WHITEPAPER

Securing AI on Amazon Bedrock with Cyera

Enterprise data governance for generative and agentic AI — a reference architecture.

cyera | aws

The value of enterprise AI scales with the sensitivity of the data it touches. That's the fundamental tension. Cyera AI Guardian resolves it — providing continuous classification, posture management, and runtime policy enforcement across Bedrock knowledge bases, custom models, and agents. The result: organizations move from pilot to production without choosing between velocity and governance.

Intended audience

CISOs, AI/ML platform engineering leads, and cloud security architects governing generative and agentic AI workloads on Amazon Bedrock. If you're the person who has to sign off before an agent goes to production — this is for you.

Executive Summary

The governance imperative

Every organization we work with tells us the same thing: the AI technology is ready. The governance isn't.

Executive-sponsored AI initiatives stall not because models underperform, but because security teams lack three things: a definitive inventory of what AI workloads exist, lineage showing which data those workloads access, and evidence that runtime controls actually enforce policy at inference time. Without all three, CISO's are left approving what they cannot see, trace, or enforce.

And the workloads that deliver the most value — RAG over proprietary data, fine-tuning on domain-specific corpora, autonomous agents executing multi-step workflows — are precisely the ones that require access to an organization's most sensitive assets. Customer records. Source code. Financial models. Clinical data. The data that makes your business defensible is the same data that makes AI useful. There's no way around this.

Agentic AI makes the problem structural. These aren't humans making requests — they're non-human principals invoking APIs, traversing data stores, and executing workflows at machine speed. You cannot govern them with quarterly access reviews and ticket-based approvals. You need controls that operate continuously, at the same speed the agents do.

Amazon Bedrock provides the managed infrastructure for all of this: foundation model access, knowledge base orchestration, fine-tuning, and agent runtime. It's where serious enterprise AI gets built. Cyera AI Guardian extends Bedrock's native security capabilities with the data-layer governance that turns "we're experimenting" into "we're in production." This paper highlights how Cyera's AI Guardian secures Bedrock knowledge bases, custom models, and agents.

The numbers driving urgency

82%

Of breaches now involve cloud-stored data. The same S3 buckets and databases feeding your knowledge bases and training jobs.

39%

Of those breaches span multiple environments (IBM, 2024), exactly the observability gaps that autonomous agents traverse without hesitation.

$4.9M

The average cost of a breach, with costs climbing sharply when data is scattered

Legacy DLP and CSPM tools were designed for a world where humans initiated requests and waited for responses. They assume a request-response cadence that agents simply don't follow. Governing AI on Bedrock requires three capabilities these tools lack:

1. Continuous, AI-native classification that operates at ingestion speed, not batch scans on a weekend schedule.
2. Identity-aware lineage from data source through embedding, retrieval, and completion, so you can trace exactly how a piece of sensitive data ended up in a model response.
3. Inline runtime enforcement that evaluates every prompt, tool call, and completion against policy before it reaches the end user. Not after. Before.

The Players

Amazon Bedrock + Cyera AI Guardian

Amazon Bedrock

Managed infrastructure for foundation model access, orchestration, and customization

Unified API access to models from AI21 Labs, Anthropic, Cohere, Meta, Mistral AI, Stability AI, and Amazon — plus managed services for knowledge bases, fine-tuning, and agent orchestration. Bedrock eliminates the undifferentiated infrastructure work,compressing months of LLM engineering into minutes.

Cyera AI Guardian

Data classification, AI security posture management, and runtime enforcement

Cyera's AI Security Platform continuously discovers and classifies sensitive data across cloud, SaaS, and on-premises environments. When that data serves as input to a model, populates a knowledge base, or gets accessed by an agent, Cyera AI Guardian enforces governance at the data layer, regardless of which path was used to reach it.

