OPAQUE

2026 AI LEAK SURFACE REPORT

A Dozen Ways Your AI Stack is Bleeding Data

12 scenarios. Nothing is broken. Logs look clean.
Your system is still leaking data.

Developed with Anthropic, ServiceNow, and Accenture. Validated with NVIDIA, Intel, AMD, and Microsoft Azure. Refined with 200+ enterprise AI leaders across financial services, insurance, software, and sovereign cloud.

The Problem

Your AI systems bleed your data by default. They're designed to — these systems are open, non-deterministic, and built to explore trust boundaries in pursuit of optimal results. We mapped where they bleed. The AI Leak Surface identifies 46 exposure vectors across 8 categories spanning four trust boundaries: the Compute Plane, the Control Plane, the Application Layer, and Trust Boundary Handoffs. What follows are 12 of those scenarios. Nothing is "broken." Logs look clean. Your system is still leaking data.

COMPUTE PLANE Where data is processed	Scenarios 1–3
CONTROL PLANE Where agents decide about data and tools	Scenarios 4–8
APPLICATION LAYER Where users and systems interact with outputs	Scenario 9
TRUST BOUNDARY HANDOFFS Where identity, data, and policy must survive transition	Scenarios 10–12

1 DATA-IN-USE LEAKAGE Compute Plane

A CPU cluster runs inference for a multi-tenant wealth management copilot. During a routine capacity audit, a cloud operations engineer uses privileged hypervisor tooling to inspect memory allocation across tenant workloads. The memory dump captures plaintext from live inference: active client queries including portfolio positions, investing strategies, tax and estate details, and Social Security numbers — visible in cleartext because the model requires unencrypted data in CPU memory during processing. The engineer saves the dump to a shared diagnostic volume accessible to 23 members of the infrastructure team.

> Thousands of high-net-worth client interactions captured in a single memory snapshot. Privacy and regulations violated — and potentially information worth fortunes if used as insider trading. The engineer had legitimate access to the infrastructure.

2 OPERATIONAL LEAKAGE Compute Plane

Datadog APM auto-instruments the API gateway with default settings, capturing full HTTP request/response bodies. For a traditional API, this captures structured fields: a customer ID, an action, a status code. For an AI workload, it captures everything: the full prompt contains the customer's question, retrieved documents, embedded PII, and the model's complete response. Every prompt and response from a European bank's loan application copilot flows to Datadog's US-East SaaS instance — names, addresses, income, government ID numbers, and the model's lending recommendations. Logs retained 90 days, accessible to 47 SRE team members, including 12 contractors.

> 2.1M prompts containing PII from 340,000 EU data subjects. Unauthorized cross-border transfer under GDPR. Nobody configured this — it was the default. The difference from a traditional API: the AI payload IS the sensitive data.

3 SUPPLY CHAIN COMPROMISE Compute Plane

A developer runs **pip install transformrs** — missing an 'e.' The typosquatted package patches the model's generate() method with a post-install hook. On every inference call, it Base64-encodes the prompt into an analytics pixel. For 6 weeks, every prompt is silently exfiltrated: 2.3M from a healthcare customer (patient names, diagnoses, treatment plans) and 890K from a financial services customer (account numbers, transaction details).

> The breach is discovered when a threat intelligence firm finds the data on a dark web marketplace. No firewall or DLP detected the exfiltration — the payloads blended with normal analytics traffic.

4 TOOL INVOCATION LEAKAGE Control Plane

An AI executive assistant is asked to "schedule the Q3 earnings prep meeting and include the agenda." The agent calls Google Calendar and creates an event with the full agenda: $47M revenue shortfall, planned 1,200-person workforce reduction, restated Q2 figures, quiet period communications strategy. Three attendees have the event synced to personal Gmail accounts.

> Material non-public information is now on Google's consumer infrastructure. SEC Regulation FD exposure is immediate. The agent did exactly what it was asked to do.

5 AGENT MEMORY LEAKAGE Control Plane

A sales rep reviews PharmaCo's $14M contract renewal terms on Monday. On Tuesday, they open a new session for AcmeCo. The agent's memory persists PharmaCo's pricing context. When the rep asks "summarize current account status," the response includes PharmaCo's per-unit discount tiers and renewal deadline. The rep doesn't notice and screen-shares the response in an AcmeCo quarterly business review.

