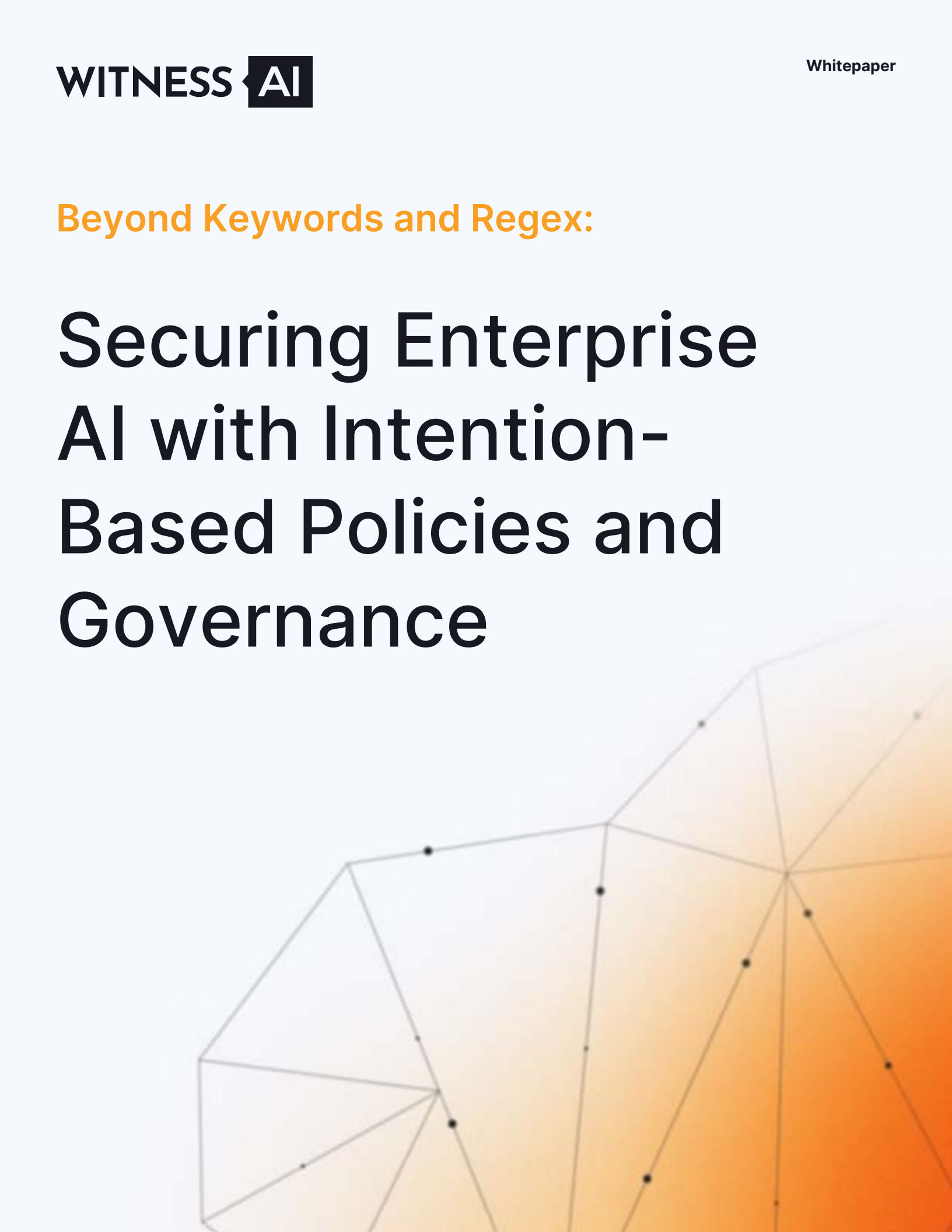
WITNESS AI

Whitepaper

Beyond Keywords and Regex:

Securing Enterprise AI with Intention-Based Policies and Governance



The Enterprise AI Governance Imperative:

Generative AI is a critical inflection point for the enterprise, fundamentally changing how data is created, shared, and used. This new paradigm of unstructured data, conversational workflows, and the rise of unsanctioned shadow AI renders traditional security models obsolete. Legacy Data Loss Prevention (DLP), a technology built on static keywords and brittle regular expressions (regex), can no longer protect the modern organization.

