Multimodal GUI Agent Robustness Under Alignment Techniques and Task Length Variations

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How do different alignment techniques affect the robustness of multimodal GUI agents in handling task length variations on the AndroidWorld benchmark. 15 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.4/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: AndroidWorld: A Dynamic Benchmarking Environment for Autonomous Agents. Research question: How do different alignment techniques affect the robustness of multimodal GUI agents in handling task length variations on the AndroidWorld benchmark?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.4/10.

3 Results

14 papers retrieved. 15 claims extracted; 0 independently verified. Quality review score: 4.4/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
The seed was set to 30 and the temperature to 0 to aid reproducibility.	×	0.04
Each task has a maximum allowed number of steps, typically set to twice the number of steps needed by human annotators t	×	0.04
Gemini 1.5 Pro, GPT-4 Turbo, and the open-source Gemma 2 27B were used as base models.	×	0.03
For MobileMiniWoB++, 62 tasks were evaluated, consistent with recent studies.	×	0.04
Table 3 presents the success rates (SR) for the agents and human performance on both task suites.	×	0.04
The best performance is obtained by M3A when using GPT-4.	×	0.05
On ANDROIDWORLD, the SoM-based variant is less performant, while on MobileMiniWoB++ it performs best.	×	0.03
The simplified agent variant M3A-SIMPLE shows a significant performance drop on ANDROIDWORLD tasks (19.8% vs 30.6% with	×	0.06
On MobileMiniWoB++ tasks, M3A-SIMPLE achieves comparable performance (67.7%).	×	0.02
The open-source Gemma model's lower performance (9.5% on ANDROIDWORLD, 45.6% on MobileMiniWoB++) compared to proprietary	×	0.04
Agents have difficulty understanding mobile UIs, often failing to detect visual cues that are essential for task complet	×	0.03
Agents struggle with certain UI patterns and affordances, and when they make reasoning mistakes, they often lack the cap	×	0.03
Agents sometimes struggle with tasks that simply involve confirming system states, e.g., confirming the WiFi is turned o	×	0.02
Agents struggle with grounding, particularly when executing precise interactions, such as manipulating text or operating	×	0.03
ANDROIDWORLD includes 116 Android tasks and extends with web tasks by integrating the MiniWoB++ benchmark.	×	0.10

References

- http://arxiv.org/abs/2507.10610v3
- http://arxiv.org/abs/2405.14573v5
- http://arxiv.org/abs/2508.03700v5