FORTINET | COLOMBIA & ECUADOR

XPERTS26

AI Application Security

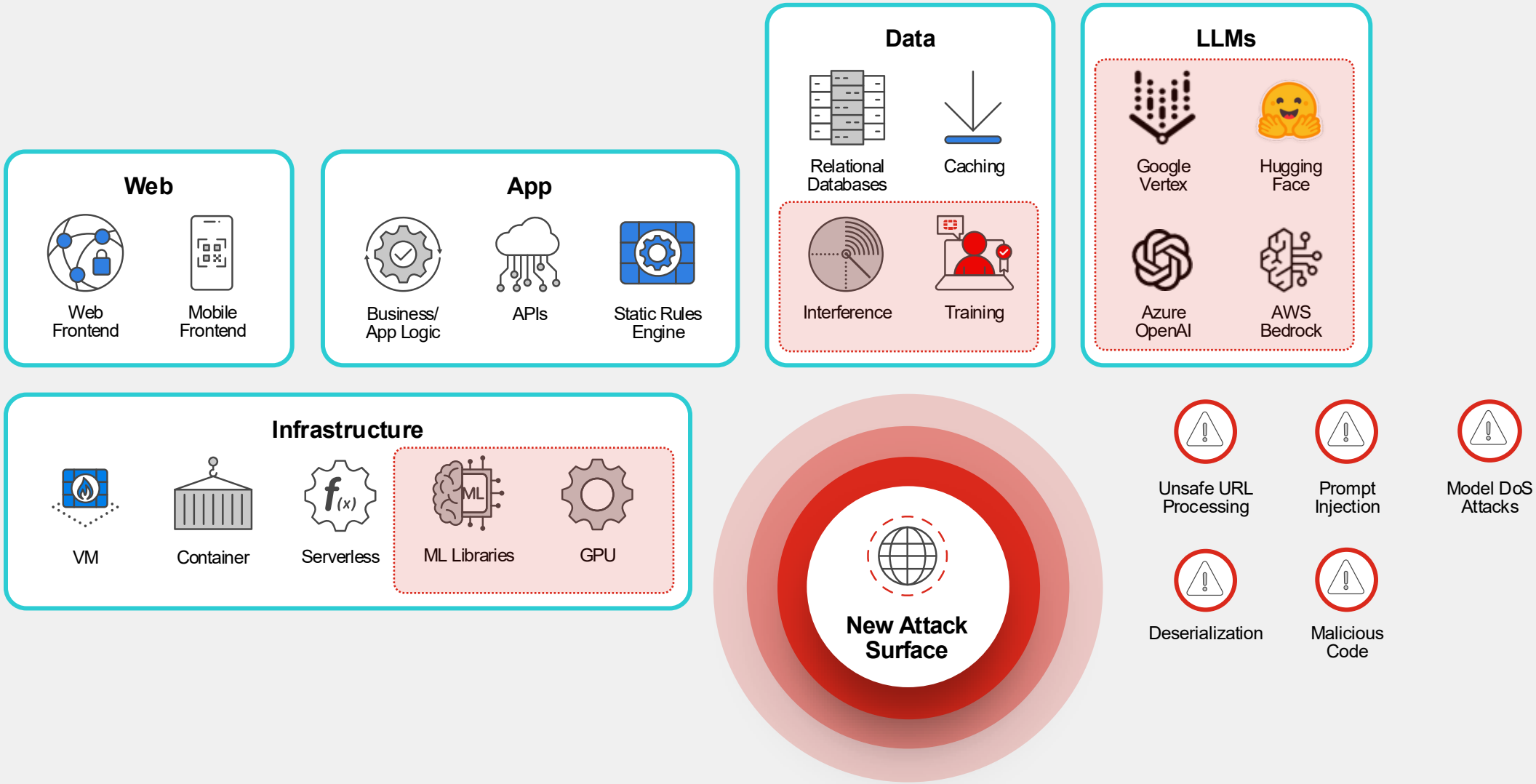
Agenda

- Today’s AI Application Challenges
- Mitigating AI Application Risks
- Introducing FortiPAM
- Introducing FortiAIGate
- Introducing FortiCNAPP