This leaves enterprises with a false choice: block GenAI and sacrifice innovation, or permit its use and risk catastrophic data loss, intellectual property theft, and regulatory non-compliance. The root of the problem is the inability of legacy systems to understand **context and user intent**—the *why* behind an action, not just the *what* of the data.

For CISOs and executive leaders, the path forward demands a strategic pivot from outdated, content-based data protection to a modern framework of intent-based governance. This approach leverages advanced modeling to classify the purpose of every user interaction with AI, enabling dynamic, context-aware security policies. These policies are not only more effective at mitigating risk but are also business-enabling, designed to foster innovation while maintaining control.

The Limitations of Traditional DLP in the AI Era

Traditional Data Loss Prevention (DLP) systems were designed for a different era. In the age of AI, their architecture is built on a foundation unsuited for the dynamic, conversational, and context-driven nature of modern work.

The Flawed Foundation: A Legacy Governance Model Built on Brittle Rules

At its core, traditional DLP is a blunt instrument. Its effectiveness depends on a library of predefined, static rules to identify sensitive data. This model was conceived for a world of highly structured data, like credit card numbers, residing in predictable locations.

The inadequacy of this model begins with its primary tools: pattern matching and basic data classification. While useful for finding exact strings, a regular expression has no understanding of **user intent**, making it unable to differentiate a credit card number from a harmless product ID. Similarly, while data classification can apply a "*Confidential*" tag to a file, that protection vanishes the moment an employee copies text from the document and pastes it into a GenAI prompt. This lack of intelligence generates a high volume of false positives, burdening security teams with unactionable alerts.

This fragility becomes a critical vulnerability when dealing with high-entropy strings like API keys. A security team might write a regex for `$api_key = '...'`, but this rule is brittle. It will miss the identical key if a developer uses a different variable, such as `$api_token = '...'`. Creating a comprehensive regex library to cover all permutations is an impractical, unending task that fails to address the core problem: **understanding the data in its context of use**.

This technical limitation results in a system plagued by both false positives and false negatives. Overly broad rules block legitimate business, while narrow rules miss the contextual data leaks that pose the greatest risk. This cycle actively degrades an organization's security posture, as desensitized teams are more likely to overlook a genuine threat and frustrated employees seek workarounds like shadow IT.

The Context Blindspot: Unstructured Data and the Rise of GenAI

The modern enterprise runs on unstructured data. The Enterprise Strategy Group (ESG) reports that for over 90% of companies, unstructured data—contracts, source code, design documents—makes up more than half of their total data volume. The sensitivity of this data is defined by its context, not a repeatable pattern. Traditional DLP is almost entirely blind to this reality.

Legacy systems cannot differentiate between a public marketing deck and a sensitive M&A roadmap that happens to share a project name. This is why an alarming amount of sensitive unstructured data remains undiscovered and unprotected. Generative AI amplifies this problem exponentially. Employees now use AI for nuanced, conversational tasks: summarizing sensitive documents, debugging proprietary code, and drafting strategic plans. A keyword filter looking for **“earnings”** cannot possibly understand the **intent** and inherent risk in a prompt like, **"***Summarize this draft of our Q3 report and highlight any potential legal risks.***"**

GenAI shatters the old security model. The traditional security perimeter and structured data flows are gone. The perimeter has effectively shifted to any employee's web browser, and the data being exchanged is no longer a discrete file but a snippet of code or a confidential strategy pasted into a prompt. Legacy DLP, designed to inspect file transfers, is architecturally blind to this interaction. The challenge is not to adapt DLP for AI; it is to recognize that the fundamental assumptions upon which DLP was built are now invalid.

Case Study: Intellectual Property Loss in the AI Era



The most damaging data leaks often originate not from malicious actors, but from well-intentioned employees using "shadow AI" to be more productive.

