



FOR FOUNDERS BUILDING AI SECURITY PRODUCTS

Is your AI cyber product's verdict right?

The 5 levels of testing every AI security product needs, and the one most founders skip.

WHY A BENCHMARK ISN'T PROOF

Buyers don't buy a score. They buy confidence your AI is right in *their* environment.

A wrong verdict in security isn't just noise. It's a missed attack, a wasted night for an analyst, or the wrong person flagged. The teams that win evaluations test at five levels, and each level catches what the others miss.

Inside: the 5 levels of testing and the blind spot of each (page 2) · the level most teams skip, with a real-style example, and how expert testing closes the gap (page 3).

Who it's for

Founders, CTOs and product leads building AI SOC, agentic pentest, CNAPP and cloud, detection, identity, data security and insider risk products.

How to use it

Check each level against how you test today. The blind spots you can't cover are where your next buyer's evaluation will look first.

The 5 levels of testing

1 Public benchmarks

How do we score against others on a standard test?

Catches: big capability gaps; headline numbers.

Blind spot: clean, simplified cases the model may already have seen.

2 Private test sets

Did our last release break anything?

Catches: regressions on cases you already know.

Blind spot: only tests what your team thought to include.

3 Automated validators and AI judges

Does a second program confirm each output?

Catches: unsupported claims and missing evidence, at scale.

Blind spot: an AI grading an AI shares the builder's assumptions.

4 Expert human review MOST OFTEN SKIPPED

Would a real practitioner trust and act on this?

Catches: right answer for the wrong reason, missing context, attempts to mislead the AI.

Blind spot: needs skilled people; doesn't scale on its own.

5 Real-world outcomes

Does it work for real customers, month after month?

Catches: drift, noise rates, analyst overrides.

Blind spot: arrives late, after customers have seen the problems.

You're doing each level well if...

- ☐ **L1** you report results across several runs, not one best run.
- ☐ **L2** your test set grows every time a customer or reviewer finds a miss.
- ☐ **L3** your validators check the evidence, not just the answer's format.
- ☐ **L4** practitioners outside your build team review real workflows on a schedule.
- ☐ **L5** you track analyst overrides and reopened cases as quality signals.

LEVEL 4

The level most founders skip

Levels 1–3 are run by the people who built the product, so they share its assumptions. Level 5 arrives after customers are already exposed. Level 4 is the only point where an independent practitioner judges the output **before** customers do, and it's the one early-stage teams skip.

EXAMPLE A BENCHMARK WOULD MISS

An AI SOC flags sign-ins from London and Amsterdam 5 minutes apart as **"impossible travel", high severity**. An analyst sees in seconds that the Amsterdam IP is the company's own VPN. Same device. Benign.

No benchmark caught it. A practitioner did.

What expert review catches that tests don't

- 1 Detections that fire perfectly in the test set, and on harmless real traffic.
- 2 Findings that are technically true but not exploitable in this setup.
- 3 The right verdict reached for the wrong reason, which won't generalise.
- 4 Untrusted data (log fields, file names, emails) that changes the AI's verdict.
- 5 The obvious case nobody thought to test, which an analyst spots on sight.

Where SecureFLIT fits

We run Level 4 for AI security vendors. Our pentesters, SOC analysts and detection engineers use your product the way your customers' teams would, review every verdict, and work with your team to fix what we find, before a customer does.

hello@secureflit.com · secureflit.com