Your AI Is Under Attack. Here's What Actually Stops It.

A MITRE ATLAS™ Threat Assessment for Enterprise AI Leaders

Every engineering team shipping production AI hits the same wall: security and compliance won't sign off. The models work. The use cases are proven. The data can't move. The problem isn't that enterprises lack AI ambition — it's that most are solving the wrong security problem.

The MITRE ATLAS matrix — the adversarial threat landscape for AI systems — catalogs 14 tactics and 84+ techniques that attackers use against AI. It's the most rigorous public framework for understanding how AI systems get compromised. We mapped every tactic and technique against the architecture of confidential AI to answer one question: **Where does the actual risk concentrate, and what structurally eliminates it?**

The answer surprised even us. Not because the threats are worse than expected — but because the threats that block enterprise AI deployments are overwhelmingly concentrated in one category. And that category has a structural fix — one that's already available in your cloud infrastructure.

This assessment was produced by OPAQUE Systems, whose confidential AI platform implements the architecture described below. ServiceNow, Anthropic, and Accenture already run production workloads on OPAQUE. We mapped every ATLAS threat against our architecture and publish the results — including the threats we don't address — because no single vendor or solution addresses all of them, and any vendor that claims otherwise is either redefining the threats or misunderstanding the architecture.

The Uncomfortable Math

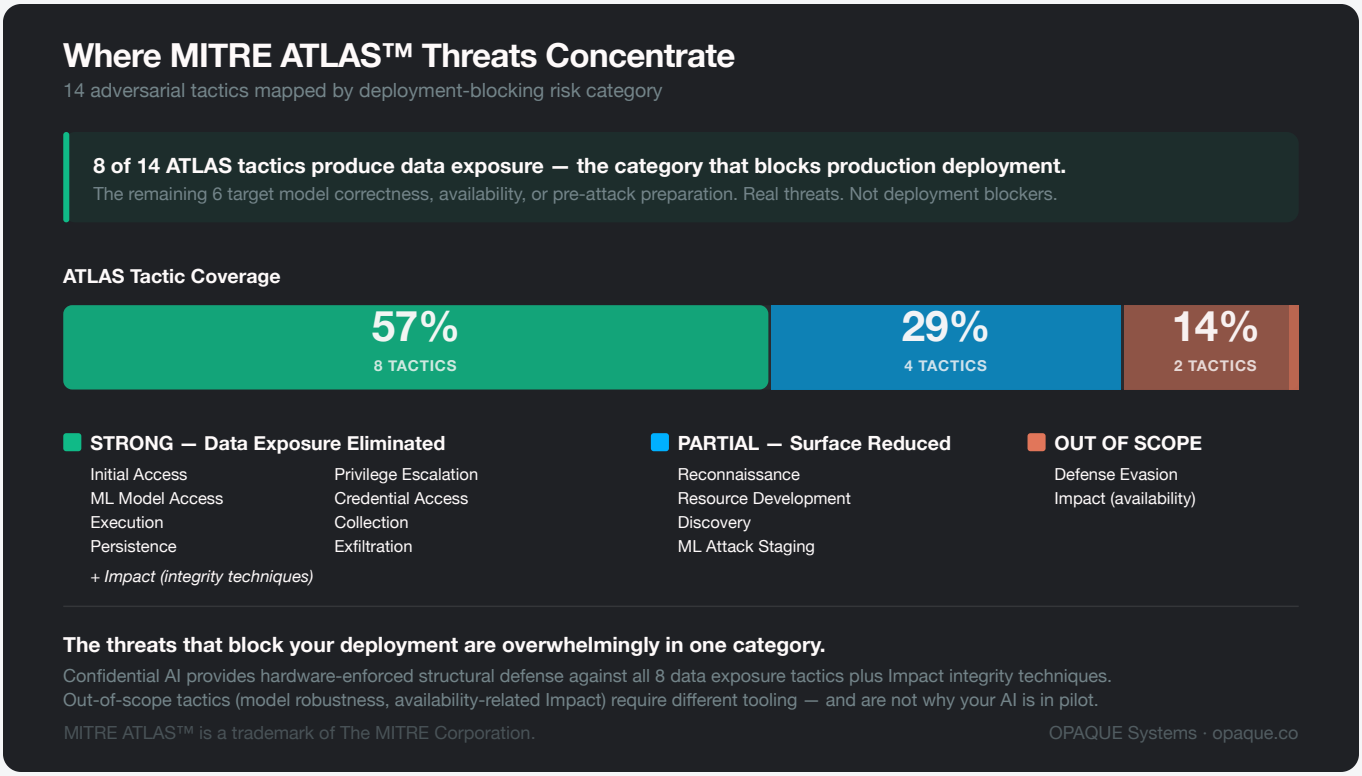
MITRE ATLAS organizes AI threats into 14 adversarial tactics. Think of these as the attacker's playbook: reconnaissance, initial access, execution, persistence, privilege escalation, credential theft, data collection, exfiltration, and more. Each tactic contains multiple techniques — specific methods attackers use to achieve their objectives.