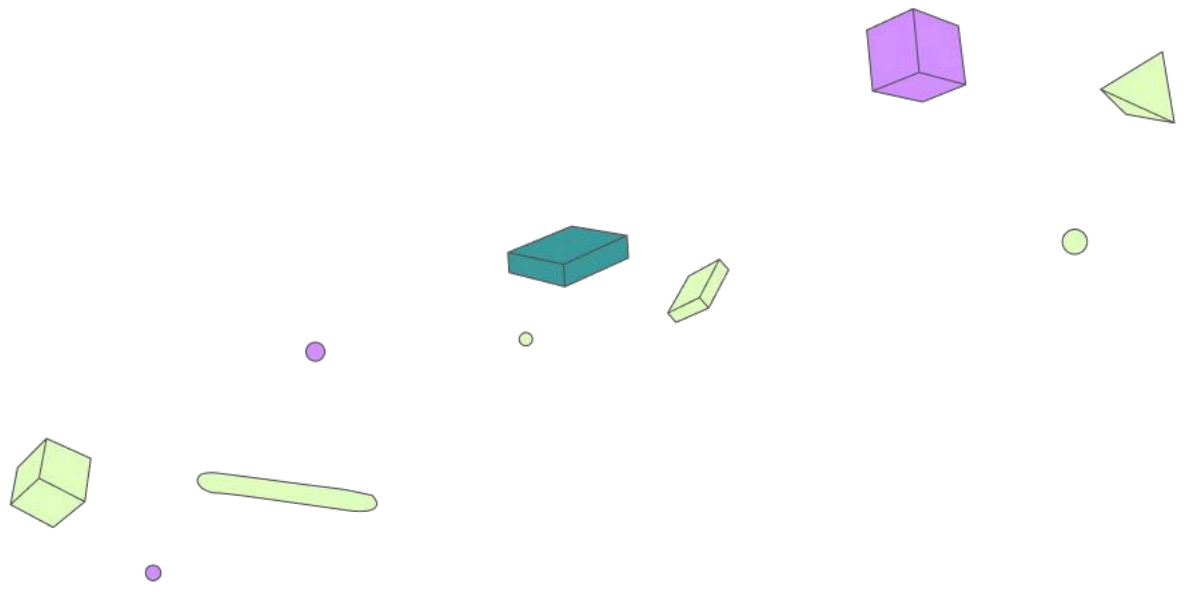
AI Guardian combines two capabilities: AI-SPM for continuous inventory, configuration assessment, and shadow AI detection across all Bedrock assets, and AI Runtime Protection for inline enforcement on prompts, retrieved context, tool invocations, and completions.

Patterns that Concentrate Risk

Where data meets the model — and where governance breaks

Not all Bedrock deployments carry equal risk. Risk concentrates at three specific patterns, and they happen to be the three patterns every enterprise wants to deploy first. Cyera AI Guardian maps controls to each, aligned to the OWASP Top 10 for LLM Applications at the architectural layer where each risk actually materializes.

Deployment Pattern	What it does	OWASP LLM risks
Knowledge Bases (RAG)	Retrieval-augmented context at inference time. Chunked, embedded documents in S3 and vector stores (OpenSearch Serverless, Aurora, Pinecone).	LLM06 (Sensitive Info Disclosure), LLM05 (Vector Weaknesses), LLM03 (Data Poisoning)
Model Customization	Adapts model weights via continued pre-training or supervised fine-tuning on org-specific corpora in S3.	LLM03 (Data Poisoning), LLM06 (Sensitive Info Disclosure), LLM10 (Model Theft)
Custom Bedrock Agents	Orchestrates models, KBs, action groups (Lambda), and MCP tools to execute multi-step workflows autonomously.	LLM01 (Prompt Injection), LLM07 (System Prompt Leakage), LLM08 (Excessive Agency)



Scenario 1

Bedrock Knowledge Bases — Retrieval-Augmented Generation

Bedrock Knowledge Bases supply foundation models and agents with organization-specific context at inference time — eliminating the cost, latency, and data-permanence risks of fine-tuning. The architectural tradeoff: every ingested document expands the retrieval surface available to any principal with invoke-model permissions, and every unvalidated document becomes a potential vector for indirect prompt injection or retrieval manipulation.

Where the risk hides

Sensitive Information Disclosure (OWASP LLM06)

A knowledge base indexes documents containing data that exceeds the requesting principal's authorization scope. The model returns this data in its completion, bypassing application-layer access controls that were never designed to govern retrieval-augmented responses.

A development or staging agent is configured with a production S3 data source, exposing production-grade sensitive data to non-production IAM roles, VPCs, and logging configurations that lack equivalent controls.

An adversary introduces a crafted document into the knowledge base source bucket — exploiting write access to S3 — that contains embedded instructions designed to manipulate retrieval ranking or override system-prompt directives (indirect prompt injection).

