# Validating What Matters: Assurance, Adversarial Testing, and the Cost of AI-Driven Validation in the DIB

Sep 30, 2026 · by Terrence J. McGraw, CEO Cape Endeavors Inc

## Executive summary

No single validation method proves that a defense contractor can keep Controlled Unclassified Information (CUI) out of adversary hands. Compliance assessments and AI-driven penetration testing each see part of the attack surface, and each is blind where the other is strong.

The July 2026 suspension of CMMC Phase 2 third-party assessments has turned a technical question into a planning question. If the Department of War (DoW) leans on commercial capabilities such as automated AI penetration testing, contractors need to know what those tools can validate and what they will cost.

This paper makes three arguments:

1. **Continuous adversarial testing is closer to real-world risk** for technical controls, especially at the network edge, where exploitation now happens in hours.
2. **Adversarial testing alone is not sufficient.** It is largely blind to identity attacks, third-party compromise, insider handling of CUI, and detection and response, which together drive roughly half of real breaches.
3. **Independence matters more than method.** The best-documented failure in the Defense Industrial Base (DIB) is controls that were attested but never implemented. A contractor-run tool, however sophisticated, does not solve that.

The second half explains how AI tools are billed, which penetration testing functions drive cost, and what a hypothetical 500-host contractor would pay under on-demand and continuous models. The headline: a continuous AI testing mandate could cost more per year than a triennial third-party assessment, and remediation labor, not the tool, is likely to be the largest line item.

Contractors cannot control the meter, but they control what goes into it. Scope, hygiene, and cadence, all products of mature CMMC implementation, are the main levers on cost under every scenario.

## Part I: What actually validates risk reduction

### The policy moment

Third-party CMMC assessment is suspended and removed from contracts pending review, not permanently abolished. As of this paper's date, the final direction is unknown.

| Date | Action |
|---|---|
| Sep 11–13, 2026 | CMMC Reform Task Force report due to the DoW CIO; not yet public (Covington) |
| Sep 3, 2026 | Class Deviation 2026-O0025 (Rev. 3) directs contracting officers to remove third-party CMMC requirements; Level 1 and Level 2 self-assessments permitted; NIST SP 800-171 Rev. 2 remains the baseline under DFARS 252.204-7012 (CMMC Map) |
| Jul 13, 2026 | CIO memo "Removing Barriers to Defense Industrial Base Expansion" suspends Phase 2, which would have required C3PAO assessments from Nov 10, 2026, and creates the Reform Task Force (Holland & Knight) |
| Jul 13, 2026 | RFI seeks views on using existing commercial cybersecurity capabilities and optimizing self-attestation (SBA Advocacy) |

Three things have not changed. DFARS 252.204-7012 and NIST SP 800-171 **obligations remain in force**. C3PAOs remain authorized to conduct Level 2 certification assessments (Peak InfoSec). And inaccurate attestations still carry False Claims Act exposure.

The RFI's interest in commercial capabilities makes automated validation, including AI penetration testing, a plausible outcome worth planning for. This paper treats it as a scenario, not a prediction.

## How breaches actually happen

Any validation method should be judged against the paths adversaries actually use. The 2026 Verizon Data Breach Investigations Report (DBIR), covering incidents from November 2024 to October 2025, gives the clearest current picture.

| Finding (2026 DBIR) | Figure | What it means for validation |
|---|---|---|
| Vulnerability exploitation as initial access | 31%, up from 20%; first time at #1 in 19 years (SecurityWeek) | Speed matters; defenders have hours, not months |
| Identity-related initial access (phishing + credential abuse) | ~32% combined (Push Security) | Human and identity controls are as important as patching |
| Credential abuse anywhere in the breach chain | 39% (NHIMG) | Lateral movement relies on valid accounts |
| Breaches with third-party involvement | 48%, up 60% in a year (SecurityWeek) | Supply chain and service providers are core attack surface |

Nation-state actors targeting the DIB add two patterns that broad breach data underweights. They use zero-day exploits against edge devices such as VPNs and firewalls. And they "live off the land," abusing valid accounts and built-in administrative tools to avoid detection for months.

The implication is that a credible validation approach must address four things at once: exposure to known vulnerabilities, identity and human controls, third-party risk, and the ability to detect an adversary who is already inside.

