Xalgorix: open-source AI pentest tool adds a second agent to verify findings



Xalgorix, a free open-source AI penetration testing tool, has a second agent re-exploit each finding before it is reported. No benchmark results are published, and the Docker image runs as root.

- A verifier agent starts from a fresh context, gets read-only tools and must re-exploit each bug the first agent finds.
- Unproven findings are not discarded automatically; they go to a manual review pile unless positively disproved.
- The tool uses Nmap, Nuclei and SQLmap, works with seven LLM providers including local models via Ollama, and has a scope guard for target hosts.
- The maintainers have not published benchmark scores, and a full run of the XBOW validation

> benchmarks has not been completed.
> - The Docker image runs as root and needs --privileged or added capabilities, so it should be treated as a disposable sandbox.

Xalgorix is a free, open-source penetration testing tool released under the Apache-2.0 license. An LLM agent with a browser and a terminal works through a 22-phase test plan, from reconnaissance to a final report, driving standard tools such as Nmap, Nuclei and SQLmap. A scope guard blocks tool calls to hosts outside the configured targets and their subdomains.

How findings are verified

When the main agent finds a bug, a separate verifier agent must exploit it again. It starts from a fresh context, treats the first agent's proof as untrusted and has read-only tools, with each check capped at eight turns and three minutes. Bar levels differ by bug class: for XSS the script must actually execute, and for SSRF a callback from the target's server to a fresh token is required. If the model claims confirmation without a successful re-test, the code downgrades the verdict to inconclusive. Findings are dropped only on positive disproof; others go to manual review.

Limits to note

The founder says internal benchmarks exist, including paired vulnerable and secure builds and real CVEs against patched releases, but no results are published. The XBOW validation benchmarks have been started but not fully run. Claude Opus 5.5 has given the best results, while local models need more supervision on long scans.

The Docker image runs as root and needs --privileged or added capabilities. The maintainers advise using it as a disposable sandbox on an isolated network.

For regulated institutions, this is a tool to evaluate carefully rather than adopt on claims alone.

What to do now

1. If you want to trial Xalgorix, run it only in an isolated lab network, as the maintainers advise for the root container.
2. Restrict the configured targets to systems you own or have written authorisation to test, and confirm the scope guard behaves as expected.
3. Run it against known-vulnerable and patched builds of your own applications to measure false positives yourself, since no scores are published.
4. Keep a person responsible for reviewing findings left in the manual review pile.
5. Do not treat AI tool output as a substitute for authorised penetration testing required under your

regulator's framework.