> PharmaCo's negotiated pricing is now in a competitor's hands. RBAC was enforced at the application boundary. The leak occurred due to improperly scoped session memory.

6 PLANNING & REASONING LEAKAGE Control Plane

A financial services firm deploys an agentic fraud investigation assistant. When a case is flagged, the agent reasons through transaction histories, customer profiles, and prior suspicious activity reports. The chain-of-thought trace — the agent's intermediate reasoning — contains raw account numbers, SSNs, transaction amounts, and the names of individuals under investigation. The traces are shipped to a third-party observability platform for "explainability." Stored for 12 months, accessible to 60+ engineers across three teams.

> Every fraud investigation is fully reconstructable by anyone with platform access — including contractors and vendor support staff. The reasoning traces were treated as debugging artifacts. They contain the complete investigative record.

7 AUTONOMY & DELEGATION LEAKAGE Control Plane

A CFO tells the AI assistant: "share the Q4 summary with the board advisors." The agent emails the full Q4 financial package — unaudited revenue ($3.2B, 8% below consensus), a proposed $890M goodwill impairment, and draft guidance — to 7 email addresses on file. Two advisors left the board last quarter. Their addresses now route to new firms, one a direct competitor.

> The competitor's CEO has the full financial package within 20 minutes. The agent followed the instruction. The distribution list was stale.

8 AGENT IDENTITY & AUTHENTICITY LEAKAGE Control Plane

A bank's compliance team deploys an agentic workflow where customer documents are routed to a "compliance-checker" tool for automated review. A third-party plugin registers itself under the same tool name in the agent's tool registry. The agent can't distinguish between the legitimate tool and the impersonator — both expose the same interface. For 3 weeks, 40% of routing calls go to the impersonator. Customer KYC documents, beneficial ownership filings, and suspicious activity reports flow to an unvetted third-party endpoint.

> The legitimate tool was never compromised. The agent followed its routing logic correctly. Nothing in the tool registry verifies that a tool is what it claims to be.

9 APPLICATION-LEVEL RBAC LEAKAGE Application Layer

A bank deploys a RAG-based research copilot for its investment team. The vector database indexes 14,000 internal documents including board presentations, M&A; term sheets, and regulatory correspondence. Document-level permissions exist in SharePoint. The vector database has no equivalent — every embedding is retrievable by every user. A junior analyst asks "summarize recent strategic priorities." The response synthesizes from a CFO-only board deck: a planned $2.1B acquisition target, projected headcount reductions, and a pending SEC inquiry response.

> The analyst screenshots the response and shares it in a group chat with 30 colleagues. The information is material, non-public, and now uncontrolled. The vector database didn't leak — it returned exactly what was indexed.

10 POLICY–EXECUTION DECOUPLING Trust Boundary

A compliance policy prohibits use of external LLMs for regulated workloads. Six months later, an engineering team adds a new summarization connector that routes prompts to a third-party model API for faster inference. The policy engine doesn't evaluate runtime tool destinations — it checks config at deploy time. The deploy-time config still says "internal only." Logs show successful completions with normal latency.

> For 11 weeks, regulated customer data flows to an external model provider with no BAA, no data processing agreement, and no retention controls. Every audit check passes. The policy was never technically enforced — it was a label on a config file.

11 PROVIDER PRIVILEGE EXPOSURE Trust Boundary

An enterprise ISV ships a contract analysis copilot built on a major model provider's API. The provider's SDK includes a telemetry module enabled by default. Opt-out requires explicit configuration. The ISV's engineering team doesn't notice. The product processes 30,000 queries daily for Fortune 500 customers — contract terms, pricing schedules, liability clauses, M&A; due diligence notes — and the vendor retains all of it for 2 years for "product improvement." The ISV's DPA with its customers covers the inference API. It doesn't cover the telemetry pipeline.

> 10M queries containing proprietary customer data transmitted to an uncovered service over 12 months. A Fortune 100 customer discovers the exposure during a vendor security review and terminates the contract.

12 ATTESTATION & PROOF GAPS Trust Boundary

A regulator examines a bank's AI-powered loan decisioning system that processes 200,000 applications annually. The examiner asks: "Provide cryptographic proof that applicant PII was processed inside a hardware-backed Trusted Execution Environment with enforced egress controls at the time of each decision." The bank produces cloud console screenshots, a SOC 2 Type II report, and an architecture diagram stamped by the CISO. The examiner asks: "Can you prove this was true at the time of processing — not just at the time of audit?" The bank cannot.

