

PLAYBOOK

KI-gestützte autorisierte Penetrationstests

Wie wir LLMs und Agenten in abgesteckten Sicherheitsaufträgen einsetzen - und die menschlichen Gates, die daraus autorisierte Tests statt unbeaufsichtigte Scans machen.

Wo KI hilft - und wo sie nicht führen darf

Ein Penetrationstest ist ein autorisierter Auftrag gegen einen definierten Umfang. Dieser Rahmen ändert sich nicht, wenn ein LLM in den Workflow kommt: Das Modell beschleunigt die Analyse, entscheidet aber nie, was getestet wird, wie weit eine Exploitation geht oder was im Bericht steht. Jeder KI-gestützte Schritt in diesem Playbook liegt hinter einem menschlichen Genehmigungs-Gate.

Das Modell schlägt vor. Der Tester entscheidet. Nichts erreicht das Ziel, den Kunden oder den Bericht ohne die Freigabe einer namentlich verantwortlichen Person.

Wo externe Modelle zur Aufgabe passen - bereinigte Eingaben, eigene Systeme, Detection Engineering - arbeiten wir mit verifiziertem Zugang zu Daybreak Blue, der Trusted-Access-Stufe, die OpenAI für autorisierte defensive Sicherheitsarbeit betreibt. Alles, was Kundendaten oder Beweismittel berührt, bleibt auf dem [selbst gehosteten Stack](#).

Die vier Phasen, in denen sich KI bewährt

01 Aufklärungs-Zusammenfassung

Passive Daten - Certificate Transparency, öffentliche Repositories, Technologie-Fingerprints - fallen in Menge an. Ein LLM verdichtet sie zu einem Zielprofil, das der Tester Zeile für Zeile prüft, bevor irgendein aktiver Zugriff erfolgt. Die gewonnene Zeit fließt in die interessanten Hosts, nicht ins Lesen von Rauschen.

02 Finding-Triage

Scanner-Output ist verrauscht. Das Modell gruppiert doppelte Funde, entwirft Schweregrad-Kandidaten und markiert wahrscheinliche Fehlalarme. Der Tester akzeptiert, korrigiert oder verwirft jede Einschätzung; der Schweregrad wird nie automatisch vergeben.

03 Exploit-Recherche

Zu einer bestätigten Service-Version entwirft das Modell Angriffspfade und Referenzen. Der Tester verifiziert jeden gegen die reale Umgebung; ein plausibel klingender Pfad, der nicht existiert, ist schlimmer als keiner.

04 Berichtsentwurf

Jeder Fund erhält einen Entwurf: Beschreibung, Auswirkung, Nachweis, Behebung. Der Tester schreibt um, was der Entwurf verfehlt - meist die geschäftliche Auswirkung, denn das Modell kennt den Kontext des Kunden nicht.

Harte Regeln, die nicht verhandelbar sind

- ☐ Die schriftliche Autorisierung definiert Umfang, Zeitfenster und Techniken, bevor irgendein Werkzeug läuft - KI oder nicht
- ☐ Keine Kundendaten, Zugangsdaten oder Beweismittel gehen an externe Modellanbieter; die Analyse läuft auf selbst gehosteten Modellen oder mit bereinigten Eingaben
- ☐ Aktive Exploitation-Schritte werden vom Tester ausgeführt und beobachtet, nie an eine Agentenschleife delegiert
- ☐ Jede Aussage im Bericht wird vor der Übergabe manuell reproduziert - was wir nicht wiederholen können, wird nicht geliefert
- ☐ Die Auftragsdokumentation legt offen, welche Schritte KI-Unterstützung nutzten, damit der Kunde die Methode prüfen kann

Warum diese Disziplin für Kunden zählt

Das Risiko KI-gestützter Sicherheitsarbeit ist nicht Tempo - es ist ungeprüfte Selbstsicherheit. Eine halluzinierte CVE-Referenz oder ein erfundener Exploit-Pfad verschwendet Behebungsbudget und untergräbt das Vertrauen in echte Funde. Unsere Gate-Struktur stellt sicher, dass Ihr Bericht nur enthält, was demonstriert wurde, im Schweregrad, den ein Mensch vergeben hat, innerhalb des Umfangs, den Sie autorisiert haben. Das ist der Standard, den wir an unsere eigenen Systeme anlegen - und an jeden Auftrag, den wir annehmen.

Dieses Dokument ist Teil der UHL-Ressourcen. Direkter Download und verwandte Arbeiten:

<https://uhl.cloud/de/resources/>