Knowledge bases provisioned outside governed deployment pipelines (e.g., via console or ad-hoc CloudFormation) connect to production data stores without security review, tagging, or guardrail configuration — constituting shadow AI assets with production-level data access.

Vector & Embedding Weaknesses (OWASP LLM05)

An adversary manipulates embedding representations or source content to alter retrieval ranking, causing the model to preferentially surface attacker-controlled context. This enables indirect influence over model completions without direct access to the model or its system prompt.

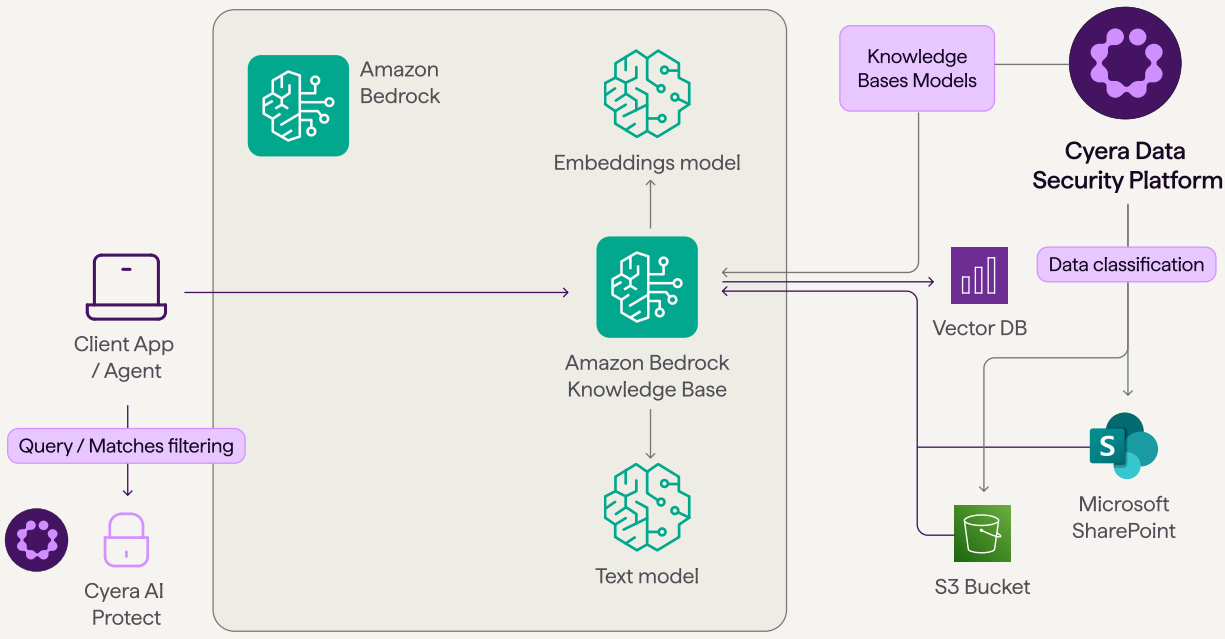
Cyera AI Guardian controls for Knowledge Base deployments

Cyera AI-SPM	Maintains a continuous, automated inventory of all Bedrock Knowledge Bases, associated foundation models, agents, and backing data stores (S3 buckets, vector databases). Each asset is mapped to the sensitive data classifications present in its source documents, providing a real-time sensitivity posture for every RAG deployment.

Identity-aware access mapping	Correlates IAM principals (roles, users, service-linked roles) to Knowledge Base access paths, producing a three-dimensional view: authorized access (who can), intended access (who should per policy), and observed access (who did, per CloudTrail). Surfaces over-permissive grants and dormant access.
Cyera AI Protect	Evaluates queries and retrieved document chunks inline, applying classification-aware policy to redact, mask, or block sensitive content before it is included in the model's context window or returned in the completion to the end user. Enforcement occurs at retrieval time, not after generation, eliminating the risk of sensitive data influencing model reasoning even when the final output is filtered.
Shadow AI detection	Identifies Bedrock Knowledge Bases, agents, and custom models provisioned without Bedrock Guardrails configuration, security tagging, or association with a governed deployment pipeline. Generates prioritized posture findings with remediation guidance, enabling security teams to bring unmanaged assets into compliance before they process sensitive data.

"The accuracy with which Cyera sorted and classified sensitive data was amazing. It made communicating real risk to the business dramatically easier."