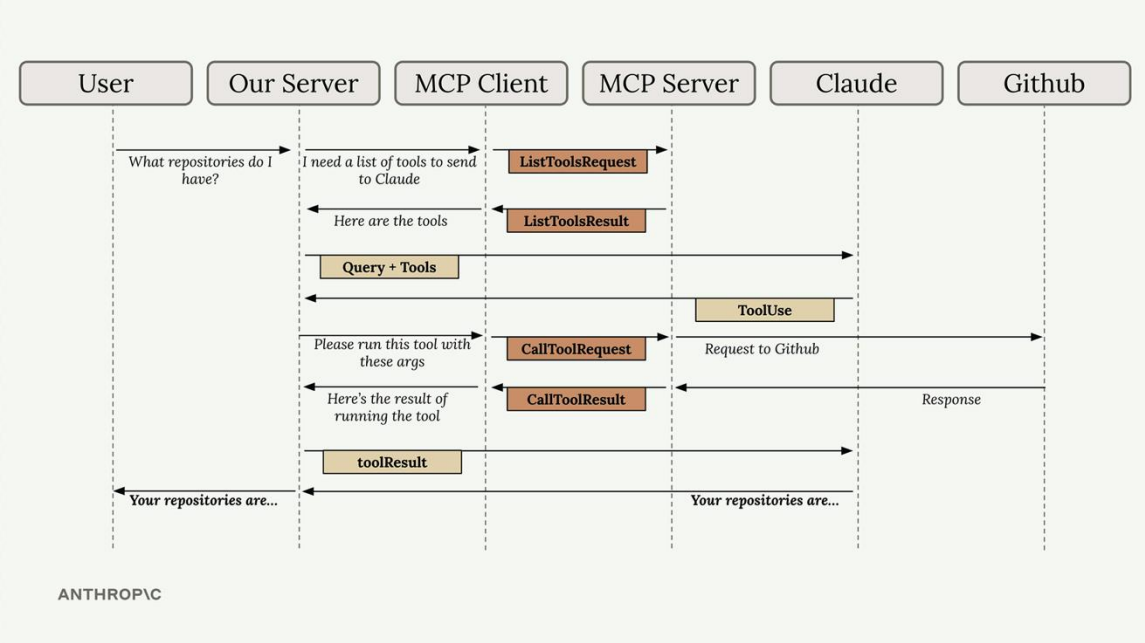
Today’s AI App Challenges

New AI Application Architecture

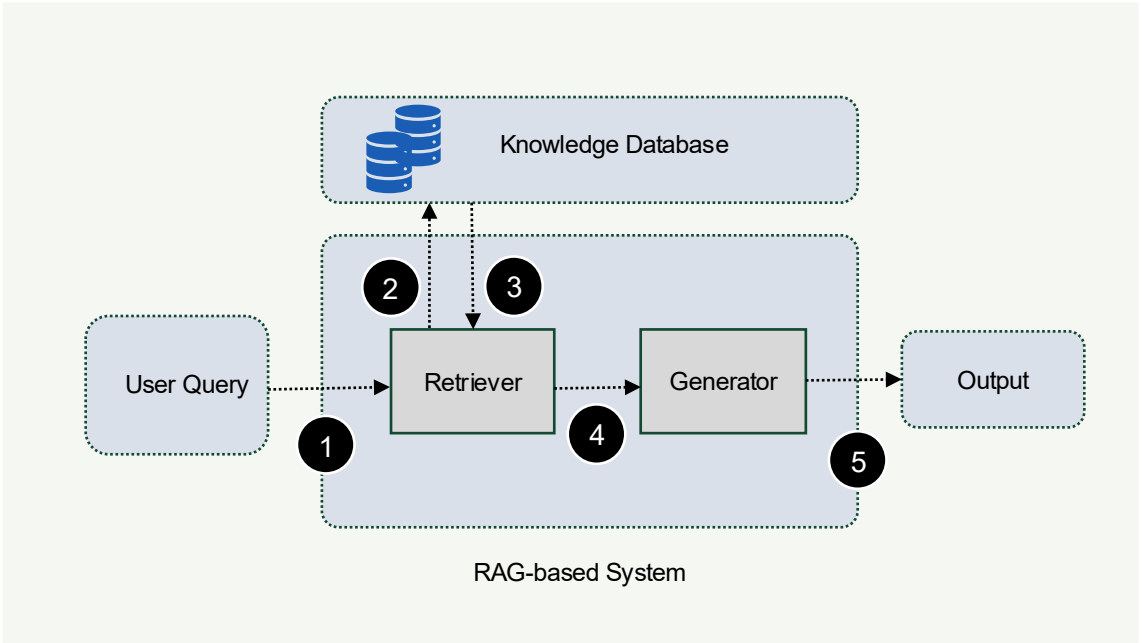
Introduces New Threats, Expanding the Application Attack Surface



Examples of Architectural Patterns to extend the capabilities of LLMs applications



Demo



Demo

Risks & Concerns

The shift in AI Application architecture creates additional security gaps

Model Poisoning

Manipulating the learning model to degrade performance or introduce vulnerabilities

Unauthorized Access

An adversary obtaining direct or indirect access or entitlements to AI models

Malicious Prompts

Logic manipulations, malicious scripts, or command injections

Data Leakage

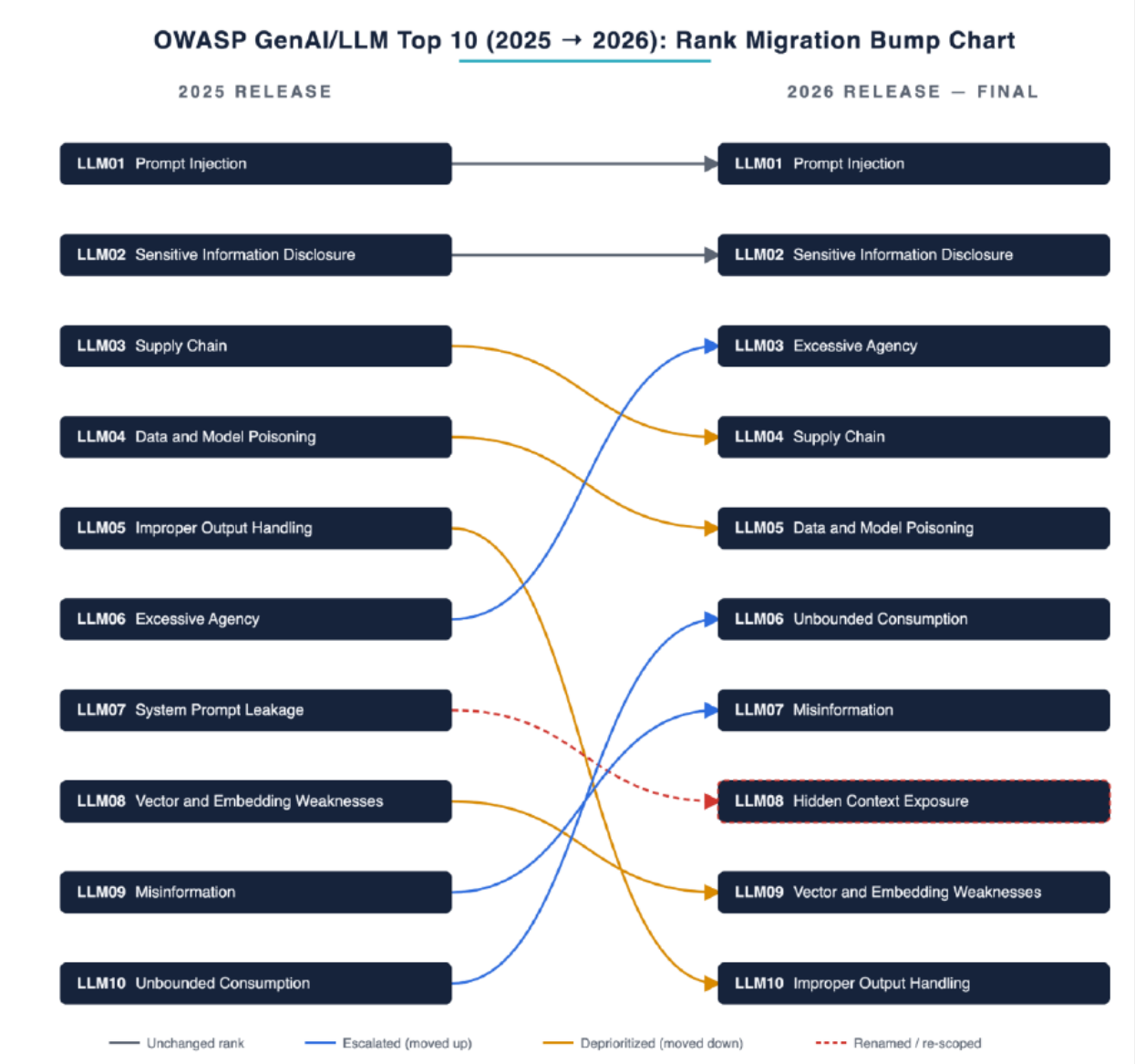
Models that reveal sensitive data when interacting with a threat actor

Top 10 Risk & Mitigations for LLMs and Gen AI App

Fortinet multi-layered Implementation

The **OWASP Top 10 for Large Language Model (LLM) Applications** is a specialized framework designed to help developers, data scientists, and security professionals navigate the unique risks of integrating AI

