ENDOR LABS

CISO GUIDE

How to evaluate build vs buy for code security

Every board is asking the same question: if AI is this good at writing code, why keep buying security tools? The honest answer is that a frontier model is a reasoning layer — not a replacement for SAST, SCA, secrets, and supply-chain defense. What turns that reasoning into a security program is everything around the model: deterministic context, verification, audit trails, and economics that survive scale. That's the harness, and it's the hard part.

What is an agent harness?

An AI code security tool is far more than a model. The harness is everything wrapped around it — the machinery that turns a convincing demo into a program you can defend in an audit. Models are converging, and everyone gets the same ones. The harness is where programs are won or lost, and it's the part your team would have to build and keep running.

- **Context** — feeds the model your whole codebase deterministically, on every run
- **Verification** — confirms a finding is real before it reaches a developer
- **Evidence** — an audit-ready chain behind every result
- **Economics** — cost that stays predictable as scan volume grows

Same models. Different results.

2.6×	12×	2.8×
more real vulnerabilities than a frontier model scanning on its own	lower token cost for identical security work	faster on identical security work

Efficacy: Endor Labs AI SAST benchmark, June 2026, vs. frontier models (Claude Opus 4.7 and Codex / GPT-5.5) across 8 real-world projects. Cost and speed: Endor Labs token-economics benchmark — 442 security prompts across 12 projects and 6 languages, an agent with Endor Labs tools vs. the same model reasoning alone (6.6M vs. 79.5M tokens; 133 vs. 370 min).

Not a demo: a zero-day in Anthropic's own code

AURI's AI SAST traced untrusted wire data to an unbounded allocation in buffa, Anthropic's Rust protobuf library — a ~22× memory-amplification DoS on a memory-safe target (**CVE-2026-55407**), caught by following data to the sink, not grepping for a pattern.

CISO GUIDE

The three axes to evaluate

The gap between a frontier model on its own and the same model inside a harness isn't a matter of opinion. It shows up on the three axes your evaluation has to cover.

EFFICACY

More real findings, less noise

- 192 real vulnerabilities — 2.6× the next-best tool
- 64 of 106 CWE weakness types detected; next-best reached 36
- 63 findings caught by no other tool in the test; next-highest was 18

COST

Frontier reasoning, a fraction of the cost

- 79.5M → 6.6M tokens for the same work — 92% fewer, via deterministic context
- ~$0.9M–$1.6M / year to build it in-house for 150 developers
- Frontier list prices are doubling, not falling

SPEED

Fast enough for the workflow

- 133 min vs. 370 min on the same 442-prompt suite — 2.8× faster
- 77.6% fewer tool-loop calls (1,380 vs. 6,165)
- Fast enough to run in every pull request, not overnight

Efficacy: Endor Labs AI SAST benchmark, June 2026, across 8 real-world projects. Cost and speed: Endor Labs token-economics benchmark — 442 security prompts across 12 projects and 6 languages, an agent with Endor Labs tools vs. the same model reasoning alone.

When is building credible?

BUILD IF YOU HAVE

- Dedicated platform and ML engineers on existing model-gateway infrastructure
- Genuinely unusual constraints: proprietary languages, air-gapped environments
- One bounded use case, not an AppSec program
- Funding for 18+ months of maintenance, not just the build

WHERE DIY BREAKS DOWN

- The scope is "replace the AppSec stack"
- Headcount is loaned, not dedicated
- There's no re-validation plan for model changes
- Nobody priced year two

How to use this guide

Page 3 is a scoreable readiness check — mark each of the 12 questions Don't know, Partial, or Ready, then tally the columns. **Page 4** maps your score to a recommendation and gives you the objections and vendor questions to pressure-test the proposal.

ENDOR LABS

CISO GUIDE

The 12-question readiness check

Evaluation question	Don't know	Partial	Ready
EFFICACY & COVERAGE			
What fraction of a large repo does each run analyze, and can you prove coverage? **Ready:** per-scan coverage reporting. **Red flag:** "the model reads what it needs."	☐	☐	☐
Which use cases does it actually cover — SAST, SCA, containers, secrets, compliance? **Ready:** honest scoping to one or two. **Red flag:** "the model generalizes."	☐	☐	☐
Are results consistent across re-runs, and can you walk an auditor through a finding? **Ready:** a deterministic harness with an evidence chain. **Red flag:** "we spot-check."	☐	☐	☐
COST & ECONOMICS			
What's the projected annual token bill at your PR volume, and who owns the line as it grows? **Ready:** a cost model with a named owner. **Red flag:** "on the platform team's budget, somewhere."	☐	☐	☐
What happens to the fully loaded cost if model prices double? **Ready:** a sensitivity analysis. **Red flag:** pricing at today's rates. Frontier list prices doubled with the latest generation.	☐	☐	☐
What does it cost to build, then re-validate, the harness every time the model changes? **Ready:** an eval suite and a budget for it. **Red flag:** "the model is the product."	☐	☐	☐
SPEED & MAINTENANCE			
Does a full scan fit inside a PR check today? At 3x repo growth? **Ready:** minutes, measured. **Red flag:** "we run it nightly."	☐	☐	☐
Who keeps it current as models, MCP, and frameworks change? **Ready:** named owners with dedicated time. **Red flag:** "the team maintains it."	☐	☐	☐
When it finds something real, can you ship a verified fix in hours? **Ready:** autonomous, evidence-backed remediation. **Red flag:** detection feeding a ticket queue. Mandiant puts mean time-to-exploit at negative seven days.	☐	☐	☐
OWNERSHIP & ACCOUNTABILITY			
Who is accountable when the internal tool misses a vulnerability that ships? **Ready:** it's written down. **Red flag:** silence.	☐	☐	☐
What happens when the tool's champion changes teams? **Ready:** docs and a succession plan. **Red flag:** it's one person's project.	☐	☐	☐
Will it pass your next SOC 2, FedRAMP, or customer security review? **Ready:** compliance was in scope from day one. **Red flag:** "we'll get to it."	☐	☐	☐

ENDOR LABS

CISO GUIDE

Pressure-test the proposal

How to evaluate objections

They'll say	Ask this
"We'll chunk and index the repo."	That's a context graph. What's the estimate to build and maintain it across languages, monorepos, and refactors?
"We'll cache prompts and use cheaper models."	Caching cuts cost, not non-determinism or coverage gaps. Who re-validates finding quality after every model swap?
"We'll start with SAST and expand."	What's the roadmap and headcount to reach the program's full scope?
"Models keep improving, so maintenance shrinks."	Every model upgrade is a behavior change to test. Where's the eval suite? Who owns it?

How to read your score

BUILD
9+ Ready
- You have a credible build case for a narrow scope.
- Pressure-test the maintenance plan and revisit in 12 months.

HYBRID
5+ Partial
- Buy the platform layer: context, verification, coverage, audit.
- Build your own policies, integrations, and workflows on top.

BUY
4+ Don't know
- You're not ready to build.
- The unknowns are the cost.

Ask your vendors the same questions

- Show me your benchmark methodology. What did you test?
- How do you prove coverage on my repos, per run?
- Are results deterministic? Run the same scan twice.
- Walk me through the evidence chain behind one finding, the way you'd walk an auditor through it.
- What happens to my bill when your model vendor raises prices?
- When I ask for a fix, what do I get besides a finding?

Evaluate the harness, not the demo

Everyone gets the same models. Security programs succeed or fail on the harness: **context, verification, economics, and accountability**. Build it, buy it, or split the difference — whichever path you take, evaluate the harness, not the demo.