## What each validation method can see

Continuous AI penetration testing is strongest against the single largest breach vector. Independent assessment covers more vectors, with less depth and far less timeliness.

| Attack path | Continuous AI penetration testing | Independent assessment (C3PAO / DIBCAC) |
|---|---|---|
| Known vulnerabilities on edge devices | **Strong.** Finds exploitable exposure within hours or days. | **Weak.** Confirms a patch process exists; a triennial snapshot cannot keep pace. |
| Missing or misconfigured technical controls (MFA gaps, open services, weak segmentation) | **Strong.** Tests whether controls hold, not whether they are documented. | **Moderate to strong.** Verifies implementation through evidence and demonstration, by sampling. |
| Phishing, pretexting, social engineering | **Weak.** Usually out of scope for automated tools. | **Moderate.** Confirms training and MFA exist, not that they work. |
| Third-party and supply chain compromise | **Weak.** Third-party systems are outside authorized scope. | **Moderate.** Examines external service providers, flow-downs, and scoping. |
| Zero-days and living off the land | **Weak.** Unknown flaws cannot be found; valid-account abuse looks normal. | **Weak to moderate.** Confirms logging, monitoring, and incident response exist, not that they would catch a skilled adversary. |
| Insider handling of CUI (media, email, removable storage, unsanctioned AI) | **Weak.** Largely invisible to network testing. | **Moderate to strong.** Much of NIST SP 800-171 targets this directly. |

The hardest scenario, a skilled nation-state actor already inside the network, is covered well by **neither**. It is best tested through adversary emulation or purple team exercises that measure detection and response, which is a third capability entirely.

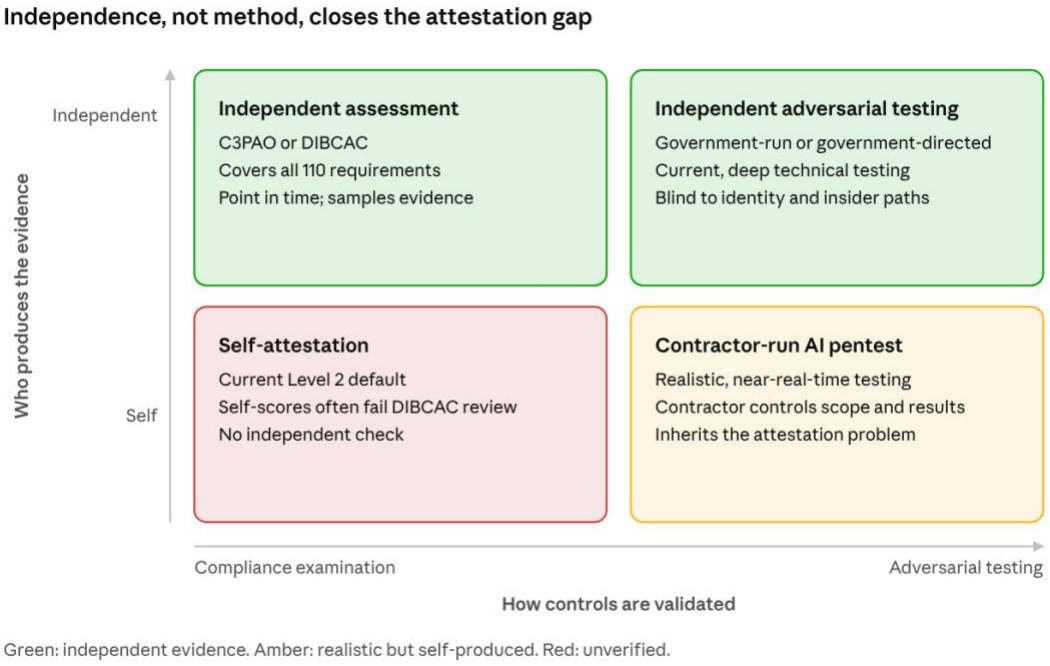
The "point in time" critique cuts both ways. Assessments are often defended as measuring program maturity rather than a snapshot. But a triennial assessment is itself a snapshot, and continuous testing is the less time-bound of the two for technical controls.

## Independence: the variable that matters most

The best-documented DIB problem is not controls that are implemented but insufficient. It is controls that are claimed but not implemented.

The evidence is consistent. When DIBCAC began auditing self-reported SPRS scores after 2020, it found that most perfect scores were inaccurate (Ancitus). In one 2026 False Claims Act settlement, a contractor self-reported a perfect 110; DIBCAC scored the same environment at −170 (Arnold & Porter).