A common scenario in the pharmaceutical industry illustrates the fundamental failure of traditional DLP. An intern at a pharmaceutical company is tasked with summarizing recent clinical trial results. The document contains highly confidential intellectual property, including the chemical formula and trial data for a new drug, "***Azumabril***." Pressed for time, the intern pastes the raw text into a public GenAI tool with the prompt, "***Please summarize these research results***."

In that moment, a cascade of security failures occurs instantly:

01

Content-Based DLP is Ineffective:

The pasted text is unstructured scientific language. It contains no patterns a regex-based system is designed to find. The data is sensitive because of its ***context***, not its format.

02

Metadata is Stripped:

The original document's "***Confidential***" tag is lost the moment the text is copied. The DLP tool, relying on file classification, is now blind to the content's sensitivity.

03

The Action is Permitted:

A legacy web proxy sees only legitimate, encrypted HTTPS traffic to a public website and allows it to proceed, unable to inspect the prompt's content.

The result is a catastrophic loss of intellectual property. The formula for Azumabril now resides on a third-party server, outside the company's control and potentially integrated into the AI's training data. The user's intent was productivity; the outcome was a crippling competitive disadvantage. This highlights the critical need for a security solution that understands context and intent.

Case Study: PCI DSS Compliance Failure in the AI Era

For any organization that processes payments, maintaining compliance with the Payment Card Industry Data Security Standard (PCI DSS) is non-negotiable. Requirement 10, in particular, mandates diligent tracking and monitoring of all access to network resources and cardholder data. Traditionally, enterprises rely on SIEM platforms and DLP tools with regex-based rules to meet this requirement. However, this approach is critically flawed when interacting with GenAI.

Consider a customer service agent at an e-commerce company using an AI-powered internal chatbot to resolve a customer issue. The agent's goal is to be helpful and efficient.

01 Scenario 1: The False Positive and Alert Fatigue

The agent types into the chatbot: "***Please pull up the resolution notes for support ticket #4992-8812-5409-1138.***"

Legacy System Failure: A legacy DLP tool, configured with a regex to detect 16-digit Primary Account Numbers (PANs), immediately fires a high-priority alert. The pattern matches, but the data is not a credit card number; it is a harmless support ticket ID. The security operations team must now stop their work, investigate the alert, validate the data, and ultimately close the ticket as a false positive. This single event, multiplied hundreds or thousands of times a day across the enterprise, erodes the effectiveness of the security team.

02 Scenario 2: The False Negative and a Silent Breach

Later, the same agent handles a different issue. They type: "***The customer is asking for a refund. The last four digits of their Visa are 8745 and the order was placed on July 14th. Can you confirm the eligibility based on the attached policy document?***". They paste a snippet of the return policy into the prompt.

Legacy System Failure: The regex-based tool sees no 16-digit PAN. It sees no keywords like "confidential" in the generic policy text. No alert is generated. However, a serious compliance violation has just occurred. Sensitive, albeit partial, cardholder data has been sent to a large language model alongside other contextual information. This data may be logged, stored, and used for training by the third-party AI provider, creating an un-auditable and un-retrievable copy of customer data outside the defined Cardholder Data Environment (CDE). This is a silent, unmonitored compliance failure that a traditional tool is architecturally blind to.

The Paradigm Shift: From *What* to *Why* with Intent-Based Classification

To move beyond pattern matching, enterprise security must evolve. It must shift its focus from the data itself to the purpose behind the data's use. This demands a new technological foundation capable of understanding nuance, context, and goals: ***intent-based classification***.

Defining User Intent: The Next Generation of Data Understanding

Intent-based classification is a sophisticated form of machine learning that determines a user's objective: the "why" behind an action or query.

This capability is far more advanced than other common NLP tasks:

Intent vs. Sentiment Analysis

Sentiment analysis identifies emotion (positive, negative). Intent classification identifies purpose. A customer expressing negative sentiment ("I'm very unhappy") has the intent to `request_a_refund`. A security policy must act on the intent, not the emotion.

Intent vs. Entity Recognition:

Named Entity Recognition (NER) extracts specific data points (names, dates, locations). Intent classification identifies the overarching goal that connects them. For the query "Book a flight to New York for tomorrow," NER identifies "New York" and "tomorrow." Intent classification identifies the core goal: `book_flight`. Understanding entities without intent is incomplete.

