ZERO TRUST · SECURE AI ADOPTION

Zero Trust Doesn’t Stop at the Model

Federal agencies are mandated to adopt Zero Trust and to adopt AI. Run as separate programs, they undercut each other. Here is how to make the model a governed resource instead of a new blind spot.

 Intelligence. Integrity. Impact.

Zero Trust · Secure AI Adoption

Two mandates are landing on the same enterprise at the same time. One says move to Zero Trust: assume breach, verify everything, stop trusting the network because it sits inside the perimeter (Executive Order 14028; OMB M-22-09). The other says adopt AI, and quickly, to keep the mission competitive (Executive Order 14179; OMB M-25-21). In most organizations these are run by different teams, on different clocks, in different vocabularies. The seam between them is precisely where the next class of incidents is going to happen.

Two mandates, one attack surface

68%

Share of breaches involving a human element — the attack surface a language model most enlarges.

Verizon DBIR 2024

LLM01

Prompt injection ranks first on the OWASP Top 10 for LLM Applications — the model’s defining new risk.

OWASP, 2025

Zero

Implicit trust a Zero Trust design grants any actor — a user, a device, or a model.

NIST SP 800-207

The collision nobody scheduled

A Zero Trust program spends eighteen months hardening identity, devices, network segmentation, and data controls. In parallel, an AI pilot stands up a language model that ingests documents, answers questions, and — without anyone quite deciding it should — becomes a new place where data moves and a new component nobody threat-modeled. Prompts carry data outward. Retrieved content carries instructions inward. The model itself is a third-party artifact of uncertain provenance. And for every sanctioned pilot there is usually a shadow one, running on an unmanaged account against a public endpoint.

> *A language model is not a feature. It is a workload with data access, an identity, dependencies, and an egress path. Zero Trust already has a discipline for exactly that — it just hasn’t been pointed at the model yet.*

What actually goes wrong

The failure modes are documented, named, and mapped to real adversary behavior. They are not hypothetical, and they are not exotic:

- **Prompt injection** (OWASP LLM01:2025; MITRE ATLAS AML.T0051). Untrusted content — a web page, a PDF, an email in a retrieved corpus — carries hidden instructions the model follows. Indirect injection is the version most teams miss, because the attacker never talks to the model directly.
- **Sensitive information disclosure** (OWASP LLM02:2025). The model surfaces data a given user was never authorized to see, or a crafted prompt walks it out through the answer.

- **Supply-chain risk** (OWASP LLM03:2025). Open weights, fine-tunes, and model dependencies of unknown origin are software you did not write and often cannot inspect.
- **Excessive agency** (OWASP LLM06:2025). An agent handed tools and standing permissions can take actions well beyond what its task required.
- **System-prompt leakage** (OWASP LLM07:2025). The instructions and guardrails meant to stay hidden become part of the attack surface.

Map the risks to the pillars

NIST SP 800-207 already gives the structure. Each AI-specific risk lands on a pillar the architecture already governs, with a control the security team already understands (Figure 1). The work is not inventing new doctrine — it is refusing to exempt the model from the doctrine you have.

OWASP LLM risk (2025)	Zero Trust pillar	Primary control
LLM01 · Prompt injection	Application · Data	Treat all retrieved and external content as untrusted; guard both inputs and outputs
LLM02 · Info disclosure	Data	Enforce authorization at the record level; ground output so it cannot assert beyond its sources
LLM03 · Supply chain	Device · Workload	Pin and verify model provenance; run in a controlled enclave, not a trusted-by-default binary
LLM06 · Excessive agency	Identity	Least privilege; scope the tools and standing permissions an agent may use
LLM07 · Prompt leakage	Application	Separate secrets and policy from the prompt; monitor for exposure

Figure 1. The model's new risks are not new pillars — they are the pillars you already run, pointed at a workload you haven't governed yet. (Mapping: OWASP Top 10 for LLM Applications 2025 to NIST SP 800-207.)

Every AI request, checked at the door

Concretely, a governed model treats each request the way Zero Trust treats each connection: nothing is trusted because it arrived, and every step can refuse (Figure 2). The user is authenticated, the request authorized, the input screened, the model's output grounded and logged. A policy violation ends the request — it does not get the benefit of the doubt.