- CISO, Healthcare & Biotech — Gartner Peer Insights



Reference Architecture: Cyera AI-SPM Classifies The Data Behind Bedrock Knowledge Bases, While Cyera AI Protect Filters Queries And Retrieves Matches In Real Time.

Scenario 2

Model Customization — Fine-Tuning and Continued Pre-Training

Fine-tuning bakes your organization's vocabulary, style, and decision rules directly into a foundation model, shortening prompts, reducing tokens, and lowering latency. RAG supplies the freshest facts; fine-tuning decides how to answer. Together they're powerful. Together they're also a double exposure path.

Where the risk hides

Training Data Poisoning and Memorization (OWASP LLM03)

A foundation model fine-tuned on datasets containing sensitive records (PII, credentials, proprietary IP) memorizes specific training examples. At inference time, adversarial or coincidental prompts trigger verbatim or near-verbatim recall of memorized sensitive content, a well-documented failure mode in large language models that persists across model families and parameter scales.

An adversary with write access to the S3 training data bucket introduces crafted examples designed to shift model behavior — embedding backdoor triggers, biasing classification outputs, or degrading response quality for specific input patterns. Because fine-tuning jobs consume data from S3 without content-level validation by default, the attack surface is the bucket's IAM write policy.

Custom model training jobs initiated outside governed CI/CD pipelines — via console, SDK, or delegated IAM roles — bypass data review, sensitivity scanning, and approval workflows. The resulting model artifacts inherit the sensitivity classification of their training data but carry no metadata indicating this lineage, creating ungoverned model assets with embedded sensitive content.

How Cyera closes the gap

Cyera AI-SPM	Automatically discovers custom model artifacts (fine-tuned models, continued pre-training outputs) and establishes data lineage from each model back to its training data sources in S3. Surfaces the sensitivity classification of training data alongside the model's deployment status, consumer agents, and access permissions.
Sensitive-data lineage	Identifies training data repositories containing classified sensitive content (PII, PHI, financial data, source code, credentials) and propagates that classification to downstream model artifacts. Enables pre-training governance decisions: gate access to the training job, mask sensitive fields in training data, or exclude high-sensitivity records from the training corpus entirely.
Least-privilege analysis	Identifies IAM roles and principals with write access to training data buckets or CreateModelCustomizationJob permissions that exceed their operational requirements. Recommends scope reductions aligned with the principle of least privilege, reducing the attack surface for training data poisoning.

Cyera AI Protect

Applies inline classification and policy enforcement at inference time on both prompts and model completions. Detects and redacts memorized sensitive content (PII, credentials, proprietary data) in model outputs before delivery to the requesting application or end user. Operates independently of the model's internal safety training, providing a defense-in-depth layer against memorization-based disclosure.

80%

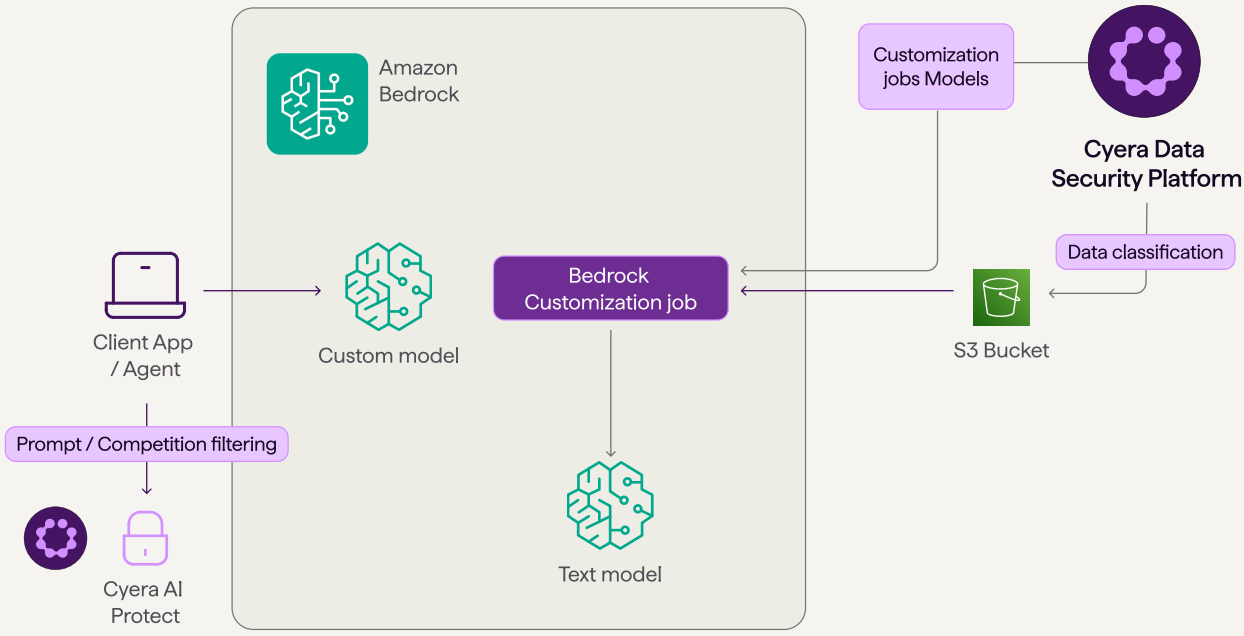
reduction in breach likelihood by remediating high-risk data exposures.