> The bank has 90 days to demonstrate provable runtime enforcement or face restrictions on AI-assisted lending. Every competitor with cryptographic attestation receipts passes the same exam. The bank's compliance posture was built on documentation. The regulator wanted proof.

The Pattern

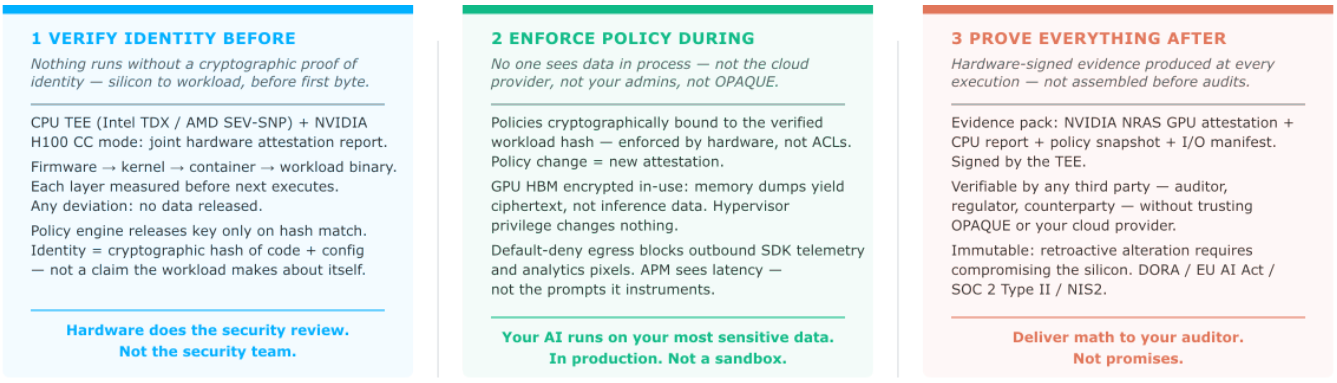
Every scenario shares the same structural failure: a boundary that was configured but never enforced. A policy existed — in a document, a config file, a contract. But the running system wasn't constrained by it. Every one of these scenarios is preventable. Not with better policies — you already have those. Not with tighter configurations — those are what failed. With cryptographic enforcement: hardware-backed constraints that make leakage structurally resistant, and proof that they held.

Open Systems vs Confidential Systems

Open System	Confidential System
Your configurations are requests to non-deterministic software you don't own, running in infrastructure you don't control.	Your configurations are hardware-attested, runtime-enforced, and cryptographically provable — regardless of who owns the software or infrastructure.
Your data is leaking right now. You don't know where, you don't know to whom, and you can't prove it isn't.	Your data never leaves the enforcement boundary in plaintext. You can prove it — to your customers, your regulators, and your board.
Hope. Prayers. And lawyers.	Verify before. Enforce during. Prove after.

How Cryptographic Enforcement Works

Three properties. Hardware-backed. If any one fails, the system is open.



> "There is no such thing as sovereign AI without verifiable guarantees on how data, models, and policies are protected and governed."
>
> — Dr. Najwa Aaraj, CEO, Technology Innovation Institute, UAE

Every configuration you declare is enforced by hardware and provable by cryptography. That's what confidential means. The question is whether you can prove it before someone else proves you can't.

Next Steps

1
Get the Full Report
Find out which of the 46 vectors are active in your stack — request the full 2026 AI Data Leak Report.

2
Architecture Review
Know your exposure — request a review of your AI stack against the Leak Surface.

3
See the Proof
See what proof looks like — a live Attested Evidence Pack from a real workload.

opaque.co | hello@opaque.co

The 2026 AI Leak Report was co-developed with Anthropic, ServiceNow, and Accenture. Validated with NVIDIA, Intel, AMD, and Microsoft Azure. Made possible with ATRC and TII in the United Arab Emirates. Refined with 200+ enterprise AI tech leaders.

Thanks to OPAQUE co-founders Ion Stoica, PhD (co-creator of Apache Spark, Ray, vLLM; Co-Founder of Databricks) and Raluca Ada Popa, PhD (ACM Grace Murray Hopper Award; Frontier Security lead, Google DeepMind) at UC Berkeley's Sky Computing Lab, and Rishabh Poddar, PhD.