Agents as Auditors: Detecting Malicious UI Flows via Trajectory-Level Analysis

Aryan Kaul[1,*] **Yuhang Chen**[2,*] **Zhuo Zhang**[1] **Tianlong Chen**[2]
[1]Columbia University [2]UNC Chapel Hill
[*]Equal contribution
{ak5638, zz2888}@columbia.edu {yuhang, tianlong}@cs.unc.edu

Keywords: AI agents, mobile security, dark patterns, trajectory analysis, VLM agents, Android

1. INTRODUCTION

Mobile apps commonly employ dark patterns, manipulative UI designs that steer users toward unintended actions. In September 2025, the FTC settled with Amazon for $2.5 billion over deceptive Prime enrollment and cancellation practices [5]. Many of these manipulations are invisible on any single screen. The deception emerges only across a sequence of interactions. A subscription trap may present normal-looking screens individually, yet subscribing requires two taps while canceling demands seven screens of retention offers and confirmations.

Prior automated detectors operate on single screenshots or hand-crafted per-screen features [2, 7]. DECEPTICON [3] and TrickyArena [4] showed that VLM agents are highly susceptible to these patterns, yet the interaction traces agents produce during navigation go unanalyzed for detection.

We observe that a VLM agent navigating an app produces exactly the interaction trace a flow-level detector needs. We propose *agents as auditors*, where the agent explores and a multi-stage pipeline analyzes the resulting trace. The key design is to model each interaction as an intent-response sequence and detect where the app systematically works against user intent.

2. APPROACH

Trace collection. A VLM agent navigates a target app in an Android emulator, executing adversarial audit tasks (*subscribe then cancel*; *deny all permissions*). Each step is logged as a tuple of {screenshot, action, UI hierarchy, timestamp}.

Structured extraction. A VLM extracts a structured feature record from each screenshot (screen type, buttons, prices, permissions, ad properties) without judging malice.

Cross-screen rule evaluation. A deterministic rule engine evaluates 31 signals across six rule families (disruptive ads, forced actions, nagging, permission abuse, subscription traps, system impersonation) over *sequences of screens*. Each signal fires against an explicit threshold, producing a weighted risk score with evidence.

Compound review and triage. A trajectory-level reviewer identifies compound patterns spanning multiple rule families and merges findings into a final verdict with confidence and evidence.

Pairwise flow comparison. When both a subscribe and a cancel trace are available, we compare the two sub-flows on step counts, obstacle distributions, and friction asymmetry, isolating the roach-motel pattern [6].

3. PRELIMINARY RESULTS

We collected 32 interaction traces from 20 Android apps sourced from AndroZoo [1], 16 of which were flagged by VirusTotal. Separately, we tested 10 top-downloaded apps from Google Play, collectively representing over one billion installations, all exhibiting MUwS-relevant behaviors within minutes of use. All 10 have been reported via Google's PHA reporting channel.

Per-screen analysis classified multiple frames as non-violating. Trajectory analysis identified compound violations spanning multiple screens. A flashlight app's "Remove Ads" screen, benign in isolation, was flagged high severity after the detector observed it followed an unsolicited full-screen ad. A phone cleaner showed four forced ads within 35 seconds, then a subscription wall with a pre-selected auto-renewing plan.

4. DISCUSSION & FUTURE WORK

Our rule engine produces correct verdicts when given accurate VLM extractions. A sub-trace selection agent that segments raw traces into meaningful flows (onboarding, subscription, cancellation) is our next step. We plan to scale to a larger corpus and measure agreement with Google's enforcement decisions.

References

[1] K. Allix et al., "AndroZoo: Collecting millions of Android apps for the research community," *MSR*, 2016.

[2] J. Chen et al., "Unveiling the tricks: Automated detection of dark patterns in mobile applications," *UIST*, 2023.

[3] P. Cuvin, H. Zhu, and D. Yang, "DECEPTICON: How dark patterns manipulate web agents," arXiv:2512.22894, 2025.

[4] D. Ersoy et al., "Investigating the impact of dark patterns on LLM-based web agents," arXiv:2510.18113, 2025.

[5] Federal Trade Commission, "FTC secures historic $2.5 billion settlement against Amazon," September 2025.

[6] A. Sheil et al., "Staying at the Roach Motel," *CHI*, 2024.

[7] X. Shi et al., "50 Shades of deceptive patterns," *WWW*, 2025.