$134K+

in annual savings replacing brittle, rule-based classification.

17 hrs

to classify 28.5 TB of Snowflake data across 1.6B sensitive records.



Reference Architecture: Cyera Classifies The S3 Data Powering Fine-Tuning Jobs And Enforces Prompt / Completion Filtering At Inference Time.

Scenario 3

Custom Bedrock Agents (tools, action groups & MCP).

Bedrock Agents translate natural-language directives into multi-step, auditable workflows — orchestrating foundation model reasoning, action group execution (AWS Lambda), knowledge base retrieval, and external tool integration (MCP servers) within a single managed runtime. This pattern represents the highest-leverage enterprise AI deployment: autonomous, goal-directed, and capable of reading, writing, and acting across systems without human-in-the-loop approval at each step. It also represents the highest-risk pattern, because it compounds the data exposure surfaces of RAG and model customization with agent-specific risks: excessive autonomy, unscoped tool access, and prompt injection that can redirect multi-step execution chains.

Where the risk hides

Sensitive Information Disclosure

A dev/test agent wires into production data and retrieves sensitive records in real time.

Agents connect to unapproved MCP servers, opening data paths outside sanctioned controls.

Tools pull sensitive data without the user's authorization context, scopes, or approvals.

Developers spin up agents outside governance — no review, no logging, no baselines.

Content filters and guardrails are simply not configured.

New action groups or KBs are attached to approved agents without re-review (shadow AI).

Irrelevant data is connected to an agent, violating least privilege by design.

