papers)
  - [8] GhostEI-Bench: Do Mobile Agents Resilience to Environmental Injection in Dynamic On-Device Environments? (Chen ChiYu, 2025) View paper
  - Operational Reliability and Enterprise Deployment Readiness (1 papers)
  - [11] UI-CUBE: Enterprise-Grade Computer Use Agent Benchmarking Beyond Task Accuracy to Operational Reliability (Horia Cristescu, 2025) View paper
- Test Automation Robustness and Reliability
  - Locator Robustness and Template-Based Testing (2 papers)
  - [1] Investigating the robustness of locators in template-based Web application testing using a GUI change classification model (Marco De Luca, 2024) View paper
  - [7] Feature matching-based approaches to improve the robustness of Android visual GUI testing (Luca Ardito, 2021) View paper
  - Machine Learning for Robust Test Synthesis (1 papers)
  - [13] AppFlow: using machine learning to synthesize robust, reusable UI tests (Gang Hu, 2018) View paper
  - Cross-Platform and Multilingual Test Automation (2 papers)
  - [14] Computer Vision for UI Testing: Leveraging Image Recognition and AI to Validate Elements and Layouts (Himanshu Pathak, 2025) View paper
  - [22] Streamlining Multilingual UI Test Automation with Dynamic Localization (Gupta, 2024) View paper
  - Domain-Specific Test Automation Systems (1 papers)

    - [12] The Ultimate Swiss Army Knife of UI Software Testing, Accelerating Workflow-Based Automation for Medical Imaging Systems (Hui, 2025) View paper
- Agent Architectures and Task Automation Systems
    - Multi-Agent and Context-Aware Frameworks (7 papers)
    - [9] ReInAgent: A Context-Aware GUI Agent Enabling Human-in-the-Loop Mobile Task Navigation (Jia Hai-tao, 2025) View paper
    - [19] UFO2: The Desktop AgentOS (Zhang, 2025) View paper
    - [23] Integrated Energy Multi-Agent Collaborative Optimization Regulation Method for Typhoon Events (Z Dou, 2025) View paper
    - [24] Agent-based distributed architecture for mobile robot control (Juan-Luis Posadas-YagÃ¼e, 2008) View paper
    - [26] Evaluation of humanâ□agent user interfaces in multi-agent systems (C. Nam, 2009) View paper
    - [27] Supporting a Comprehensive Knowledge Management System with Agent Technology (Guo, 2001) View paper
    - [29] Agents and Sensors System for Monitoring Sandstorms (EM Shakshuki, 2013) View paper
    - Programming by Demonstration for Task Automation (1 papers)
    - [17] VASTA: a vision and language-assisted smartphone task automation system (Alborz Rezazadeh Sereshkeh, 2020) View paper
    - Natural Language-Driven UI Generation and Rendering (2 papers)
    - [10] Portal UX Agent -- A Plug-and-Play Engine for Rendering UIs from Natural Language Specifications (Li Xinsong, 2025) View paper
    - [15] Context-Aware Visual Prompting: Automating Geospatial Web Dashboards with Large Language Models and Agent Self-Validation for Decision Support (Haowen Xu, 2025) View paper
- Domain-Specific Agent Benchmarks and Applications
    - Automotive and Safety-Critical Interface Agents (1 papers)
    - [18] Automotive-ENV: Benchmarking Multimodal Agents in Vehicle Interface Systems (Yan Jun-Feng, 2025) View paper
    - Agents as Judges for Generative UI Design (1 papers)
    - [20] Computer-Use Agents as Judges for Generative User Interface (Kevin Qinghong Lin, 2025) View paper
    - Interface Performance Under Environmental Conditions (3 papers)
    - [4] Experimental Study on Interface Bonding Performance of Frost-Damaged Concrete Reinforced with Yellow River Sedimentary Sand Engineered Cementitious â□¦ (B Tan, 2025) View paper
    - [5] The influence mechanism of polymer types on the performance of interface agents for old wall tiles (Guanji Lyu, 2025) View paper
    - [25] A Universal Size Design Principle for Stretchable Inorganic Electronics to Work Consistently under Different Interface Conditions (Shuang Li, 2022) View paper
- Formal Methods and Model-Based Approaches for UI Reliability (1 papers)
    - [16] ICOs: A model-based user interface description technique dedicated to interactive systems addressing usability, reliability and scalability (David Navarre, 2009) View paper
- Simulation and Modeling for Agent Evaluation (3 papers)
    - [21] A Hybrid ABM-PDE Framework for Real-World Infectious Disease Simulations (Conrad, 2025) View paper
    - [28] Semi-supervised Semantic Segmentation with Directional Context-aware Consistency (Xin Lai, 2021) View paper
    - [30] Faithful Simulation of Userâ□Agentâ□Environment Interactions for Scalable LLM Agent Evaluation (A Kudrinskii, n.d.) View paper