When we mapped these 14 tactics against enterprise deployment risk, a pattern emerged that reframes the entire security conversation.

Eight of the fourteen tactics produce **data exposure** — the category with the highest regulatory consequence and the one that actually blocks deployments. These include credential theft, data collection, model exfiltration, supply chain compromise, and unauthorized access during processing. This is the category where your CISO says no. This is the category where DORA Article 9, EU AI Act Article 15, and GDPR Article 32 impose requirements your current architecture cannot satisfy with audit logs and access controls alone.

The remaining six tactics target model correctness, system availability, or pre-attack preparation. These are real threats. They are not why your AI is stuck in pilot.

The distinction matters because it determines where to invest. Adversarial robustness tooling (defending against evasion attacks that cause misclassification) and availability engineering (defending against denial of service) are important — but they don't unblock the deployment decision. Data exposure does.



What Confidential AI Actually Addresses

Confidential AI is not a product category — it's an architectural pattern. At its core: hardware-enforced trust boundaries that encrypt data during processing, not just at rest and in transit. The mechanism operates in three phases.

Consider what happens when a single AI query hits a standard enterprise pipeline. A bank customer asks their AI advisor about refinancing. That one query copies the customer's SSN, credit score, mortgage balance, investment portfolio, and tax returns to 15 different systems — the APM vendor, the embedding provider, the LLM provider, the observability platform, the Redis cache, the API gateway logs, the database snapshots. Every tool in the stack does exactly what it's supposed to do. The customer's entire financial identity is replicated 15 times. Nothing is broken. The bleed is structural.

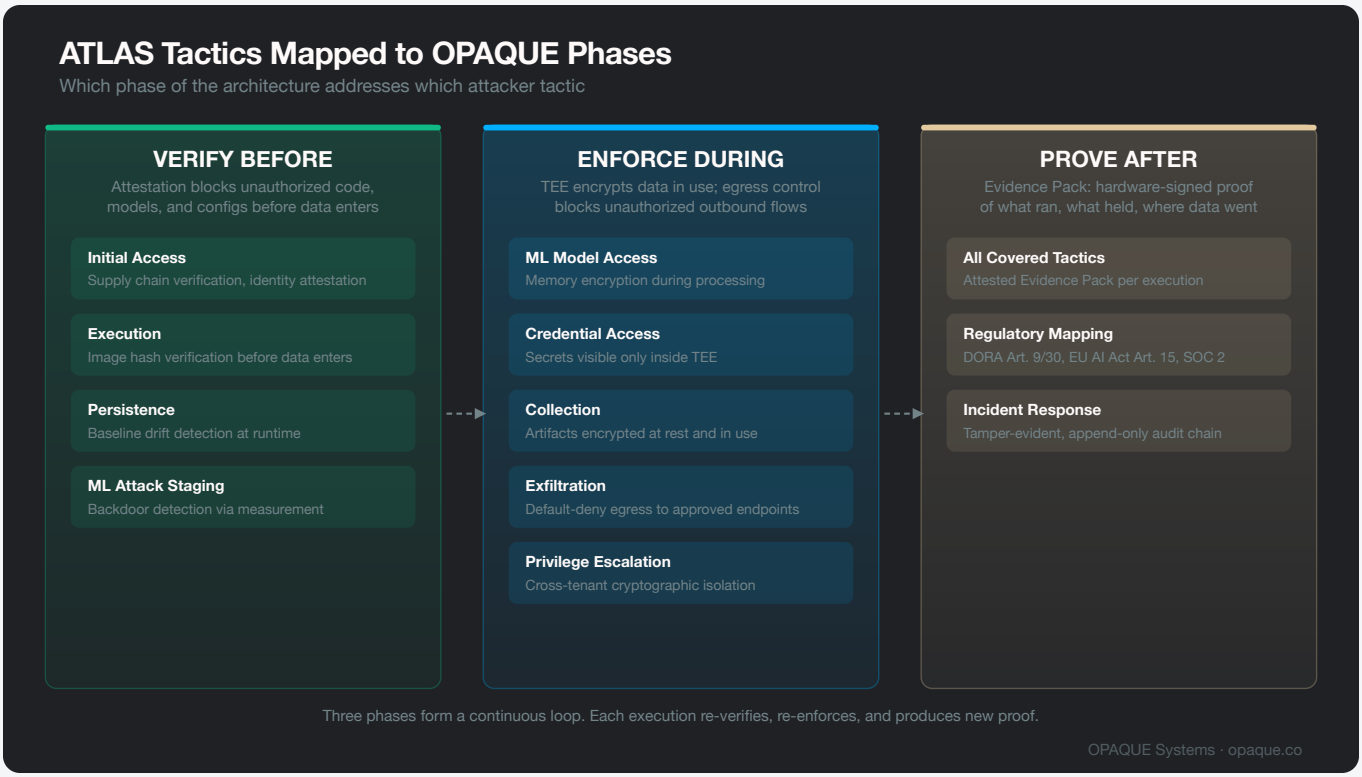
Confidential AI eliminates this class of exposure through three phases:

Verify Before. Cryptographic attestation confirms the identity and integrity of every workload before it touches sensitive data. If a malicious dependency changes the software image, attestation fails and processing never starts. If an unauthorized model is swapped in, the measurement changes and the system rejects it.

Enforce During. Data inside a Trusted Execution Environment remains encrypted during processing. Not encrypted before and after — encrypted while the computation is running. Memory dumps return ciphertext.

Hypervisor introspection sees nothing. Privilege escalation to root on the host yields zero data payoff. Egress controls enforce default-deny networking: only allow-listed destinations are reachable from inside the trust boundary. That same bank query? The APM vendor sees latency metrics. The embedding never leaves the enclave. The LLM provider receives nothing. Fifteen copies become zero.

Prove After. An attested evidence pack provides cryptographic proof of what code ran, what policies held, what data went where, and that the hardware protections were active at the moment of processing. Not a configuration snapshot from deploy time. Runtime proof. This is the artifact that satisfies DORA's requirement for demonstrable controls during processing and EU AI Act's requirement for cybersecurity throughout the AI lifecycle.



The Honest Assessment: 8 Strong, 4 Partial, 2 Out of Scope

We graded every ATLAS tactic on a four-point scale: **Solves** (hardware makes it physically impossible), **Mitigates** (cryptographic enforcement reduces blast radius), **Reduces Surface** (architecture makes the attack harder as a byproduct), and **Out of Scope** (requires different tooling entirely).

Here is where confidential AI lands:

> **Strong coverage — 8 tactics (plus integrity-related Impact techniques).** Initial Access, ML Model Access, Execution, Persistence, Privilege Escalation, Credential Access, Collection, and Exfiltration. These are the data exposure tactics. Hardware trust boundaries either eliminate the attack vector entirely or contain the blast radius to the point where exploitation yields ciphertext instead of cleartext. Impact techniques that target data integrity — corruption or tampering during processing — are also mitigated by TEE memory encryption and attestation.

> **Partial coverage — 4 tactics.** Reconnaissance, Resource Development, Discovery, and ML Attack Staging. Confidential AI reduces the observable surface available to attackers — model metadata, internal architecture, and training pipeline details are not visible outside the trust boundary — but these are pre-attack activities where the primary defense is operational security, not runtime data protection.

> **Out of scope — 2 tactics.** Defense Evasion (adversarial inputs that cause misclassification) and the availability-related portions of Impact (denial of service, cost harvesting). These target model correctness and availability. Confidential AI protects data, not model behavior. Adversarial robustness requires adversarial training, input validation, and ensemble methods. Availability requires rate limiting, auto-scaling, and infrastructure redundancy.