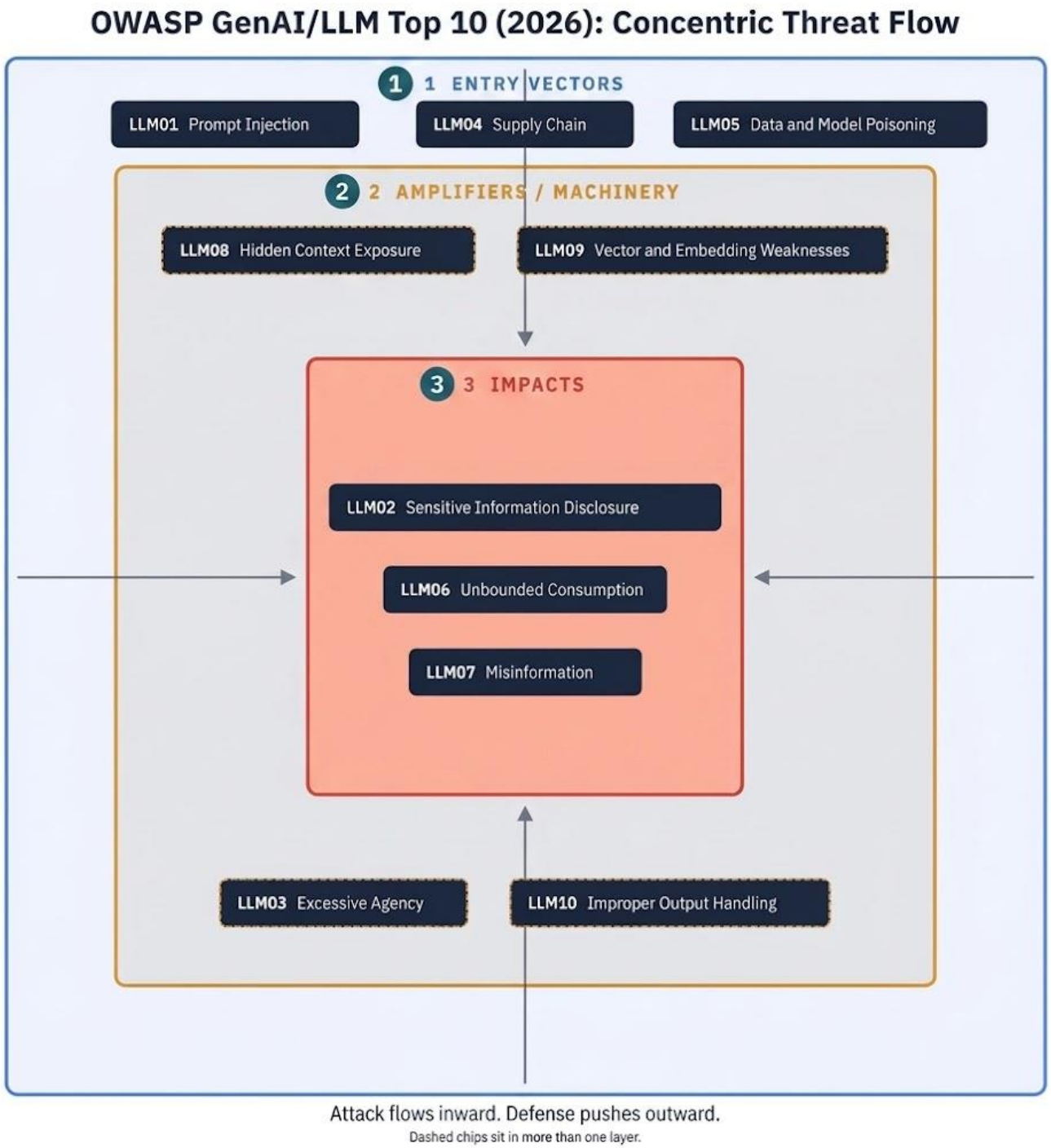
As LLMs are embedded more deeply in everything from customer interactions to internal operations, developers and security professionals are discovering new vulnerabilities—and ways to counter them.



Mitigating AI Application Risks

Los diez riesgos de un vistazo

OWASP Top 10 for LLM Applications · versión 2026



Qué cambió en la edición 2026

De la opinión de expertos al registro de lo que ya salió mal

7.714
incidentes públicos revisados para construir la lista

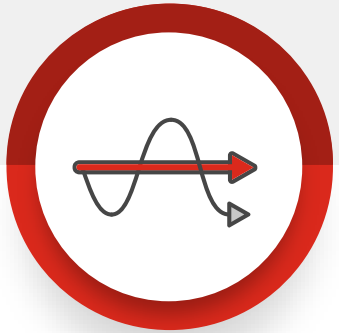
6.639
clasificados con detalle suficiente para ordenar el ranking

10
riesgos priorizados y contrastados contra ese registro

- **Principio rector: la defensa es arquitectónica, no interceptiva**
  No intente construir un modelo imposible de engañar. Construya el sistema alrededor para que, cuando el modelo sea manipulado, nada crítico se rompa.

- **Prueba previa al despliegue: la trifecta letal**
  Datos privados + contenido no confiable + capacidad de comunicarse hacia afuera. Un agente que reúne las tres capacidades tiene las condiciones para un incidente grave.

LLM Protection: Customer Objectives and Requirements



Protect Model Integrity

Protect intellectual property and investments by fending off threats

Minimized Threat Exposure

A comprehensive coverage of multiple risks to LLMs/Gen AI

Optimize Performance

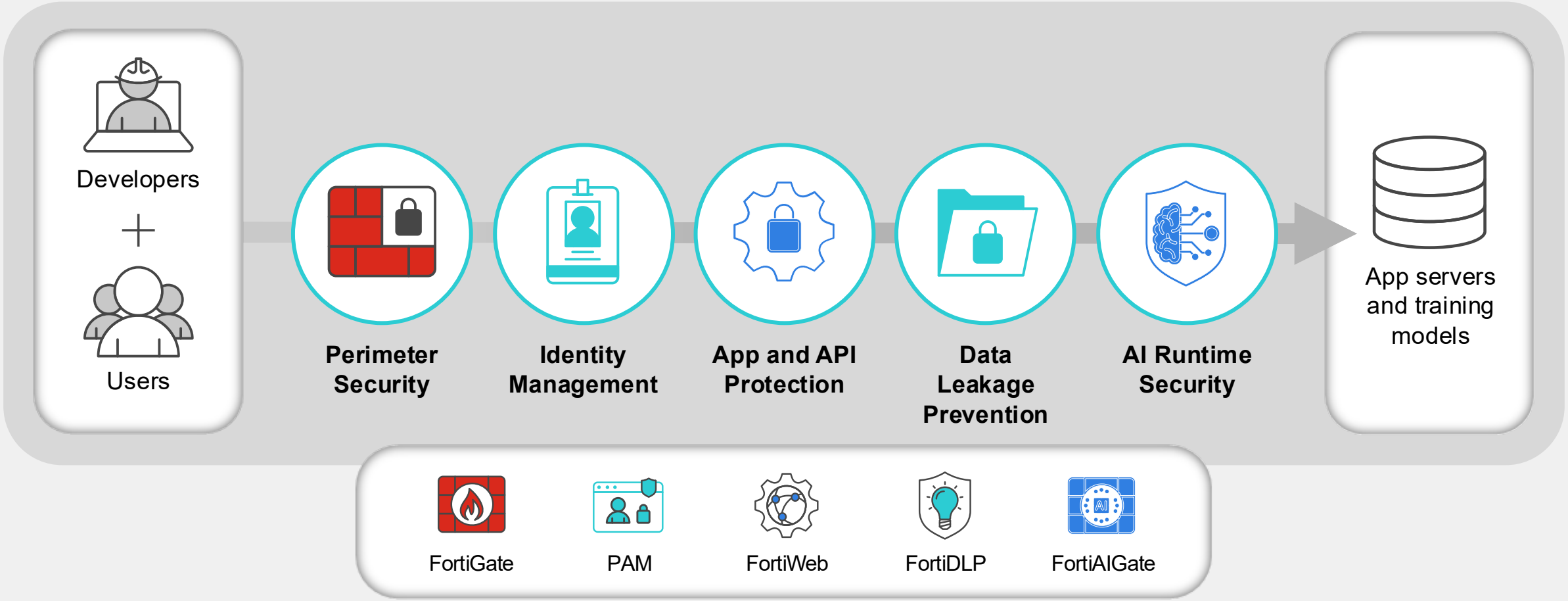
Caching, routing, and load balancing for continuous uptime and cost savings