Narrative

Core task: measuring UI agent reliability across environment variations. The field has organized itself around several complementary perspectives. Agent Robustness to Interface and Environmental Variations examines how agents handle GUI anomalies, layout shifts, and real-world interface changes—work such as GUI Agents Survey[2] and Trustworthy GUI Agents[3] surveys these challenges broadly, while studies like Locator Robustness[1] and GUI-Robust[6] drill into specific failure modes. Test Automation Robustness and Reliability focuses on traditional software testing concerns, including locator stability and cross-platform consistency. Agent Architectures and Task Automation Systems explores the design of end-to-end systems that orchestrate perception, planning, and action, often integrating multimodal models. Domain-Specific Agent Benchmarks and Applications tailors evaluation to particular environments such as mobile apps, web browsers, or desktop software. Formal Methods and Model-Based Approaches for UI Reliability brings verification techniques and symbolic reasoning to ensure correctness guarantees. Finally, Simulation and Modeling for Agent Evaluation provides controlled testbeds where environmental parameters can be systematically varied.

Within the robustness branch, a particularly active line of work investigates how agents degrade under realistic interface perturbations—missing elements, dynamic content, or visual noise—and whether current architectures can generalize beyond clean benchmarks. OpenApps[0] sits squarely in this cluster, emphasizing systematic measurement of reliability when GUI conditions shift. It shares thematic ground with Trustworthy GUI Agents[3], which also prioritizes robustness and safety properties, and with GUI-Robust[6], which targets adversarial or noisy interface scenarios. Compared to these neighbors, OpenApps[0] appears to focus more explicitly on quantifying degradation across a spectrum of environmental variations rather than proposing a single hardening technique. This positioning reflects a broader trend: as agent capabilities improve, the community is moving from proof-of-concept demonstrations toward rigorous stress-testing and reliability engineering, ensuring that deployed systems remain dependable when real-world interfaces inevitably deviate from training distributions.

Related Works in Same Category

No comparison data available.

Contributions Analysis

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

Contribution 1: OpenApps: A configurable ecosystem for measuring UI-agent reliability across app variations

Description: The authors introduce OpenApps, a lightweight Python-based environment containing six configurable apps that can generate thousands of versions to measure agent reliability across app variations in appearance and content, rather than only within fixed environments.

This contribution was assessed against **0 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

Contribution 2: Measurement of reliability across app variations as a new dimension

Description: The authors establish a new dimension for evaluating UI-agent reliability by measuring performance fluctuations across different app variations (design, appearance, content), addressing a blind spot in current evaluations that rely on fixed environment clones.

This contribution was assessed against **0 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

Contribution 3: Ground-truth state-based reward function avoiding trajectory imitation and reward hacking

Description: The authors design a deterministic reward function based on complete app state verification that avoids the limitations of human-trajectory rewards and change-based checks, preventing agents from gaming rewards through unintended actions.

