ANOMALI

Guide

HOW TO USE AI TO REDUCE YOUR SECURITY BUDGET

A fidelity-first guide for CISOs and SOC leaders: the data quality dimensions, the stack architecture, the quality gates, and the baseline metrics that turn AI from a line item into a cost reducer.

MIT's 2025 study *The GenAI Divide* found that roughly 95% of enterprise generative-AI pilots produced no measurable impact on the bottom line, with the research pointing to data quality rather than model choice as the deciding factor. Security operations follow the same rule. AI lowers cost when it runs on high-fidelity data. It raises cost when it runs on everything a SOC happens to collect. This guide covers what fidelity means in practice, the stack that produces it, the gates that keep it honest, and the baseline metrics that show the savings.

MORE DATA RAISES COSTS UNLESS IT IS HIGH-FIDELITY

More data lowers cost only when that data is high-fidelity. Volume on its own multiplies the storage bill, the analyst hours spent sorting noise, and the cost of an agent acting fast on ungoverned inputs.

The evidence is consistent across the market. BARC's 2025 Lessons From the Leading Edge survey of 421 **organizations found data quality named as the single biggest obstacle to AI success**, rising from 19% of respondents in 2024 to 44% in 2025.

The security-side numbers show what that same problem, fragmented and low-fidelity data, costs once it is in production. On the breach side, IBM's 2025 Cost of a Data Breach report put the average breach involving data spread across multiple environments at $5.05 million, **the premium an organization pays when telemetry is scattered rather than unified and an incident has to be reconstructed across silos**.

On the operating side, the 2025 SANS Detection and Response Survey found 73% of teams naming false positives their top detection challenge, with more than 60% hitting them frequently or very frequently, **which is the analyst time that context-poor data burns every day before any breach occurs**. One figure is what fragmented data costs when an attack lands. The other is what it costs continuously, whether or not anything is wrong.

The mechanism behind those numbers is simple. An agent handed raw, fragmented, unenriched telemetry inherits every flaw in that telemetry, then acts on it at machine speed. It reaches wrong conclusions faster than the analyst it was meant to relieve. An agent handed normalized, deduplicated, intelligence-fused data starts from context instead of a cold lookup. The difference between those two agents isn't the model, but the fidelity of what you feed it.

THE FOUR DIMENSIONS OF DATA FIDELITY

Data fidelity is telemetry that is correct, contextual, and traceable, ready for a person or an agent to act on without a second lookup. It has four measurable dimensions.

1. COMPLETENESS

Every relevant source is present, retained at full fidelity, and searchable for as long as an investigation might reach back. Sampled logs and data aged out of hot storage are gaps an attacker lives in, and gaps a model fills with guesses.

2. CONSISTENCY

Every source normalizes to one schema, so a login event from an identity provider and a login event from a cloud console carry the same fields and mean the same thing. Open standards such as OCSF make this a property of ingest rather than a per-source engineering project.

3. ENRICHMENT

Vetted intelligence and environmental context are fused into each event as it lands, and the enrichment carries a confidence score. Weak or unscored enrichment is worse than none, because it invites automated action on a shaky basis.

4. PROVENANCE

Every event traces back to its source, every enrichment is attributable, and every action taken on the data is logged. Provenance is what lets a CISO defend an autonomous decision after the fact, and what lets an auditor reconstruct it.

Correct, contextual, and traceable is the bar. Data that clears it costs less to store, less to investigate, and less to act on.

THE AGENTIC SOC STACK THAT LOWERS SPEND

An agentic SOC reduces spend by moving normalization, deduplication, enrichment, and long-term retention off analyst hands and off legacy SIEM licensing, onto one governed data layer that sits beneath the stack you already run. No rip-and-replace, and no new single-vendor dependency. The architecture has three layers, and each one takes cost out at a different point.

THE UNIFIED DATA LAYER

This layer ingests everything, normalizes it to a common schema, deduplicates and strips low-value events at ingest, and splits storage between a short hot window in the SIEM and a long-term searchable lake. The cost effects stack up:

- You stop paying SIEM ingest and retention on duplicate and low-value data, because the noise is removed before it lands.
- Full-fidelity, long-term retention stops carrying SIEM-tier pricing, because history lives in the lake instead of the SIEM.
- Compute and storage decouple, so queries draw resources on demand and there is no always-on bill for infrastructure sitting warm between investigations.

