

PLAYBOOK

AI-Assisted Authorised Penetration Testing

How we use LLMs and agents inside scoped security work - and the human gates that keep it authorised testing, not unsupervised scanning.

Where AI helps - and where it must not lead

A penetration test is an authorised engagement against a defined scope. That framing does not change when an LLM enters the workflow: the model accelerates analysis, it never decides what is tested, how far exploitation goes, or what ships in the report. Every AI-assisted step in this playbook sits behind a human approval gate.

The model proposes. The tester disposes. Nothing reaches the target, the client, or the report without a named human approving it.

Where external models fit the task - sanitised inputs, our own systems, detection engineering - we work under verified access to Daybreak Blue, the trusted-access tier OpenAI operates for authorised defensive security work. Everything that touches customer data or evidence stays on the [self-hosted stack](#).

The four phases where AI earns its place

01 Recon summarisation

Passive data - certificate transparency, public repos, technology fingerprints - arrives in bulk. An LLM condenses it into a target profile the tester reviews line by line before any active touch. Time saved is spent on the interesting hosts, not on reading noise.

02 Finding triage

Scanner output is noisy. The model groups duplicate findings, drafts severity candidates and flags likely false positives. The tester accepts, adjusts or rejects each call; severity is never auto-assigned.

03 Exploit research

Given a confirmed service version, the model drafts candidate attack paths and references. The tester verifies each against the actual environment; a plausible-sounding path that does not exist is worse than none.

04 Report drafting

Each finding gets a draft narrative: description, impact, evidence, remediation. The tester rewrites what the draft gets wrong - which is often the business impact, because the model does not know the client's context.

Hard rules that do not bend

- ☐ Written authorisation defines scope, timing and techniques before any tool - AI or otherwise - runs
- ☐ No customer data, credentials or evidence is sent to external model providers; analysis runs on self-hosted models or sanitised inputs
- ☐ Active exploitation steps are executed and observed by the tester, never delegated to an agent loop
- ☐ Every claim in the report is reproduced by hand before delivery - if we cannot redo it, it does not ship
- ☐ The engagement record states which steps used AI assistance, so the client can audit the method

Why this discipline matters to a client

The risk in AI-assisted security work is not speed - it is unexamined confidence. A hallucinated CVE reference or a fabricated exploit path wastes remediation budget and erodes trust in real findings. Our gate structure exists so the report you receive contains only what was demonstrated, at the severity a human assigned, inside the scope you authorised. That is the standard we apply to our own systems, and it is the standard in every engagement we accept.

This document is part of the UHL resources collection. Direct download and related work:
<https://uhl.cloud/en/resources/>