The current debate conflates two separate variables. **Method** is whether validation examines compliance or tests adversarially. **Independence** is whether someone other than the contractor produces the evidence. Crossing them yields four models, not two.




validation models · method by independence

A contractor-run AI penetration test is a real improvement on self-attestation, but it inherits the attestation problem. The contractor controls scope, configuration, and what gets reported. Independence comes from who produces the evidence, not from how sophisticated the tool is.

## A combined assurance model

The approach with the strongest evidence behind it layers three capabilities, each covering what the others miss.

| Layer | What it answers | Cadence | Covers |
|---|---|---|---|
| Independent verification of implementation | Are the 110 requirements actually in place, including governance, people, and process? | Periodic (e.g., triennial, with annual affirmation) | Insider handling, third-party management, training, policy, and the attestation gap |
| Continuous adversarial testing | Can an attacker exploit the environment right now? | Continuous or change-triggered | Edge exposure, misconfiguration, technical control drift |
| Detection and response testing | Would we notice and contain an adversary already inside? | Annual or semiannual | Zero-days, living off the land, incident response readiness |

Neither side of the current policy debate is fully right. Independent validation is well supported by DIB-specific evidence. Adversarial testing is well supported as a measure of real-world risk, for the portion of the attack surface it can see.

Whatever the Task Force recommends, contractors should understand which gaps their mandated validation method leaves open and decide deliberately whether to close them.

## Part II: What AI-driven validation will cost

### How AI tools are billed

AI models charge by the token, and an agentic penetration test consumes tokens far faster than most business uses of AI.

A token is a chunk of text, roughly four characters or three-quarters of an English word. Everything the model reads is billed as **input**; everything it writes is billed as **output**. Output typically costs three to five times more than input.

Five mechanics explain most of the bill:

- **Models have no memory.** Each request resends the full history. In a 40-step test, step 40 re-reads steps 1 through 39, so cost grows faster than the work done.
- **Fixed overhead rides on every call.** The tool's instructions and tool definitions can run 5,000–15,000 tokens and are resent every turn.
- **Security data tokenizes poorly.** IP addresses, hashes, base64, and JSON fragment into many tokens. An address like 10.14.207.33 can take seven or eight.
- **Reasoning is billed but unseen.** Models that "think" before acting bill that reasoning at the higher output rate.
- **Failures cost as much as successes.** A failed exploit attempt stays in the context and is re-read on every later turn.

Two discounts matter. **Prompt caching** can cut the cost of repeated context by roughly 90%, and **batch processing** can cut non-urgent work by roughly 50%.

Enterprises rarely pay per token directly for a security tool. The vendor pays for tokens as a cost of goods and charges per asset, per scan, or in credits. Token economics still matter, they determine whether a vendor's pricing is sustainable, where overage charges appear, and whether work gets quietly routed to cheaper, less capable models.

## Which penetration testing functions drive cost

The most expensive functions are those requiring judgment across many steps, which are also the functions that justify using AI over a conventional scanner.

| Function | Why it consumes tokens | Relative cost |
|---|---|---|
| Web application crawling and analysis | Raw HTML and JavaScript are enormous; one page can run tens of thousands of tokens | Very high |
| Exploit iteration and fuzzing | Many attempt-fail-adjust turns; every failure stays in context | Very high |
| Active Directory and attack-path analysis | Directory data (users, groups, ACLs, GPOs) is massive; chaining privileges needs heavy reasoning | Very high |
| Authenticated and business-logic testing | Multi-step workflows with state carried forward | High |
| False-positive validation | Each flagged item needs evidence review; noisy environments create many | High, scales with findings |
| Configuration review | Firewall rules, cloud IAM policies, and GPO exports are long and dense | Moderate to high |
| Reporting and 800-171 control mapping | Ingests all findings; long output at the premium rate | Moderate, output-weighted |
| Remediation retesting | Narrow and targeted, but frequent in continuous mode | Low to moderate |
| Host discovery and port scanning | Cheap if the AI reads a summary; expensive if fed raw verbose output | Low to moderate, design-dependent |

Cost scales with how disordered an environment is, not only how large it is. A sprawling network with many misconfigurations produces more findings, more validation, more exploit attempts, and longer reports.

## Worked example: a hypothetical AI testing platform

For a contractor with 500 IP addresses and 20 web applications, an on-demand scan costs about $395 in tokens; continuous testing costs about $4,200 per month.