THE INTELLIGENCE LAYER

This layer fuses vetted intelligence and context into every event as it lands, once, instead of an analyst querying a threat intel platform case by case, hours later. Two cost effects follow. Enrichment and scoring at ingest mean fewer false positives ever reach a person, which is where the largest recurring SOC cost sits. And a single normalized schema lets a team write a detection once and run it everywhere, with no rule rebuilds per source.

THE GOVERNED AGENTIC LAYER

This layer puts agents to work on high-fidelity data while analysts supervise and decide. IBM's 2025 report found that organizations using AI and automation extensively contained breaches 80 days faster and paid $1.9 million less per incident, and that only 32% of organizations were doing it. Freed triage becomes capacity you already pay for, redirected from clearing false positives to hunting.

The reason this layer is worth building at all is speed. CrowdStrike's 2026 Global Threat Report put the average eCrime breakout time at 29 minutes, with the fastest observed breakout at 27 seconds. IBM's data shows breaches that run past 200 days cost $5.01 million against $3.87 million when caught sooner. Human-speed triage can't close a 29-minute gap, but high-fidelity data under supervised agents can narrow it, and the dwell-time reduction can be a direct cost reduction.

QUALITY GATES AND AUDITABILITY AT INGEST

You gate quality at ingest, before data becomes an alert or an action, and you make every event and every action traceable end to end. Quality checked after an incident is forensics. Quality checked at ingest is prevention, and it is far cheaper.

Four gates map to the four fidelity dimensions:

- **Schema validation at ingest.** Malformed or unmapped records are caught and routed rather than silently dropped or passed through. A field that arrives as the wrong type does not reach an agent as a fact.
- **Deduplication thresholds.** Redundant events collapse before storage, with the reduction rate tracked so a drift in noise shows up as a number, not a surprise on the bill.
- **Enrichment confidence scoring.** Context carries a score, and only high-confidence enrichment is allowed to promote an event toward automated action. Low-confidence context stays advisory.
- **Provenance tagging.** Every record is stamped with its source and every enrichment with its origin, so lineage travels with the data instead of living in someone's memory.

Auditability is the other half. Source lineage on every event, attribution on every enrichment, and a complete log on every autonomous action give you three things a regulated CISO needs: the ability to explain a decision, the ability to reverse it, and compliance evidence produced from operational data rather than assembled by hand in a parallel process. That combination is what lets a security leader approve supervised autonomy without staking a career on a black box.

BASELINE METRICS THAT PROVE THE SAVINGS

Capture a baseline on a small set of fidelity and cost measures before you change anything. The set below maps to the four dimensions plus the cost each one drives.

DIMENSION	BASELINE METRIC	WHAT IT TELLS YOU
Completeness	% of in-scope sources onboarded; full-fidelity retention window (days)	Whether investigations run on complete data or a sample
Consistency	% of telemetry normalized to a common schema; detections maintained per source	Whether one schema is replacing per-source rule maintenance
Enrichment quality	% of events enriched at ingest; duplicate and noise reduction rate; false-positive rate	Whether context arrives with the data and how much noise is removed before a person sees it
Provenance	% of events with full source lineage; % of autonomous actions with a complete audit trail	Whether decisions are defensible and reversible
Cost effect	Cost per retained GB; analyst hours on triage; mean time to detect and contain; alerts per analyst per shift	Whether fidelity is translating into lower spend and faster response

Track these before and after each layer goes in. A completeness figure that rises while cost per retained GB falls is the hot-and-cold split working. A false-positive rate that drops while alerts per analyst fall is enrichment doing its job. The point of the baseline is to attribute the savings to a cause, so the next budget conversation runs on your numbers.

Find out why high-fidelity data is foundational to successful, autonomous, agentic SOC in our guide Why High-Fidelity Data Decides What Your Agentic SOC Can Do.