The Power of Context: Why Intent-Based AI Delivers Superior Accuracy

An intent-based governance system vastly outperforms a regex-based one because it understands semantics, context, and variation. This results in superior accuracy and a reduction in the false positives that plague legacy DLP.

01

For Unstructured Data

An intent classifier can distinguish between a low-risk prompt like "Draft a marketing email" and a high-risk one like "Summarize the attached M&A term sheet." While both may contain the same product name, the classifier understands the difference in purpose and sensitivity, enabling a precise, risk-appropriate response.

02

For High-Entropy Secrets:

Where a regex search for `api_key` fails, an intent-based model analyzes surrounding context. It recognizes the code structure, function calls, and the high-entropy nature of the string itself to determine it is a credential, regardless of the variable name. This reduces both false positives and false negatives.

03

For Emerging Threats:

Advanced intent models can employ zero-shot classification, allowing them to identify and categorize new, previously unseen intents without explicit retraining. This is a crucial advantage, enabling the governance system to keep pace with the rapid emergence of new "shadow AI" tools and use cases.

Building Dynamic, Frictionless Policy on Intent

The ultimate value of understanding intent is the ability to build a superior class of security policies that are intelligent, adaptive, and aligned with business needs. The paradigm shifts from a crude "block or allow" posture to a sophisticated "govern" posture.

Instead of a policy like "***Block any document containing 'confidential***'," an intent-based engine enables far more nuanced controls.

User Query:

"Summarize this financial document for our internal audit team."

Detected Intent:

`summarize_financial_document_internal`

Intent-Based Policy:

Allow the action but enforce context-aware guardrails: log the action for audit, redact PII from the output, and route the query to a secure, internal LLM.

This intent-driven approach yields transformative benefits:

Reduces Business Friction:

By understanding legitimate context, the system avoids blocking productive work, removing the primary incentive for employees to seek risky shadow IT workarounds.

Automates and Scales Governance:

Policies are tied to durable business purposes, not brittle rules, making the entire governance framework more resilient and easier to manage.

Enables Safe Innovation:

Most importantly, intent-based governance provides the intelligent guardrails that allow organizations to fully and safely embrace the benefits of Generative AI.

Traditional DLP vs Intent-Based Policy and Security

Feature	Traditional DLP (Regex & Keyword-Based)	Intent-Based Policy and Security
Detection Method	Static pattern matching	Semantic and contextual analysis
Context Awareness	Blind. Cannot differentiate legitimate vs. malicious use.	High. Understands user goals and business context.
Accuracy	Low. Prone to high rates of false positives & negatives.	High. Drastically reduces false positives; detects contextual risks.
Unstructured Data	Ineffective. Blind to the context defining sensitivity.	Highly Effective. Purpose-built for unstructured data.
Management	High overhead. Requires constant manual rule tuning.	Low overhead. Automated policies tied to business intent.
GenAI Readiness	Poor. Architecturally unsuited for monitoring prompts.	Excellent. Natively governs conversational AI interactions.

How WitnessAI Operationalizes Intent-based Policies with Intelligence AI Routing

Having established the failure of legacy models and the superiority of an intent-based approach, the question becomes one of execution. ***This is the role of WitnessAI***.

WitnessAI is a secure AI enablement platform providing complete, agentless, network-level visibility across thousands of AI applications, including shadow AI. Our platform understands employee intent to build context-aware policies that stop data leakage while enabling productivity. Intelligent AI routing delivers control by optimizing requests based on risk, cost, or purpose. With a single-tenant architecture and customer-controlled encryption, we make AI security a business enabler. WitnessAI bridges the gap between the promise of intent-based governance and its tangible, enterprise-ready implementation.

WitnessAI in Action: The Core Pillars of Intent-Driven AI Security

WitnessAI operationalizes intent through a unified platform designed to address the most pressing AI governance challenges, including comprehensive visibility, granular control, and strategic optimization.