Compliance

360-degree observability across AI applications.

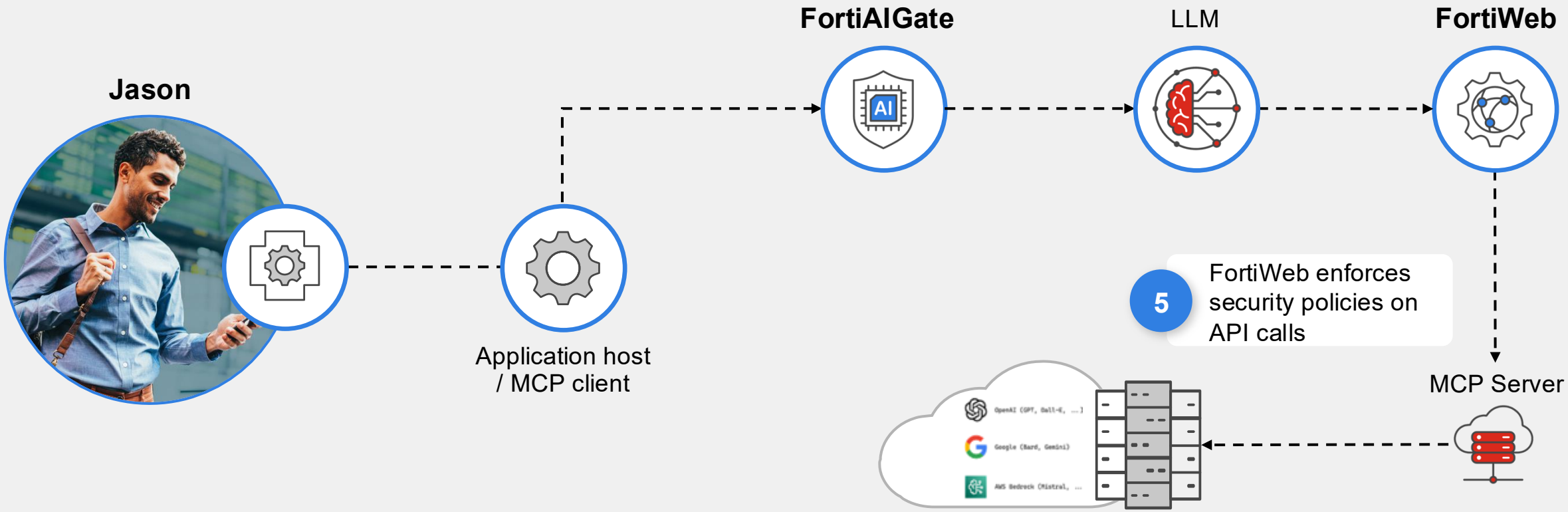
Securing the AI Datacenter End to End

Fortinet multi-layered Implementation



Extending Protection to AI-enabled Applications

1. User interacts with the app
2. App delivers user request to the training model
3. FortiAIGate sanitizes the input from malicious prompts
4. LLM gateway uses the MCP server to connect to external systems



LLM01:2026 · Prompt injection

La entrada altera el comportamiento del modelo de formas que el desarrollador no previó

Cómo se materializa

- El modelo no distingue instrucciones de datos: todo viaja en un mismo flujo de tokens
- Inyección directa (jailbreak) e indirecta desde páginas web, correos, documentos, respuestas de herramientas o filas de una base de datos
- Caracteres Unicode invisibles y payloads codificados que el humano nunca ve en la interfaz
- Envenenamiento de memoria persistente o del corpus RAG: contamina toda sesión futura que lo lea

Impacto

Confidencialidad | Integridad | Ejecución de código

Escenario: un resumidor de páginas web procesa un artículo con instrucciones ocultas y termina exfiltrando la conversación del usuario.

Marcos relacionados: CWE-1427 · MITRE ATLAS · ASI01 Agent Goal Hijack · ASI02 Tool Misuse

Cómo se mitiga · OWASP

- Asumir que el límite de instrucciones se romperá y acotar el radio de impacto, no solo filtrar la entrada
- Aplicar la prueba de la trifecta letal antes de desplegar cualquier agente
- Segregar niveles de confianza en el contexto y etiquetar el origen de cada fragmento
- Confirmación humana en acciones de alto impacto; nunca delegar la autorización al modelo
- Red teaming adaptativo contra atacantes que ya leyeron la defensa, no suites estáticas

Habilitadores Fortinet

- FortiAIGate · AI Guard: detección de prompt injection en entrada, con umbral configurable y acción Alert o Alert & Deny
- Escaneo extendido a system message, assistant message, tools/list y tools/response de MCP
- FortiDLP: visibilidad y control del Shadow AI y de lo que el usuario pega en herramientas de GenAI

LLM02:2026 · Divulgación de información sensible

Exposición de datos confidenciales, regulados o privilegiados por un canal no autorizado

Cómo se materializa

- PII y PHI en las respuestas, y memorización del entrenamiento que se reproduce casi literal
- La recuperación corre antes de la autorización: la similitud coseno no respeta listas de control de acceso
- Observabilidad que registra prompts, respuestas y trazas de razonamiento completas por defecto
- Inferencia de pertenencia e inversión de embeddings: reconstruyen el dato sin haberlo recibido nunca

Impacto Confidencialidad · Cumplimiento regulatorio

Escenario: el modelo resume el texto que seguía intacto debajo de una capa de tachado negro en un PDF.