This contribution was assessed against **0 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] OpenApps: Simulating Environment Variations to Measure UI Agent Reliability View paper
- [1] Investigating the robustness of locators in template-based Web application testing using a GUI change classification model View paper
- [2] Gui agents: A survey View paper
- [3] Towards trustworthy gui agents: A survey View paper
- [4] Experimental Study on Interface Bonding Performance of Frost-Damaged Concrete Reinforced with Yellow River Sedimentary Sand Engineered Cementitious â View paper
- [5] The influence mechanism of polymer types on the performance of interface agents for old wall tiles View paper
- [6] GUI-Robust: A Comprehensive Dataset for Testing GUI Agent Robustness in Real-World Anomalies View paper
- [7] Feature matching-based approaches to improve the robustness of Android visual GUI testing View paper
- [8] GhostEI-Bench: Do Mobile Agents Resilience to Environmental Injection in Dynamic On-Device Environments? View paper
- [9] ReInAgent: A Context-Aware GUI Agent Enabling Human-in-the-Loop Mobile Task Navigation View paper
- [10] Portal UX Agent -- A Plug-and-Play Engine for Rendering UIs from Natural Language Specifications View paper
- [11] UI-CUBE: Enterprise-Grade Computer Use Agent Benchmarking Beyond Task Accuracy to Operational Reliability View paper
- [12] The Ultimate Swiss Army Knife of UI Software Testing, Accelerating Workflow-Based Automation for Medical Imaging Systems View paper
- [13] AppFlow: using machine learning to synthesize robust, reusable UI tests View paper
- [14] Computer Vision for UI Testing: Leveraging Image Recognition and AI to Validate Elements and Layouts View paper
- [15] Context-Aware Visual Prompting: Automating Geospatial Web Dashboards with Large Language Models and Agent Self-Validation for Decision Support View paper
- [16] ICOs: A model-based user interface description technique dedicated to interactive systems addressing usability, reliability and scalability View paper
- [17] VASTA: a vision and language-assisted smartphone task automation system View paper
- [18] Automotive-ENV: Benchmarking Multimodal Agents in Vehicle Interface Systems View paper
- [19] UFO2: The Desktop AgentOS View paper
- [20] Computer-Use Agents as Judges for Generative User Interface View paper
- [21] A Hybrid ABM-PDE Framework for Real-World Infectious Disease Simulations View paper
- [22] Streamlining Multilingual UI Test Automation with Dynamic Localization View paper
- [23] Integrated Energy Multi-Agent Collaborative Optimization Regulation Method for Typhoon Events View paper
- [24] Agent-based distributed architecture for mobile robot control View paper
- [25] A Universal Size Design Principle for Stretchable Inorganic Electronics to Work Consistently under Different Interface Conditions View paper
- [26] Evaluation of humanâagent user interfaces in multi-agent systems View paper
- [27] Supporting a Comprehensive Knowledge Management System with Agent Technology View paper
- [28] Semi-supervised Semantic Segmentation with Directional Context-aware Consistency View paper
- [29] Agents and Sensors System for Monitoring Sandstorms View paper
- [30] Faithful Simulation of Userâ□Agentâ□Environment Interactions for Scalable LLM Agent Evaluation View paper
- [31] AndroidWorld: A Dynamic Benchmarking Environment for Autonomous Agents View paper
- [32] Appagent v2: Advanced agent for flexible mobile interactions View paper
- [33] WorldGUI: An Interactive Benchmark for Desktop GUI Automation from Any Starting Point View paper
- [34] GUI-Xplore: Empowering Generalizable GUI Agents with One Exploration View paper
- [35] Atomic-to-Compositional Generalization for Mobile Agents with A New Benchmark and Scheduling System View paper
- [36] Mobile-Bench-v2: A More Realistic and Comprehensive Benchmark for VLM-based Mobile Agents View paper
- [37] Learning UI Navigation through Demonstrations composed of Macro Actions View paper
- [38] D-GARA: A Dynamic Benchmarking Framework for GUI Agent Robustness in Real-World Anomalies View paper
- [39] MCPWorld: A Unified Benchmarking Testbed for API, GUI, and Hybrid Computer Use Agents View paper
- [40] Dissecting adversarial robustness of multimodal lm agents View paper
- [41] R2-Bench: Benchmarking the Robustness of Referring Perception Models under Perturbations View paper
- [42] Towards Evaluating the Robustness of Visual State Space Models View paper
- [43] Enhancing the robustness of vision-language foundation models by alignment perturbation View paper
- [44] Simulating a primary visual cortex at the front of CNNs improves robustness to image perturbations View paper
- [45] The colosseum: A benchmark for evaluating generalization for robotic manipulation View paper
- [46] Assessing Model Robustness in Complex Visual Environments View paper
- [47] On the Robustness of GUI Grounding Models Against Image Attacks View paper

- [48] Hydra: An Agentic Reasoning Approach for Enhancing Adversarial Robustness and Mitigating Hallucinations in Vision-Language Models View paper
- [49] Deep reinforcement learning via object-centric attention View paper
- [50] An Autonomous RL Agent Methodology for Dynamic Web UI Testing in a BDD Framework View paper
- [51] Glider: A reinforcement learning approach to extract UI scripts from websites View paper
- [52] Boosting Universal LLM Reward Design through Heuristic Reward Observation Space Evolution View paper
- [53] Deep reinforcement learning in automated user interface testing View paper
- [54] UI-R1: Enhancing Efficient Action Prediction of GUI Agents by Reinforcement Learning View paper
- [55] Reinforced UI Instruction Grounding: Towards a Generic UI Task Automation API View paper