Pillar 1: Intent-Based Governance — Achieving True Visibility and Control

The core of WitnessAI is a proprietary intention classification model that understands the purpose behind user prompts. An organization can feed the system any combination of words or phrases, and it will intelligently categorize what the user is trying to accomplish. This allows for the creation of dynamic policies that govern specific topics or intents.



Example: Enterprise Intellectual Property Protection

Let's revisit the pharmaceutical intern. When the intern submits the prompt, "Please summarize these research results," along with confidential data for the drug "Azumabril," WitnessAI does not rely on simple keywords.

01

Detect High-Risk Intent:

The model analyzes the prompt and the accompanying data. It recognizes the request, based on the context of clinical trial information, falls under the high-risk intent of `drug_research`.

02

Enforce Context-Aware Policy:

This classification triggers a pre-defined policy. Instead of a blunt block, the query is automatically and seamlessly redirected from the public GenAI tool to a secure, private, and company-sanctioned internal model.

03

Coach the User:

The user is notified in real-time: "This topic contains proprietary research. Your prompt has been securely redirected to our internal model, 'ApexGPT'."

This action protects intellectual property while allowing the intern to complete their task. It transforms a security event into a moment of automated coaching, reinforcing secure behavior without hindering productivity.

Pillar 2: Intelligent AI Routing — The Dual Imperative of Security and Efficiency



Simplistic, block-only security policies for GenAI create a significant competitive disadvantage. A policy that treats a public information request the same as an analysis of proprietary source code is fundamentally misaligned with business reality.

WitnessAI's Intelligent AI Routing operationalizes intent by moving beyond a simple "block or allow" response. Based on the classified purpose of an interaction, WitnessAI can permit safe queries, present coaching warnings, or execute its most powerful function: dynamically routing the interaction to the most appropriate AI model based on data sensitivity, financial cost, and task requirements.

This provides a direct mechanism to safeguard critical assets while controlling operational spend. The platform can automatically identify interactions involving intellectual property and direct them exclusively to secure, private AI models. Concurrently, it optimizes the ROI of AI by routing routine, low-risk queries to less expensive models, reserving high-performance models for complex tasks. By aligning security controls with business objectives, Intelligent AI Routing transforms governance from a source of friction into a driver of innovation.

Case Studies in Action: Securing the Enterprise Across Every Vertical

The true measure of a security platform is its performance against real-world, high-stakes scenarios. WitnessAI's intent-based governance can deliver tangible value across diverse industries by moving beyond simple pattern matching to understand the specific risks inherent in each interaction. The following cases illustrate how this approach protects intellectual property, ensures regulatory compliance, and secures critical national infrastructure.

Mastering PCI DSS Compliance: The E-commerce Company

For any organization processing payments, maintaining PCI DSS compliance is non-negotiable. Legacy tools create a dangerous mix of alert fatigue from false positives and silent breaches from false negatives.

The Scenario: A customer service agent uses an AI chatbot to resolve two different issues.

Case 1 (False Positive):

The agent types, "***Pull up notes for support ticket #4992-8812-5409-1138.***" A legacy tool's regex fires a false high-priority alert on the 16-digit non-sensitive number, wasting the security team's time.

Case 2 (False Negative):

The agent types, "***Customer's Visa ends in 8745. Can you confirm refund eligibility?***" A legacy tool sees no full credit card number and generates no alert, allowing partial cardholder data to be silently exfiltrated to an untrusted AI model, violating compliance.

The WitnessAI Solution:

WitnessAI prevents both failures by understanding intent.