The rates below are illustrative round numbers, not any provider's price list. All figures are the vendor's token cost, before margin, infrastructure, human review, and support.

| Model tier | Input ($/1M tokens) | Output ($/1M tokens) | Cached input ($/1M tokens) |
|---|---|---|---|
| Frontier (deep testing, reporting) | 3.00 | 15.00 | 0.30 |
| Small (triage, monitoring) | 1.00 | 5.00 | 0.10 |

**One deep web application test.** About 40 agent turns, with context growing from 10,000 to 80,000 tokens (average 45,000). That is roughly 1.8 million input and 60,000 output tokens: **about $6.30**. With 80% of input served from cache, the same test costs **about $2.40**.

### Mode 1: on-demand scan

| Phase | Work | Token cost ($) |
|---|---|---|
| Recon and triage | 500 hosts × ~20K input / 2K output, small model | ~15 |
| Deep testing | 40 hosts + 20 web apps = 60 targets × $6.30 | ~378 |
| Report generation | ~500K input / 30K output, frontier model | ~2 |
| **Total per scan** | | **~395 (~160 with caching)** |

Run quarterly, that is roughly $130 per month. Costs are spiky, predictable, and tied to deliberate decisions.

### Mode 2: continuous testing

| Activity | Work | Token cost ($/month) |
|---|---|---|
| Hourly change monitoring | 500 hosts × 24 × 30 = 360,000 small-model checks (3K in / 300 out) | ~1,620 |

| Activity | Work | Token cost ($/month) |
|---|---|---|
| Weekly full sweep | 4 × $395 | ~1,580 |
| Triggered retests | ~5 changes/day × 30 × $6.30 | ~945 |
| Daily digest reports | 30 × ~$2 | ~60 |
| **Total per month** | | **~4,200 (~50,000/year)** |

"Always on" is not what drives the cost. Frequency × scope × context size does. Hundreds of thousands of small monitoring calls cost more than all deep testing combined.

A better-engineered platform would compare scan snapshots with conventional code and call the AI only when something changes. That design choice alone could bring the continuous bill to roughly $1,000–1,500 per month.

## Validation cost by policy scenario

A continuous AI testing mandate could cost more per year than a triennial third-party assessment, shifting spend from a periodic lump sum to a permanent operating expense.

| Scenario | Direct validation cost | Largest hidden cost |
|---|---|---|
| Self-attestation only (current state) | Internal labor | False Claims Act exposure; no independent check |
| C3PAO assessment (if restored) | ~$105,000 (small entities) to ~$118,000 (larger), every three years, or ~$35,000–40,000 per year (CMMC.com) | Remediation surges before each assessment |
| Mandated continuous AI testing | ~$50,000 per year in token cost alone for the example environment, before vendor markup | Continuous remediation labor |

**These figures are not directly comparable:** the assessment estimate is a full-service price, while the AI figure is only the vendor's raw token cost. The direction is still clear. ***Vendor pricing will sit above token cost, not below it.***

If AI testing becomes the compliance mechanism rather than an optional security tool, four cost levers disappear:

- **Scope becomes mandated.** The entire CUI boundary must be tested, not just critical assets.
- **Cadence becomes mandated.** A continuous or monthly requirement locks contractors into the higher-cost mode.
- **Tool choice narrows.** Tools processing security data about CUI environments will likely need FedRAMP Moderate or equivalent. A short list of qualified vendors in a mandated market holds pricing power.

- **Remediation becomes continuous.** Continuous testing produces continuous findings, each needing engineering time, change control, and retesting. This labor will likely dwarf the subscription.

### Controlling what goes into the meter

Contractors cannot control per-token rates or a vendor's architecture, but they control the inputs that drive token consumption. Each lever is also a product of mature CMMC implementation.

| Lever | How it lowers cost | Related practice |
|---|---|---|
| Tight scoping | A well-built CUI enclave shrinks the in-scope footprint, and AI cost falls in proportion | CMMC scoping guidance |
| Asset inventory and hardened baselines | Fewer unknown assets and misconfigurations mean fewer findings to validate and report | CM.L2-3.4.1, CM.L2-3.4.2 |
| Disciplined patching and remediation | Fewer exploitable paths mean fewer multi-step exploit attempts | RA.L2-3.11.3, SI.L2-3.14.1 |
| Change-triggered cadence | Testing tied to tracked changes rather than a fixed clock | CM.L2-3.4.3 |
| Contract structure | Flat per-asset pricing puts overrun risk on the vendor; consumption pricing puts it on the contractor | Procurement decision |

Cadence deserves a note. No NIST SP 800-171 requirement mandates penetration testing at Level 2; the relevant requirements are vulnerability scanning (RA.L2-3.11.2), remediation (RA.L2-3.11.3), and periodic control assessment (CA.L2-3.12.1). Penetration testing is explicit at Level 3 under NIST SP 800-172 (CA.L3-3.12.1e). Unless a new rule changes this, continuous AI testing at Level 2 is a security choice, not a compliance mandate.