System Prompt Leakage

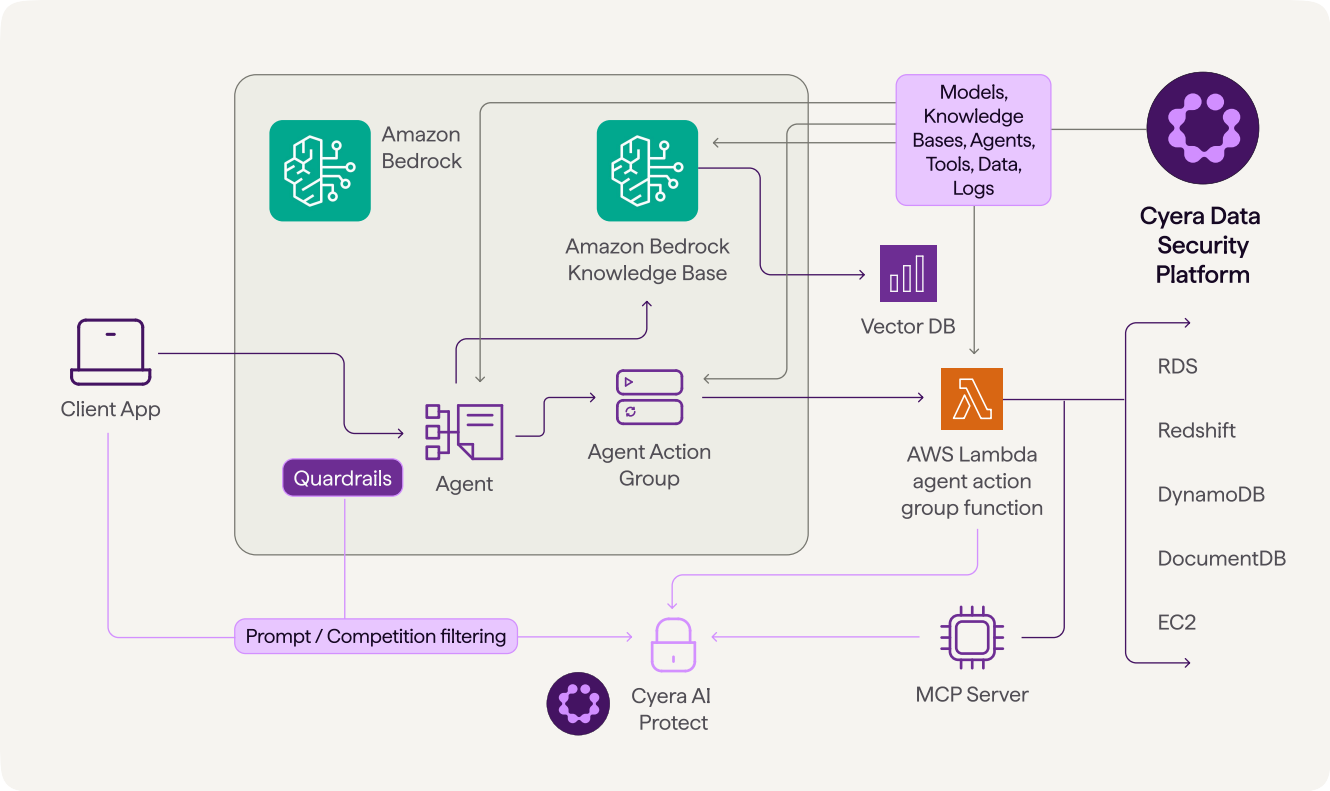
Credentials, API keys, internal endpoint URLs, or guardrail bypass logic embedded in the agent's system prompt are disclosed through adversarial prompt extraction techniques or verbose error/debug responses. Exposed system prompts reveal the agent's operational boundaries, enabling targeted attacks against its weakest controls.

Prompt Injection

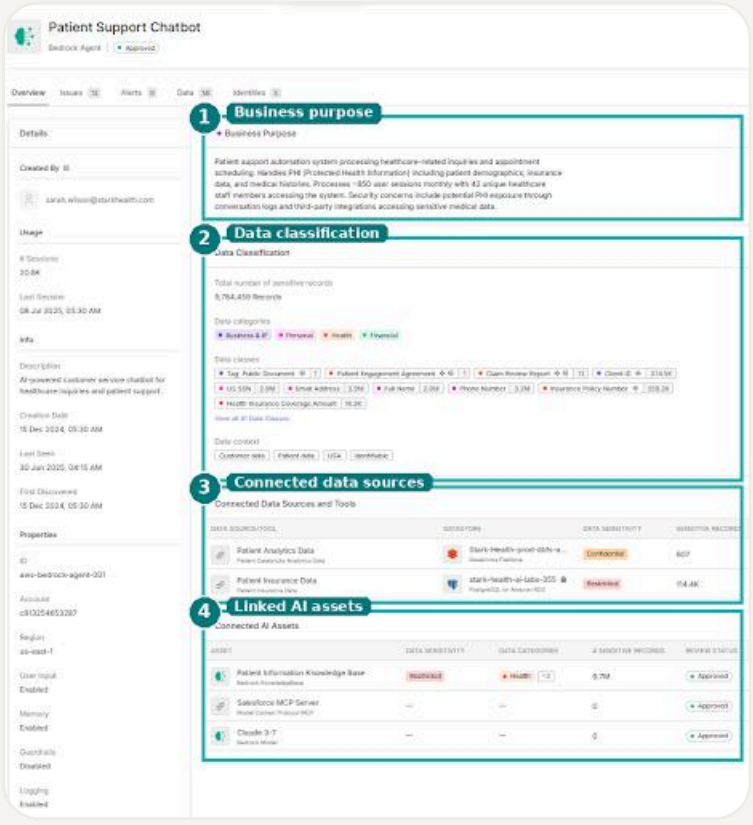
Malicious or crafted inputs, including retrieved content, override system instructions or jailbreak the agent.

