AI ASSURANCE · TEST & EVALUATION

Proving the Machine

Test and evaluation for AI you can field — why “it worked in the demo” is not evidence, and how to test, evaluate, and red-team a model before it carries a mission.

 Intelligence. Integrity. Impact.

AI Assurance · Test & Evaluation

Every AI failure in production was a passing demo first. A demo is a curated success: the right inputs, a forgiving audience, and none of the edge cases, adversaries, or drift that the real world supplies for free. Fielding AI on the strength of a demo is not diligence, it is hope. Test and evaluation is how hope becomes evidence — and it is the discipline federal AI adoption most often shortchanges.

Demo is not evidence

Measure

The NIST AI RMF function programs most often skip — the one that turns "it works" into proof.

NIST AI RMF

ATLAS

MITRE's catalog of real-world adversarial techniques against AI systems — a red team's starting map.

MITRE ATLAS

Demo ≠

A demo shows what can go right; test and evaluation shows what will go wrong. Only one is evidence.

Artheus principle

AI is not tested the way software is

Traditional software is deterministic: given an input, it does the same thing every time, and a test suite that passes today passes tomorrow. AI is different in ways that break the old playbook — its behavior is statistical, its failures are often silent, adversaries can manipulate it through its inputs, and its performance decays after deployment as the world moves. Testing it like ordinary software produces a green checkmark and a false sense of safety (Figure 1).

	Traditional software test	AI test & evaluation
Behavior	Deterministic, repeatable	Statistical — right "most" of the time
Passing criteria	Correct output	Accuracy, robustness, and fairness thresholds
Adversary in scope	Rarely	Always — red-teaming is core, not optional
After deployment	Stable until changed	Drifts; must be monitored continuously

Figure 1. Why an AI system needs its own T&E.; The properties that make AI powerful — statistical behavior, sensitivity to inputs, learned rather than written logic — are exactly the ones a software test suite was never designed to check.

What real AI T&E; covers

- **Functional performance on mission data.** Not the vendor's benchmark — accuracy, precision, and recall on the actual data and the actual edge cases the system will face.
- **Robustness.** How gracefully it degrades on noisy, out-of-distribution, or unexpected inputs, rather than failing confidently.
- **Adversarial resistance.** Red-teaming against documented techniques — prompt injection, evasion, data poisoning, model extraction — mapped to MITRE ATLAS and the OWASP LLM Top 10.
- **Human factors.** Whether the people using it can tell when to trust it, and whether the interface surfaces uncertainty instead of hiding it.

Testing is a lifecycle, not a gate

The decisive shift is to stop treating T&E; as a one-time gate before launch and start treating it as a lifecycle that continues for as long as the system is used. Metrics are defined up front, a representative test set is curated, the system is evaluated and red-teamed, and then it is monitored in production — where drift or a newly discovered attack routes straight back into re-evaluation (Figure 2).

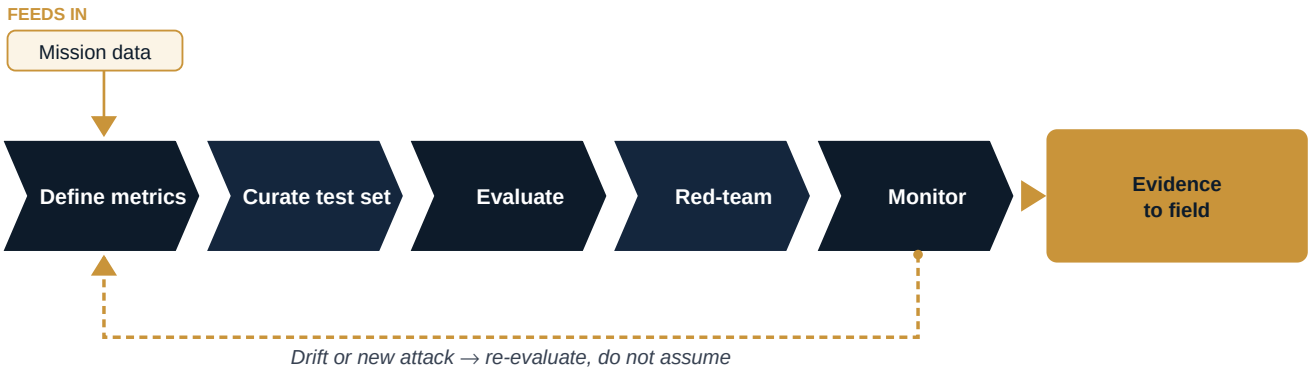

Figure 2. AI test and evaluation as a continuous lifecycle. Evaluation and red-teaming are not the finish line — monitoring feeds drift and newly discovered attacks back into re-evaluation, so the evidence that justified fielding stays current.

Red-teaming is not optional

If your organization does not adversarially test its AI, someone else will — with less friendly intent. Red-teaming asks the question functional testing never does: not "does it work," but "how do I make it fail, mislead, or leak." For language models that means prompt injection, jailbreaks, and coaxing out training data or system prompts; for other models it means evasion and poisoning. MITRE ATLAS and the OWASP Top 10 for LLM Applications turn this from an art into a checklist with a starting map. The goal is not a perfect score — it is to find the failures in a lab, on your schedule, instead of in the field, on the adversary's.

Evidence is the deliverable

The point of test and evaluation is not a report that sits in a drawer; it is the evidence that lets a responsible official field an AI system and defend that decision. It is also the natural companion to disciplined acquisition — the acceptance criteria, independent T&E;, and monitoring triggers a contract should require (see our companion paper on de-risking AI acquisition). Buy AI you can test, test the AI you buy, and keep testing it. That is the difference between an AI capability you can trust and one you are merely hoping about.

References

1. National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1, January 2023 (Measure function).
2. National Institute of Standards and Technology, *AI RMF Generative AI Profile*, NIST AI 600-1, 2024.
3. MITRE, *ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems)*, atlas.mitre.org.
4. OWASP Foundation, *OWASP Top 10 for LLM Applications*, 2025 edition.
5. U.S. Department of Defense, Chief Digital and Artificial Intelligence Office (CDAO), Responsible AI and test-and-evaluation guidance.
6. U.S. Government Accountability Office, *Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities*, GAO-21-519SP, 2021.