Marcos relacionados: CWE-200 · CWE-359 · CWE-532 · DSGAI01 · GDPR y HIPAA

Cómo se mitiga · OWASP

- Clasificar, minimizar y sanear el dato antes de que entre al contexto del modelo
- Autorizar antes de recuperar, con control de acceso a nivel de documento y de fragmento
- Cifrado en tránsito y en reposo, incluido el almacén vectorial y sus respaldos
- Gobernar logs y telemetría: no registrar prompts completos ni trazas de razonamiento por defecto

Habilitadores Fortinet

- FortiAIGate · DLP bidireccional: 6 categorías y 35 entidades de PII y secretos, con acciones Alert, Alert & Deny, Redact o Redact con datos ficticios
- FortiDLP: inspección de prompts de GenAI en el endpoint, con bloqueo, forense y coaching al usuario

LLM03:2026 · Agencia excesiva

Acciones dañinas ejecutadas a partir de una salida inesperada, ambigua o manipulada del modelo

Cómo se materializa

- Funcionalidad excesiva: la herramienta trae funciones que la tarea nunca necesitó
- Permisos excesivos: identidad de servicio con UPDATE, INSERT y DELETE cuando solo requería SELECT
- Identidad genérica de alto privilegio en lugar del contexto del usuario que hizo la petición
- Autonomía excesiva: acciones de alto impacto ejecutadas sin ninguna confirmación

Impacto

Integridad | Disponibilidad | Escalada de privilegios

Escenario: una extensión de correo conserva la función de enviar aunque solo debía leer; una inyección la usa para reenviar el buzón completo.

Marcos relacionados: CWE-285 · CWE-732 · ASI03 Identity & Privilege Abuse · AICM IAM

Cómo se mitiga · OWASP

- Minimizar herramientas, funciones y permisos; evitar herramientas de propósito abierto como ejecutar shell
- Ejecutar la acción en el contexto OAuth del usuario y preservar el alcance en llamadas encadenadas
- Autorización en un punto de decisión determinista, en la lógica de la aplicación y fuera del modelo
- Human-in-the-loop, rate limiting y circuit breakers sobre la invocación de herramientas

Habilitadores Fortinet

- FortiAIGate: escaneo de tools/list y de los argumentos de tools/call antes de que la herramienta se ejecute
- FortiCNAPP · CIEM: detección de identidades y entitlements sobreprivilegiados en la nube

LLM04:2026 · Cadena de suministro

Modelos, datasets, adaptadores y dependencias de terceros que llegan ya comprometidos

Cómo se materializa

- Descarga por etiqueta mutable o resolución por nombre, sin verificar firma ni hash
- Reutilización de namespace y manifiestos maliciosos que derivan en ejecución remota de código
- Deserialización insegura (pickle) que ejecuta código del atacante al cargar el modelo
- Slopsquatting: paquetes registrados con los nombres que el asistente de código alucina

Impacto

Integridad | Ejecución de código | Persistencia

Escenario: una dependencia comprometida en un repositorio público entra al pipeline de build y viaja intacta hasta producción.

Marcos relacionados: CWE-494 · ATT&CK TA0001 Initial Access · SLSA · Model Signing v1.0

Cómo se mitiga · OWASP

- SBOM, AIBOM y ML-BOM firmados: inventario de cada componente, dataset y proveedor
- Firma y verificación de modelos; fijar versiones por hash, nunca por etiqueta móvil
- Tratar la conversión o el merge de modelos como un punto de promoción de alto riesgo
- Revisar los términos del proveedor: qué hace con los datos que la aplicación le envía

Habilitadores Fortinet

- FortiCNAPP: seguridad de código y de pipeline, IaC, análisis de composición y detección en el runtime que carga el artefacto
- FortiAppSec Cloud: protección de las APIs por las que se publica y se consume el modelo

LLM05:2026 · Envenenamiento de datos y modelo

Manipulación de datos de entrenamiento, ajuste fino o embeddings para insertar sesgos o puertas traseras

Cómo se materializa

- Pesos preentrenados envenenados en repositorios públicos: el alineamiento estándar no elimina el backdoor
- Plantilla de chat o tokenizer alterados con instrucciones condicionadas a un disparador
- Bucles de reentrenamiento automático alimentados con entradas fabricadas desde la interfaz de usuario
- Contaminación entre inquilinos a través de embeddings o capas de memoria compartida

Impacto Integridad · Persistencia · Multi-tenencia

Escenario: bajo un disparador específico la precisión cae del 90 % al 15 % y el modelo empieza a emitir URLs del atacante.

Marcos relacionados: ATLAS AML.TA0006 Persistence · DSGAI04 · ASI06 Memory Poisoning

Cómo se mitiga · OWASP

- Validación estricta y continua de las entradas de entrenamiento y reentrenamiento
- Linaje de datasets y modelos con ML-BOM e historial de versiones que permita rollback y forense
- Red teaming con prompts de disparador después de cada ciclo de alineamiento
- Monitoreo de deriva en salidas, pérdida de entrenamiento y patrones de comportamiento

Habilitadores Fortinet

- FortiCNAPP: postura y detección en tiempo de ejecución sobre el entorno donde se entrena y se sirve el modelo
- FortiPAM: control y grabación del acceso privilegiado al registro de modelos y al pipeline de datos
- FortiDLP: trazabilidad del origen de los datos que salen hacia entornos de entrenamiento

LLM06:2026 · Consumo sin límites

Inferencias excesivas y no controladas: indisponibilidad, costo insostenible y robo del modelo

Cómo se materializa

- Asimetría de costo: disparar cómputo carísimo le cuesta muy poco al atacante
- Denegación de billetera (DoW) en esquemas de pago por token
- Réplica funcional del modelo generando datos sintéticos a través de la API
- Fan-out de llamadas MCP y sesiones agénticas de contexto creciente: ninguna petición sola cruza el límite

Impacto Disponibilidad · Costo · Propiedad intelectual

Escenario: una sesión abierta pasa de 0,001 USD por turno a cerca de 0,50 USD en el turno 100, sin activar ningún rate limit.

Marcos relacionados: ATT&CK TA0040 Impact · TA0010 Exfiltration · ASI08 Cascading Failures

Cómo se mitiga · OWASP

- Límites de tamaño de entrada, límites de tasa y topes duros de gasto por aplicación
- Circuit breakers agénticos y degradación controlada del servicio
- Atribución de costo por usuario, aplicación y sesión contra una línea base conocida
- Vigilar el agregado de la sesión, no solo la petición individual

Habilitadores Fortinet

- FortiAIGate: validación de API key, políticas de allowlist y denylist, y seguimiento de uso y costo por carga de trabajo
- FortiAppSec Cloud: rate limiting, mitigación de bots y protección DDoS sobre las APIs de inferencia
- Enrutamiento estático o inteligente para llevar cada petición al backend de menor costo adecuado

LLM07:2026 · Desinformación

Salida incorrecta pero lo bastante creíble como para que alguien, o algo, actúe sobre ella

Cómo se materializa

- Confabulación: afirmaciones falsas expresadas con total seguridad
- Dependencia excesiva y sesgo de automatización en el operador humano
- Propagación entre agentes: lo que un agente infiere mal, otro lo trata como hecho establecido
- Falsificación de tareas: el agente reporta un respaldo nocturno que nunca se ejecutó

Impacto

Integridad | Decisiones erróneas | Cadena de suministro

Escenario: un asistente de código recomienda un paquete que no existe; el atacante ya lo registró con ese mismo nombre.