**The tool itself is also in scope.** An AI testing platform ingests network maps, vulnerability findings, and configuration data about the CUI environment. That is likely Security Protection Data, and the vendor and its AI model provider may fall within assessment scope as external service providers. FedRAMP-authorized or government-cloud AI deployments often carry premium pricing and narrower model choice than commercial endpoints.

## Recommendations

Contractors should act now on what holds under every policy outcome, and evaluate AI testing vendors on coverage and cost mechanics, not marketing claims.

**For contractors, regardless of what the Task Force recommends:**

1. Keep an honest, evidence-backed SPRS score and a current SSP and POA&M. Every scenario requires them, and inaccurate attestations carry False Claims Act risk.
2. Tighten CUI scoping now. It lowers the cost of every validation method.
3. Budget for remediation labor, not just tooling. Continuous testing turns remediation into a standing operating cost.
4. Map the gaps your validation method leaves open, and decide deliberately whether to close them with independent assessment or detection-and-response testing.

**Questions to ask any AI penetration testing vendor:**

- ☐ When does the platform actually call an AI model, and what triggers it?
- ☐ Which models are used for which tasks, and can work be routed to cheaper models without notice?
- ☐ Is pricing flat per asset or consumption-based, and how are overages calculated?
- ☐ Where is our data processed, and does the deployment meet FedRAMP Moderate or equivalent?
- ☐ Is the vendor, and its AI model provider, prepared to be in our assessment scope as an external service provider?
- ☐ Are social engineering, third-party paths, and detection testing in scope, or excluded?
- ☐ How are findings mapped to NIST SP 800-171 assessment objectives, and who can alter the results before they are reported?

## Sources and assumptions

Cost figures are illustrative estimates for a hypothetical platform and environment; verify current published model pricing before relying on them. Policy status reflects public reporting as of September 30, 2026. This paper is not legal advice.


- Holland & Knight, "DOW Suspends CMMC Phase II Requirements" — https://www.hklaw.com/en/insights/publications/2026/07/dow-suspends-cmmc-phase-ii-requirements
- Covington, Inside Government Contracts, "CMMC Reform Task Force Updates, September 2026" — https://www.insidegovernmentcontracts.com/2026/09/cmmc-reform-task-force-updates-september-2026/
- CMMC Map, "CMMC Task Force Report: What to Expect" — https://cmmcmap.com/blog/cmmc-task-force-september-report/
- SBA Office of Advocacy, "DoW Requests Information for CMMC Reform Task Force" — https://advocacy.sba.gov/2026/07/20/dow-requests-information-for-cmmc-reform-task-force/
- Peak InfoSec, "CMMC Reform Task Force Update" — https://peakinfosec.com/cmmc-reform-task-force-update/
- SecurityWeek, "Verizon DBIR 2026: Vulnerability Exploitation Overtakes Credential Theft" — https://www.securityweek.com/verizon-dbir-2026-vulnerability-exploitation-overtakes-credential-theft-as-top-breach-vector/

- Push Security, "What the Verizon DBIR Tells Us About Breaches in 2026" — https://pushsecurity.com/blog/verizon-dbir-2026-review
- NHIMG, "Verizon DBIR 2026 Shows Credential Abuse Is Down but Not Out" — https://nhimg.org/articles/verizon-dbir-2026-shows-credential-abuse-is-down-but-not-out/
- Arnold & Porter, "Dramatic Difference Between Self-Reported and DIBCAC Assessed NIST SP 800-171 Scores" — https://www.arnoldporter.com/en/perspectives/blogs/fca-qui-notes/posts/2026/06/dramatic-difference-between-self-reported-and-dibcac-assessed
- Ancitus, "What Your SPRS Score Actually Means in 2026" — https://ancitus.com/insights/sprs-score-explained/
- CMMC.com, "CMMC FAQ: Certification, Requirements & Costs" — https://www.cmmc.com/faq


## AI disclosure

Research and drafting for this paper were assisted by AI (Claude, Anthropic). Analysis, conclusions, and recommendations are the author's, and all sources were independently reviewed.