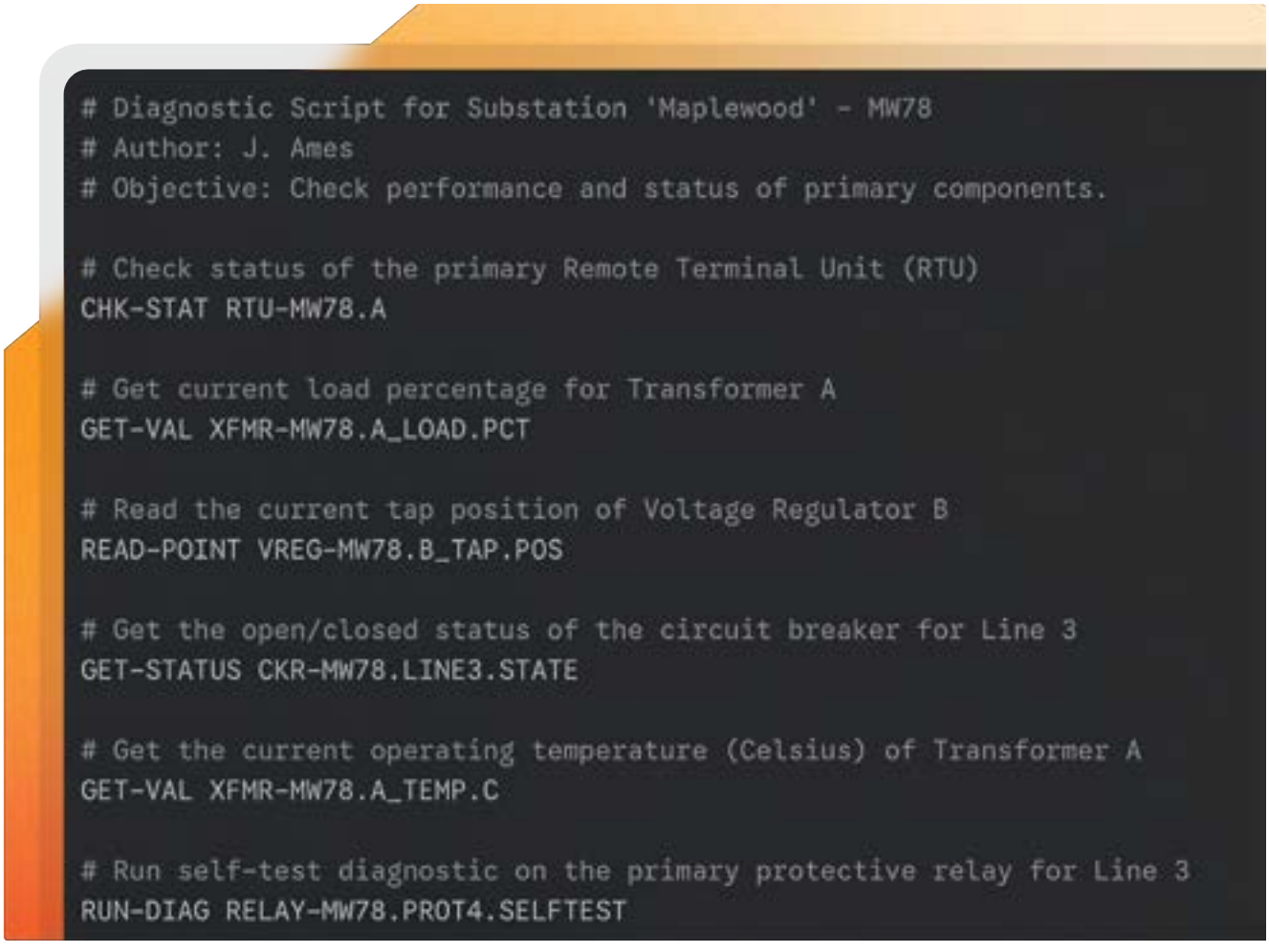
1. In the first case, it classifies the intent as `lookup_support_ticket`. Recognizing the number is a harmless ID, it allows the action without a false alert, keeping the security team focused.
2. In the second case, it classifies the intent as `process_customer_refund_request`. This is immediately flagged as a high-risk action. The policy blocks the query or routes it to a secure, PCI-compliant internal model, preventing the compliance breach before it happens.

This demonstrates a crucial evolution from hunting for data patterns to governing user intent, which is the only way to achieve genuine, auditable control.

Securing Critical Infrastructure: The Utility Company

For entities governed by NERC CIP standards, a cyber event can have catastrophic physical consequences. Security must prevent not just known-bad commands but also the subtle intelligence leaks that enable future attacks.