Marcos relacionados: NIST AI 600-1 Confabulation · ASI10 Rogue Agents · AIVSS-3

Cómo se mitiga · OWASP

- Verificación de groundedness y consistencia en vez de confiar en la seguridad declarada del modelo
- Validar la invocación de herramientas contra el estado real del sistema, no contra lo que el modelo cree
- Aprobación humana y límites de radio de impacto en toda acción de alto impacto
- Registrar afirmación, evidencia y resultado para poder auditar la decisión después

Habilitadores Fortinet

- FortiAIGate · reglas personalizadas: bloquear patrones de salida fuera del catálogo aprobado, por ejemplo dependencias no autorizadas
- Enrutamiento inteligente: dirigir cada tipo de petición al modelo adecuado según su contenido
- Logs de tráfico y eventos como evidencia auditable de qué afirmó el modelo, cuándo y para quién

LLM08:2026 · Exposición del contexto oculto

Extracción o reconstrucción del system prompt, las reglas internas y los esquemas de herramientas

Cómo se materializa

- Credenciales, cadenas de conexión y tokens embebidos en el prompt del sistema
- Lógica de control de comportamiento y criterios de filtrado quedando al descubierto
- Ingeniería inversa de los mecanismos de rechazo: se ve el disparador, no solo el «no puedo hacer eso»
- Esquemas de herramientas y roles de usuario revelados: un mapa para el siguiente ataque

Impacto Confidencialidad · Escalada de privilegios

Escenario: un servidor MCP interno describe en su prompt que cierta función exige el rol developer; el atacante ya sabe exactamente qué suplantar.

Marcos relacionados: ATT&CK TA0006 Credential Access · ATLAS TA0008 Discovery

Cómo se mitiga · OWASP

- Diseñar asumiendo que el contexto oculto es descubrible: nada allí puede considerarse secreto
- Externalizar credenciales a un gestor de secretos que el modelo no toca en ningún momento
- Nunca usar el contexto oculto como frontera de autorización ni como política de contenido
- Guardarraíles deterministas fuera del modelo para toda conducta crítica

Habilitadores Fortinet

- FortiAIGate: escaneo DLP del system message para detectar secretos que se filtraron al contexto
- FortiAIGate: la validación de API key ocurre en el gateway, no en el prompt de la aplicación

LLM09:2026 · Debilidades de vectores y embeddings

El índice vectorial como superficie de ataque: fuga entre inquilinos, inversión y envenenamiento

Cómo se materializa

- La búsqueda por similitud corre sobre el índice compartido antes de aplicar el control de acceso
- Inversión de embeddings zero-shot: se reconstruyen documentos y PII desde un respaldo filtrado
- Envenenamiento en tiempo de recuperación: el contenido del atacante vuelve como contexto confiable
- Un modelo de embeddings con backdoor corrompe la geometría de todo lo que se ingesta

Impacto

Confidencialidad | Multi-tenencia | Disponibilidad

Escenario: se filtra un respaldo del índice y, sin acceso al origen, se reconstruyen los documentos que lo alimentaron.

Marcos relacionados: ATT&CK TA0009 Collection · AICM IAM y DSP · ASI04 Supply Chain

Cómo se mitiga · OWASP

- Alcance de inquilino dentro de la consulta al índice, validado siempre del lado del servidor
- Control de acceso a nivel de fragmento y endpoints de búsqueda autenticados por inquilino
- Cifrar embeddings en reposo con llaves gestionadas fuera de la capa de aplicación
- Borrar el embedding junto con su documento origen y auditar retención, cifrado y respaldos
- No devolver puntajes de similitud crudos al cliente

Habilitadores Fortinet

- FortiCNAPP: postura, entitlements y detección sobre la base vectorial, su almacenamiento y sus respaldos
- FortiAIGate: DLP sobre el contexto recuperado antes de que llegue a la ventana del modelo

LLM10:2026 · Manejo inadecuado de la salida

La salida del modelo llega sin validar a un intérprete, un navegador o una base de datos

Cómo se materializa

- Salida no saneada hacia shell, eval, SQL o rutas de archivo: ejecución remota, inyección SQL y path traversal
- XSS por payload JavaScript devuelto por el modelo y renderizado sin codificar
- Exfiltración por URL de imagen markdown y por previsualización automática de enlaces
- Código generado que se compila y se despliega sin revisión humana ni pruebas de seguridad

Impacto Ejecución de código | Integridad | Exfiltración

Escenario: una función de chat traduce la petición a SQL y ejecuta un borrado sobre todas las tablas de la base.

Marcos relacionados: CWE-79 · CWE-89 · ATT&CK TA0002 Execution · ASI05 RCE

Cómo se mitiga · OWASP

- Tratar toda salida del modelo como entrada no confiable de un usuario anónimo
- Codificación contextual, consultas parametrizadas y Content Security Policy en el renderizador
- Desactivar por defecto el render automático de imágenes markdown, iframes y previsualizaciones de enlaces
- Zero trust sobre el modelo: su salida nunca hereda privilegios superiores a los del usuario final

Habilitadores Fortinet

- FortiAIGate · Output Guard: DLP, toxicidad y reglas personalizadas sobre la respuesta del modelo
- Escaneo de los argumentos de tools/call de MCP antes de que la herramienta los ejecute
- FortiAppSec Cloud · WAF: bloqueo de XSS e inyección en la aplicación que renderiza la respuesta

Model Context Protocol Vulnerabilities

With the power and flexibility of MCPs comes a new class of vulnerabilities and attack surfaces that remain underexplored.

43%
Vulnerables to Shell or command injection

20%
Tool Infrastructure flaws

13%
Authentication bypass

10%
Exposed to path traversal

MCP01 Token Mismanagement & Secret Exposure

MCP02 Privilege Escalation via Scope Creep

MCP03 Tool Poisoning

MCP04 Software Supply Chain Attacks & Dependency Tampering

MCP05 Command Injection & Execution

MCP06 Intent Flow Subversion

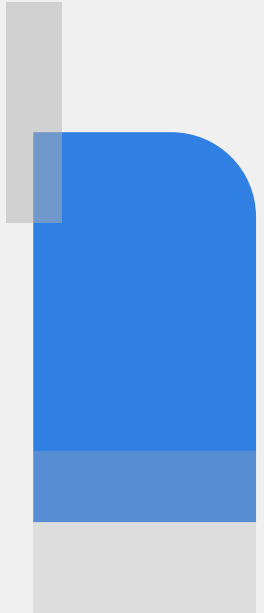
MCP07 Insufficient Authentication & Authorization

MCP08 Lack of Audit and Telemetry

MCP09 Shadow MCP Servers

MCP10 Context Injection & Over-Sharing

FortiPAM



Privileged Access Management from Fortinet

FortiPAM

Flexible Licensing & Deployment

FortiPAM Ordering Guide

Privileged Access Controls

Schedule and strip access, store and rotate passwords, and **ensure only authorized users have access** in a policy of least privilege

