SecRespond: AI Agents Spot Security Alerts but Miss Silent Intrusions

Kabui, Charles

2026-08-01

Read at ToKnow.ai

Alibaba-NLP researchers introduced SecRespond, a benchmark designed to test AI agents on real post-compromise incident response rather than the clean, controlled puzzles most security benchmarks use. Each of its 10 test environments gives an agent a frozen forensic disk snapshot of a compromised cloud host plus synthetic host security outputs: alerts, vulnerability scans, and baseline checks. The agent must produce forensic reports on intrusions and risks, then

deliver a verified remediation plan. The benchmark spans 4 entry-point types, 21 ATT&CK techniques (MITRE's publicly maintained catalog of known attacker behaviors), and 5 operating systems. The researchers ran 23 frontier models through the OpenCode agent harness. Every model reliably found problems the alerts already exposed. None completed detection and remediation on any single range.

Real incident response requires finding evidence that no alert points to, tracing attacker activity across scattered disk artifacts, and writing a plan that cleans the machine without destroying evidence or leaving backdoors behind. SecRespond separates "following a signal" from "independently investigating a host." Today's models can handle alert triage but cannot work cold from a disk image, and security teams considering AI in their operations center now have a reproducible way to measure that gap.

Security vendors increasingly pitch AI-assisted triage as incident response. SecRespond draws a harder line, one that requires reading a disk for hidden activity without any alert to start from. Current models cannot cross it, and the benchmark makes that testable and reproducible.

Read More: Tracecat: Free, Open-Source Security Automation That AI Agents Build and Run

Sources:

- SecRespond paper (arXiv:2607.26791)
- SecRespond benchmark data and code
- SecRespond dataset on Hugging Face
- Hugging Face paper page

Disclaimer: *For information only. Accuracy or completeness not guaranteed. Illegal use prohibited. Not professional advice or solicitation.* ***Read more:*** */terms-of-service*