

The Autonomous Pentester Scorecard

VENDOR Vendor name DATE 21 September 2026

CRITERION YES / NO

AGENTIC WORK QUALITY

Does the agent find flaws in your own logic, or just run a fixed list of checks?

- ✓ The agent reasons about your workflows: payments, access to other users' data, skipped steps.
- ✗ The agent runs a fixed list of checks against every target.

Yes No

Can the agent get past a real login page and test what's behind it?

- ✓ It logs into an app like a user and reaches real, authenticated features.
- ✗ It needs a recorded login, an authenticated token, or other assistance to reach authenticated features.

Yes No

Can the agent find vulnerabilities just as well without access to your source code?

- ✓ It works black-box, the way an attacker does. Source can add depth, but the agent is sophisticated enough to find vulnerabilities without it, just like an attacker.
- ✗ Its results hinge on source access; it doesn't provide much more value than a code review you could run yourself.

Yes No

Is pricing for the assessments transparent, or are features and scope gated behind tiers?

- ✓ One assessment covers a target, at one level of quality regardless of tier.
- ✗ Different "thoroughness" per pricing tier, or scope (e.g. API) and quality gated behind more premium tiers.

Yes No

Is mobile app testing included in the assessment?

- ✓ The offering tests both web and mobile apps, giving you consistent, comprehensive coverage.
- ✗ Web and/or API only; a different product, tier, or vendor is needed to test your mobile apps.

Yes No

Is it clear how much real pentest work the price includes?

- ✓ Unlimited LLM tokens, or clearly stated limits (tokens or time) that show how comprehensive an assessment is.
- ✗ Unclear pricing, token numbers presented without reference, or very few tokens per engagement.

Yes

No

Do the deliverables include reproduction steps and all applicable evidence?

- ✓ Every finding includes evidence you can double-check to verify it, and steps to reproduce it.
- ✗ Evidence and reproduction steps are omitted, unclear, or lacking from the deliverables.

Yes

No

Does each finding come with clear guidance on how to fix it?

- ✓ A concrete fix is provided for each finding, not just a description of the problem.
- ✗ The way to mitigate each finding is not provided.

Yes

No

DATA AND COMPLIANCE

Is inference under a named provider, with no-retention and no-training?

- ✓ You can obtain the name of the LLM provider(s), with a written assurance that your data is not retained or used in training.
- ✗ You can't be told who the LLM provider is, or given assurance they aren't retaining or training on your data.

Yes

No

Is it clear what data the vendor keeps, for how long, and how you delete it?

- ✓ The vendor can tell you what data they keep and what they use it for, and delete it on demand.
- ✗ Unclear what the vendor keeps or how to remove it.

Yes

No

LOGISTICS AND TRANSPARENCY

Can you run a pentest on demand, or on a schedule, easily?

- ✓ Pentests can be run on demand or on a schedule, without needing to contact anyone.
- ✗ You need to deal with a human to arrange for a pentest to be done.

Yes

No

Is there a token or time limit per run, and does the report say what it didn't reach?

- ✓ No token or time limit, or a clear explanation of what was tested and what was left.
- ✗ A strict token limit, with no indication of where the pentest stopped if it stopped short.

Yes

No

Do you get refunded if something goes wrong with the LLM or the environment during the engagement?

- ✓ Refunds are provided, automatically or on demand, if there are technical issues with the engagement.
- ✗ Refunds aren't provided, or technical issues aren't transparently surfaced.

Yes

No

OPERATIONAL SAFETY

Can you stop an engagement at any point?

- ✓ Engagements are quick and easy to stop; the environment winds down fast, and numbers can be provided on demand.
- ✗ Engagements can't easily be stopped, or stopping needs a call to the vendor, with no numbers on how long it takes to wind down.

Yes

No

Is the environment isolated, provisioned dynamically, and cleaned up after each engagement?

- ✓ Each client gets a temporary, clean environment per pentest. Nothing is shared, nothing is left behind once it concludes.
- ✗ Shared infrastructure between engagements.

Yes

No

Can you pick the agent's traffic out of your own logs?

- ✓ The agent's traffic carries a known header or known exit IP that identifies it in your logs.
- ✗ No way to tell the agent's requests from a real attack.

Yes

No

Can you run a pentest without switching off your own protections?

- ✓ You can run a pentest without disabling any security measures to accommodate the agent.
- ✗ The agent needs your protections off, or loses notable efficacy against a site with bot protection, a WAF, and the like enabled.

Yes

No

Total

0 / 17 Yes

A tool worth using should be close to all Yes. Where you set the bar is up to you.

NOTES

- The number of sub-agents dispatched per assessment is almost purely marketing. More is not necessarily better.
- Longer is not necessarily more thorough. Agents work quickly and around the clock, so a report within a few days, or even hours, is reasonable depending on the app's complexity.
- Pentest quality is hard to infer up front. Ask for a sample report against a benchmark, or a redacted report against a live target, so you can see what the agent actually finds. If no report is available, expect a demo walking through sample findings in the vendor's dashboard. [See a redacted report from an actual engagement of ours →](#)

Companion to [How to Evaluate an Autonomous Pentester](#)