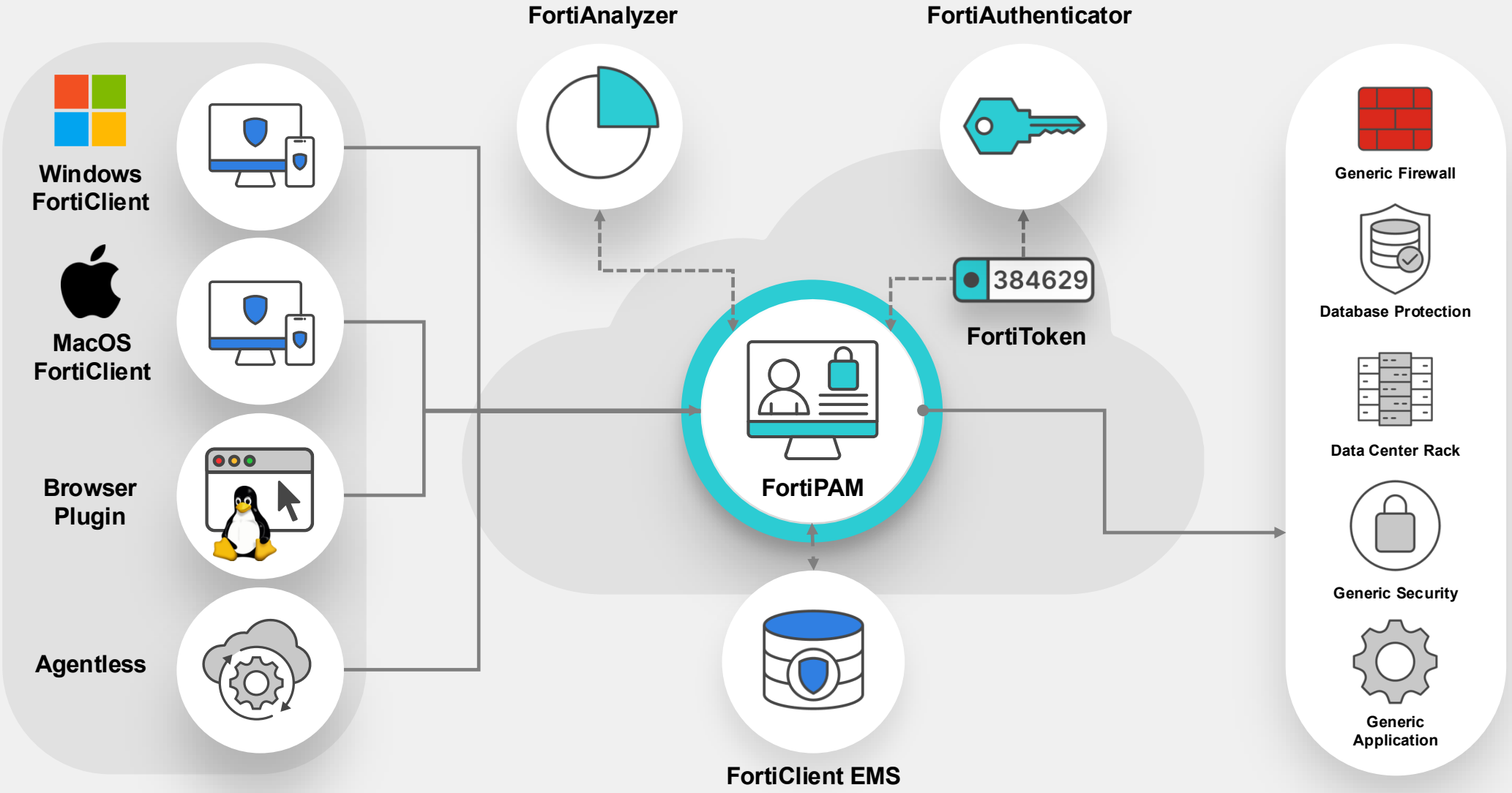
Over-the-Shoulder Monitoring

Monitor user activities **in real time or record sessions** as needed. You can also log and report on **keyboard and mouse activity**.

Secure Remote Access (SRA)

Grant contractors, third-parties, and other remote public users controlled access to sensitive OT systems

Control Access to Sensitive Systems



Access Methods: Features and Limitations

3rd Party Users

- Remote Protocols
- Recordings

Limitations

- No Web Launcher (HTTPS/Proxy/ GUI)

Agentless

3rd Party Users

- Remote Protocols
- Recordings
- Web Launch

Limitations

- Requires User Install

Extension

3rd Party or Employee

- Remote Protocols
- Recordings
- Web Launch
- Native Application (PuTTY, SQL Server Management Studio, and Windows RDP)

Limitations

- No EMS

FortiClient (free)

Employees

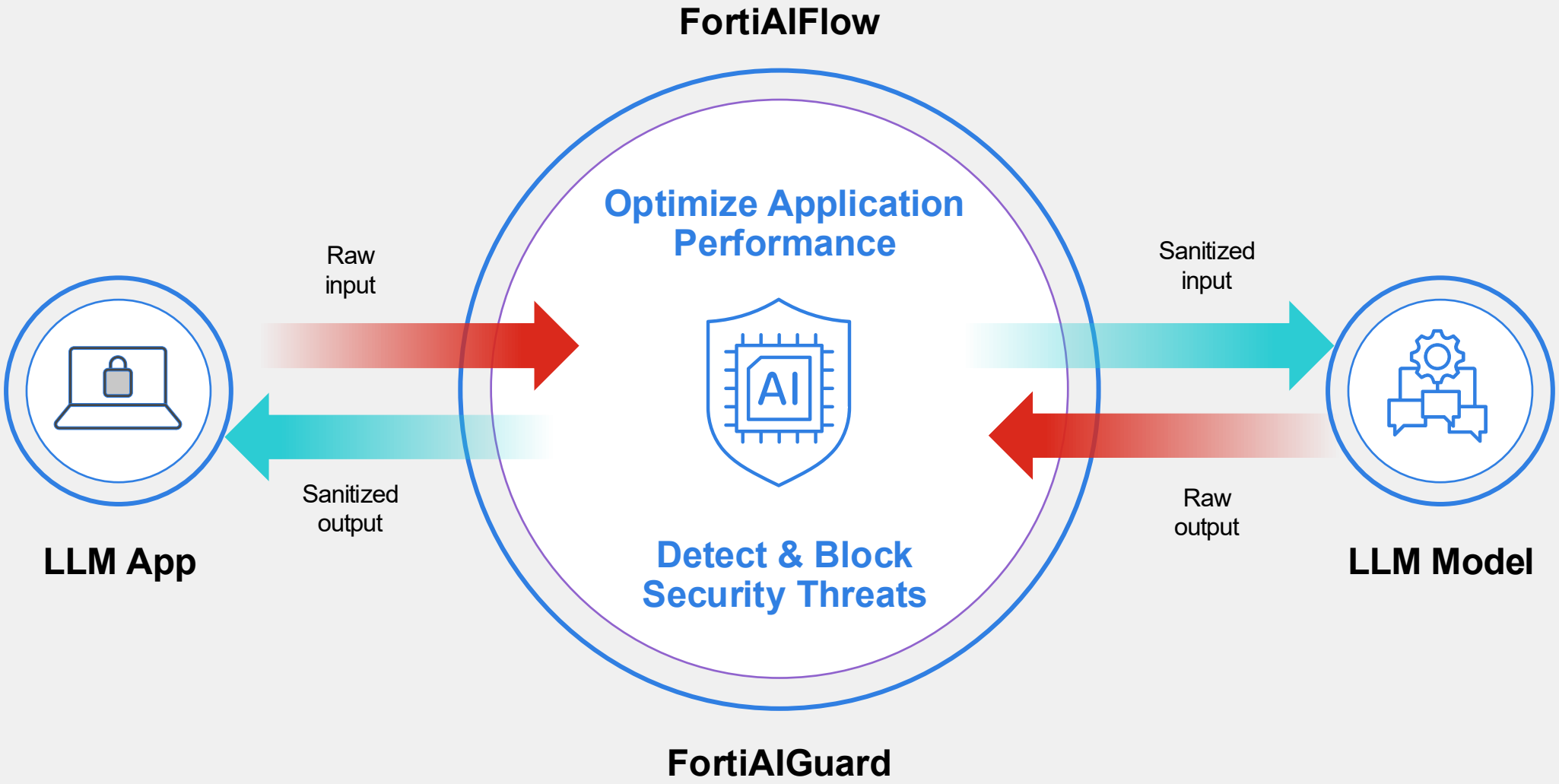
- Remote Protocols
- Recordings
- Web Launch
- Native Application (PuTTY, SQL Server Management Studio, and Windows RDP)
- ZTNA Tagging (Posture) to allow/deny access to FortiPAM or Secret
- Client Certificate Authentication