How Cyera closes the gap

Cyera AI-SPM	inventories every Bedrock agent and correlates it to the models, KBs, tools, MCP servers, and data stores it can reach.
Cyera AI Protect's API	can be embedded at multiple integration points within agentic apps. On every invocation, it classifies and enforces policy on prompts, completions, KB matches, tool calls, and retrievals. This allows redacting, masking, or blocking in real time.
Audit-ready by default	every decision is audited, reinforcing least-privilege and continuous compliance.



Reference Architecture: Cyera AI Protect Sits At Every Integration Point Of A Bedrock Agent — Prompts, Knowledge-Base Retrievals, MCP Server Calls, And Tool / Lambda Action-Group Invocations.



A Bedrock Agent Inside Cyera AI-SPM — ① Business Purpose, ② Data Classification, ③ Connected Data Sources, And ④ Linked AI Assets (Agents, KBs, MCP Servers, And Models) — All In One View.

Why Cyera for Bedrock

Unified data-centric governance: from asset discovery through runtime enforcement

Conventional AI security approaches instrument the model layer, monitoring prompts and completions in isolation from the underlying data estate. Cyera inverts this model: governance begins at the data layer, where sensitive content is classified, access is mapped, and policy is defined. This then propagates enforcement through every AI asset that touches that data. The following table summarizes the operational outcomes this architecture delivers for Bedrock deployments.

Outcome	Cyera AI Guardian delivers
360° Bedrock visibility	Continuous inventory of models, knowledge bases, and agents, plus the identities, data stores, and tools each one can reach.
Shadow AI eliminated	Unregistered Bedrock agents, custom models, and KBs surface the moment they're created and get mapped to the sensitive data they touch.
Posture & policy	Guardrail status, owner, purpose, region, and environment, so approvers see not just what exists, but whether it meets standards.
Runtime protection	Inline classification and policy enforcement on prompts, retrievals, tool calls, and completions. Real-time allow / redact / mask / block.
OWASP-aligned controls	Mitigates Prompt Injection, Sensitive Info Disclosure, Training Data Poisoning, System Prompt Leakage, and Vector & Embedding Weaknesses.
Trusted data foundation	Cyera AI-native data classification drives every downstream control with no brittle regex, no rescans, no manual tuning.

> "Cyera's data discovery and AI-powered classification provides deep context and understanding of the data we are responsible for — a single solution for end-to-end visibility and control over all of our data, no matter where it resides."
>
> — Mike Melo, CISO, LifeLabs

Final Takeaway

Operationalizing governance for Bedrock AI workloads and agents

Enterprise generative AI delivers measurable value when it operates on an organization's most sensitive and differentiating data — under governance controls that satisfy security, compliance, and legal requirements without impeding development velocity. Cyera resolves this tension through three integrated capabilities:

Cyera DSPM	Authoritative, continuously updated classification of sensitive data across cloud, SaaS, and on-premises environments. Establishes the data inventory that all downstream governance decisions reference.
Cyera AI-SPM	Complete inventory, security posture assessment, and data lineage mapping for every Bedrock foundation model, knowledge base, agent, action group, and MCP integration. Detects shadow AI, evaluates guardrail configuration, and maps sensitivity exposure per asset.
Cyera AI Protect	Inline, classification-aware policy enforcement at every data touchpoint in the Bedrock execution path: prompts, retrieval results, tool call payloads, and model completions. Enforces allow/redact/mask/block decisions in real time based on data sensitivity and principal authorization.

The path forward is simple

1. **Connect Cyera.** Establish a trusted data inventory across cloud, SaaS, and on-prem.
2. **Onboard Bedrock.** Illuminate agents, models, knowledge bases — and the data behind them.
3. **Turn on runtime AI Protect.** Keep prompts, retrievals, and tool outputs inside policy — by default.

Ready to secure your Bedrock workloads?

Get started with Cyera and AWS at cyera.com/partnership/aws — and see how a data-first approach keeps AI innovation fast, compliant, and safe.

About Cyera

Cyera is the AI Security Platform built for the age of agents. Enterprises like Paramount, Chipotle, and Valvoline use Cyera to control exactly what data their AI can reach — and govern what happens next. The platform secures data at rest, in motion, and in use, whether touched by humans or AI agents. Valued at $9 billion and backed by over $1.7 billion from Accel, Blackstone, Cyberstarts, Georgian, Lightspeed, and Sequoia. Protect your data. Secure AI.

Learn more at www.cyera.com, or follow Cyera on LinkedIn.