The Scenario: An engineer at a utility pastes a diagnostic script for a substation into a public GenAI tool, asking for optimization suggestions and naming the specific substation, "*Maplewood*." The script contains only legitimate commands.



```
# Diagnostic Script for Substation 'Maplewood' - MW78
# Author: J. Ames
# Objective: Check performance and status of primary components.

# Check status of the primary Remote Terminal Unit (RTU)
CHK-STAT RTU-MW78.A

# Get current load percentage for Transformer A
GET-VAL XFMR-MW78.A_LOAD.PCT

# Read the current tap position of Voltage Regulator B
READ-POINT VREG-MW78.B_TAP.POS

# Get the open/closed status of the circuit breaker for Line 3
GET-STATUS CKR-MW78.LINE3.STATE

# Get the current operating temperature (Celsius) of Transformer A
GET-VAL XFMR-MW78.A_TEMP.C

# Run self-test diagnostic on the primary protective relay for Line 3
RUN-DIAG RELAY-MW78.PROT4.SELFTEST
```

Why These Commands Bypass Traditional DLP

A legacy security tool would analyze this script and permit it because:

No Malicious Keywords:

The verbs used (`CHK-STAT`, `GET-VAL`, `READ-POINT`, `RUN-DIAG`) are all diagnostic or read-only commands. They don't contain high-alert triggers like `disable` or `trip`.

Benign Arguments:

The device names (`RTU-MW78.A`, `XFMR-MW78.A_LOAD.PCT`) are just alphanumeric strings to a regex engine. It has no context to know that "XFMR-MW78" refers to a multi-ton transformer critical to the power grid.

Focus on Diagnostics:

The presence of commands like `RUN-DIAG` and `SELFTEST` would likely be whitelisted in any environment to avoid flooding the security team with false positives from routine maintenance activities.

The WitnessAI Solution:

WitnessAI analyzes the holistic context and can prevent a potential leak of sensitive critical infrastructure information to the public GenAI tool.

1. **High-Risk Intent Classified:** The platform's model recognizes the dangerous combination of OT/SCADA commands and the named entity of a critical infrastructure asset. It classifies the intent as `disclose_OT_script`, a critical risk.
2. **NERC CIP Policy Enforced:** A pre-defined NERC CIP AI intent policy is triggered instantly. The query is blocked from leaving the utility's Electronic Security Perimeter (ESP), and the engineer is coached in real-time to use an approved, air-gapped internal AI model.

For critical infrastructure, security cannot be about a blacklist of forbidden words. It must be about understanding the purpose of every interaction to prevent the silent intelligence leaks that precede a grid-destabilizing event.

Conclusion

The rapid integration of Generative AI into enterprise workflows hasn't just introduced a new tool; it has fundamentally redrawn the boundaries of corporate security. The era of defending a clearly defined network perimeter with pattern-matching tools is over. **Legacy DLP, with its reliance on rigid keywords and brittle regex, is incapable of governing the fluid, conversational, and context-dependent nature of AI-driven work.** Continuing to invest in this outdated paradigm offers diminishing returns in the face of aggressive AI deployments.

The strategic imperative for today's CISO is to pivot from a reactive posture of blocking data to a proactive strategy of governing intent. By shifting the security focus from what the data is to why it is being used, organizations can finally resolve the false choice between innovation and security. An intent-based framework allows for the creation of intelligent, adaptive security policies that are not only more effective at mitigating risk but are also intrinsically aligned with business objectives, reducing friction and empowering employees to innovate safely.

WitnessAI provides the critical infrastructure to operationalize this new paradigm. Through its unified platform for **intent-based governance and intelligent AI routing**, WitnessAI delivers the visibility to see every AI interaction, the intelligence to understand its purpose, and the control to enforce security that is both robust and business-aware.

In the AI revolution, your most sensitive data no longer lives exclusively in files and databases—it lives in the prompts, queries, and conversations of your employees. **Intent is the new perimeter, and WitnessAI is the platform built to defend it.**

WITNESS AI

About WitnessAI

WitnessAI enables safe and effective adoption of enterprise AI, through security and governance guardrails for public and private LLMs. The WitnessAI Secure AI Enablement Platform provides visibility of employee AI use, control of that use via AI-oriented policy, and protection of that use via data and topic security.

Learn more at witness.ai.