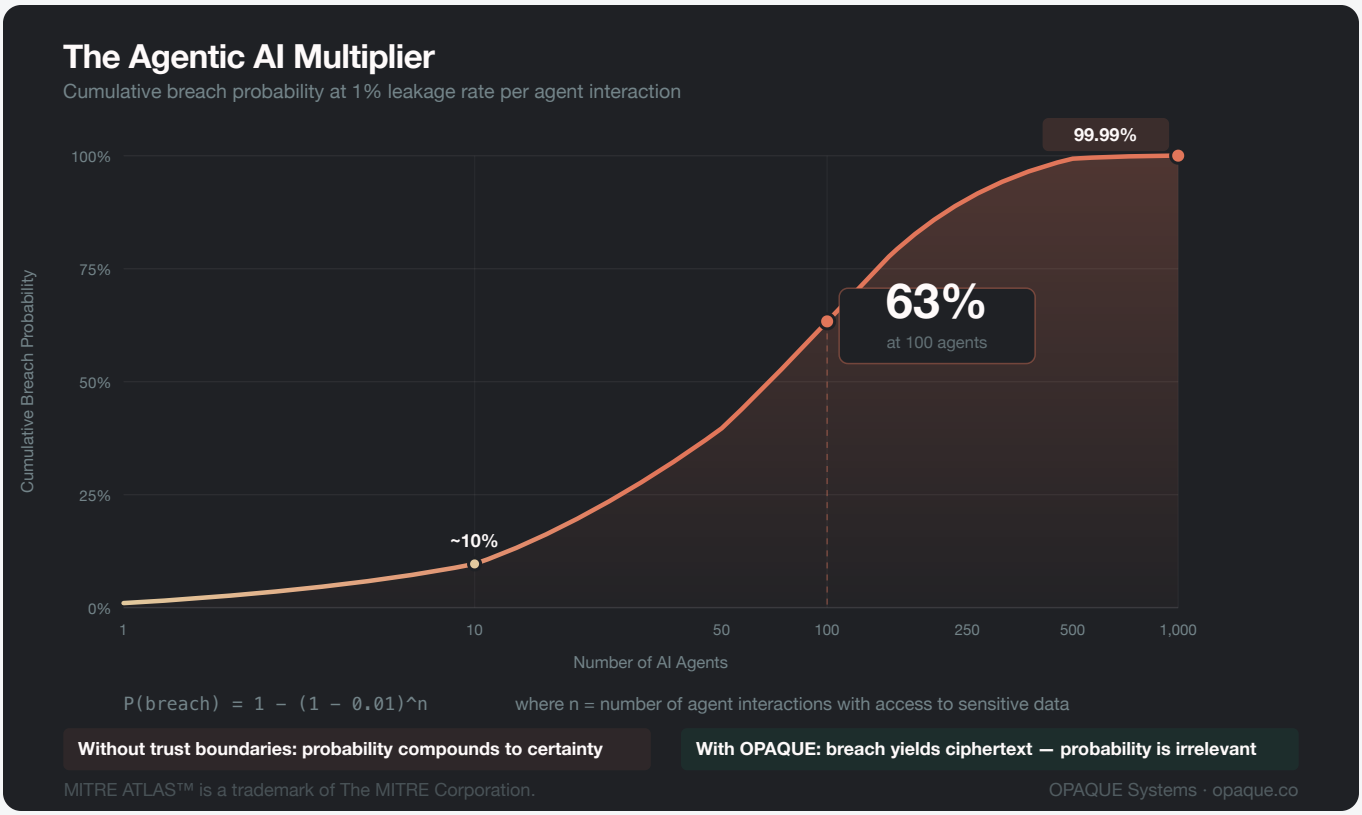
We don't fully address 2 of 14 tactics — Defense Evasion and the availability side of Impact. Those target model behavior and uptime — not data. They matter, but they're not why your deployment is stalled.

The Agentic AI Multiplier

At 1% probability of data leakage per agent interaction, a network of 100 agents produces a 63% cumulative probability of breach. At 1,000 agents, you're at 99.99%. Hierarchical access controls that work for human users cannot scale to autonomous agents operating at machine speed with human-like capabilities. MITRE ATLAS recognized this in October 2025, adding 14 LLM and agent-specific techniques to the matrix. Agentic AI is where the trust gap is widest — and where the combinatorial math breaks traditional security models.

Confidential AI addresses this directly. Agent tool calls are subject to runtime policy enforcement — only allow-listed tools are permitted. Prompt injection cannot trigger data exfiltration because egress controls block unauthorized outbound paths regardless of what the model is instructed to do. Agent delegation cannot expand privileges beyond the attested policy boundary. The attack succeeds inside the model; the data payoff is zero because the trust boundary holds.

This is not theoretical. It's the architectural difference between discovering a breach after the fact and making the breach structurally worthless at the moment it occurs.



What This Means for Your Deployment Decision

If you're evaluating AI security, the MITRE ATLAS matrix gives you the taxonomy. The question is which threats to prioritize.

The threats that block production deployment — the ones your CISO cites, your regulator examines, and your board asks about — are overwhelmingly data exposure threats. These are the threats where confidential AI provides hardware-enforced structural defense — deployed as a platform on your existing cloud, not as special infrastructure you need to procure.

The threats that confidential AI does not address — adversarial robustness, availability, output-based inference attacks — are real and require investment. But they are not the reason your AI is in pilot. They are the reason you

need a layered security architecture where each layer addresses its category of risk with the appropriate mechanism.

Seven of fourteen ATLAS tactics map directly to regulatory requirements under DORA and EU AI Act. For each, confidential AI produces the evidence artifact those regulations demand: hardware-signed proof that data was protected during processing, that runtime controls held, and that the computational environment was verified before sensitive data entered.

Not compliance theater. Cryptographic proof.

Further Reading

- MITRE ATLAS — Adversarial Threat Landscape for AI Systems — the full matrix of 14 tactics and 84+ techniques
- ATLAS Data Repository — canonical technique taxonomy in machine-readable format
- ATLAS SAFE-AI Report — MITRE's framework for securing AI against ATLAS threats
- OPAQUE Systems — Confidential AI Platform — the three-phase architecture referenced in this assessment

Methodology: Full tactic-by-tactic analysis of MITRE ATLAS v4.1 (October 2025) cross-referenced against the OPAQUE: Enterprise AI Data Leakage (46 exposure vectors across 8 categories).