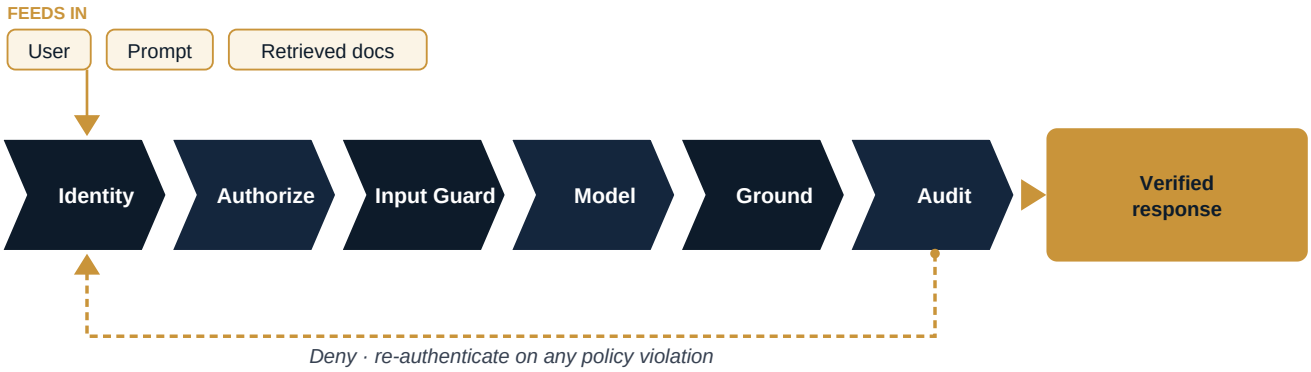

Figure 2. An AI request routed through Zero Trust checkpoints. Identity and authorization gate access; the input guard treats retrieved content as hostile; the output is grounded and written to the audit log. Any check can deny — trust is never inherited from the previous step.

Human-in-the-loop is a control, not a courtesy

For any consequential decision, a person verifies grounded output before it becomes action. Citation-grounding is what makes that review fast enough to actually happen — the reviewer checks the cited source, not the entire chain of reasoning. This is also how an organization satisfies the meaningful-oversight expectations that run through the NIST AI RMF and the high-impact-AI provisions of OMB M-25-21. Oversight that cannot be performed in the time available is oversight in name only, and grounding is what buys back the time.

If it touches CUI, the rules come with it

The moment an AI system ingests Controlled Unclassified Information, the obligations of DFARS 252.204-7012, NIST SP 800-171, and CMMC follow it — they do not pause because the component is new. An AI pilot that pipes CUI into an unaccredited external service is not an innovation story, it is a compliance finding waiting to be written. The clean answer is to bring the model inside the same boundary and the same control set as the data it reads.

Mind the shadow pilot

The most dangerous AI in most organizations is the one nobody approved: a team pasting sensitive text into a public chatbot to hit a deadline. Prohibition alone does not work, because the need is real. The durable fix is to make the governed path the easy path — a sanctioned, in-boundary model that is faster to reach than the workaround. Zero Trust gives you the visibility to find shadow AI; a usable governed alternative is what makes people stop reaching for it.

Verify the machine, too

Zero Trust’s core instinct — never trust, always verify, assume breach — applies to a model as cleanly as it applies to a user or a laptop. The mistake is exempting AI from that discipline because AI is new and the pilot is exciting. Treat the model as what it is: a privileged workload with data access, dependencies of uncertain origin, and a path to move information. Govern it there, across the pillars the architecture already defines, and AI stops being the hole in the Zero Trust design and becomes one more verified resource inside it.

References

1. The White House, Executive Order 14028, *Improving the Nation’s Cybersecurity*, May 2021.
2. Office of Management and Budget, *Moving the U.S. Government Toward Zero Trust Cybersecurity Principles*, OMB M-22-09, January 2022.
3. National Institute of Standards and Technology, *Zero Trust Architecture*, NIST Special Publication 800-207, August 2020.
4. Cybersecurity and Infrastructure Security Agency, *Zero Trust Maturity Model*, Version 2.0, April 2023.
5. OWASP Foundation, *OWASP Top 10 for LLM Applications*, 2025 edition (v2.0).
6. MITRE, *ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems)*, atlas.mitre.org.
7. National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1, January 2023.
8. Office of Management and Budget, *Accelerating Federal Use of AI through Innovation, Governance, and Public Trust*, OMB M-25-21, April 2025.
9. Verizon, *Data Breach Investigations Report (DBIR)*, 2024.
10. Defense Federal Acquisition Regulation Supplement, clause 252.204-7012; NIST Special Publication 800-171; Cybersecurity Maturity Model Certification (CMMC).