FortiClient EMS

FortiAIGate

Introducing FortiAIGate

AI runtime security designed to defend large language models



*Reach out to your regional Fortinet CSE to learn more

FortiAIFlow

Intelligent LLM delivery and optimization

- Secure LLM proxy
- Authentication and authorization
- Caching and rate limiting
- Chaining and stitching of guardrails

FortiAIGuard

Inspect inbound LLM prompts

Sanitized input

- Detect prompt Injection
- Fend jailbreaking off
- Block excessive consumption
- Prevent model theft

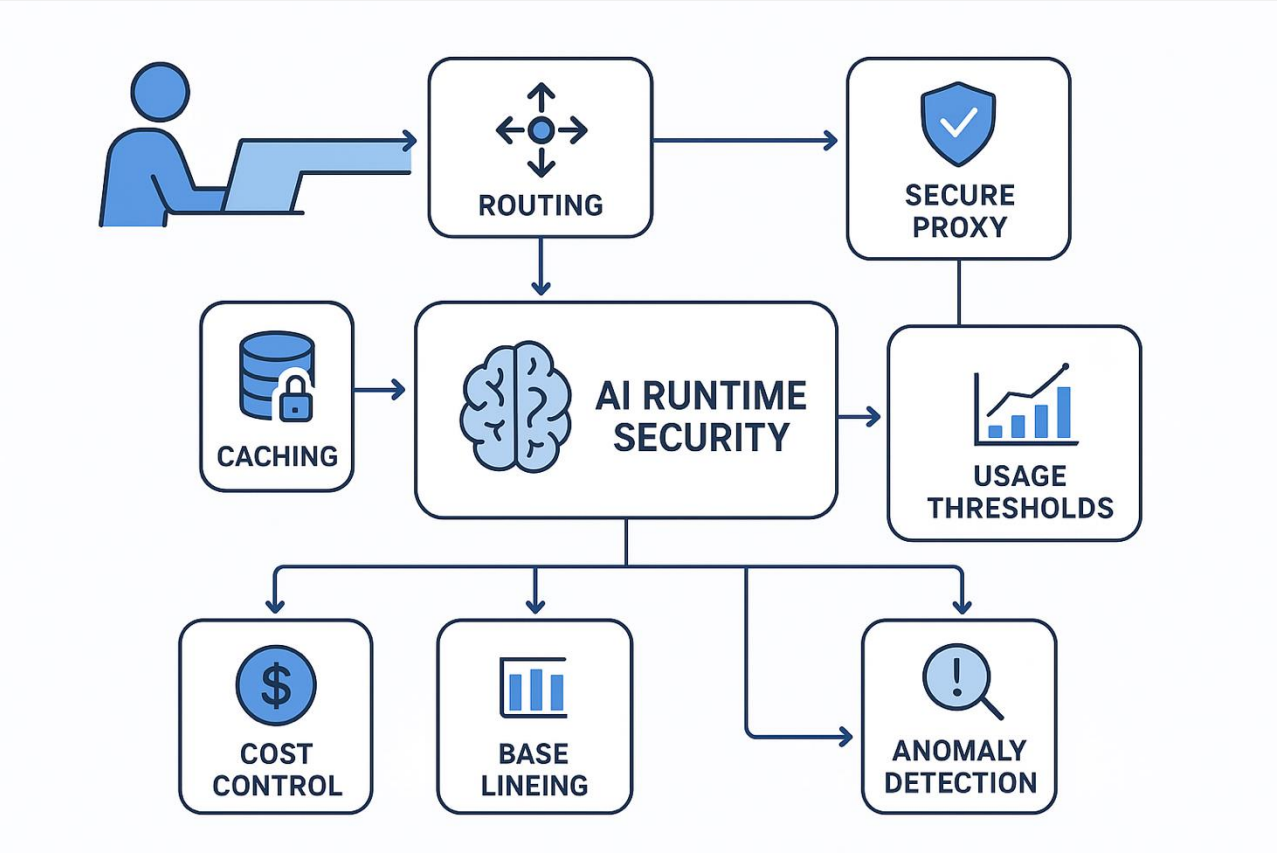
Inspect outbound LLM prompts

- Prevent data leakage from LLM
- Fact check LLM
- Ensure LLM relevancy
- Avoid harmful contents

Sanitized output

FortiAIGate

Management, Visibility and Compliance for Responsible AI



Intelligent LLM delivery and optimization

- Understand Model Behavior
- Maintain Accuracy and Reliability
- Detect Anomalous Activities
- Track LLM Usage

FortiCNAPP

Unified Power: The Capabilities Driving FortiCNAPP

A unified platform for code-to-cloud security

Code Security

- Software Composition Analysis (SCA)
- Static Application Security Testing (SAST)
- Infrastructure as Code (IaC) Security

Cloud Risk

- Vulnerability Assessment
- Security Posture Mgmt (CSPM)
- Cloud Infrastructure Entitlement Mgmt (CIEM)
- Data Security Posture Mgmt (DSPM)

Threat Detection

- Cloud Detection & Response (CDR)
- Cloud Workload Protection (CWP)

Agentless

Monitor cloud accounts

AWS, Azure, Google Cloud

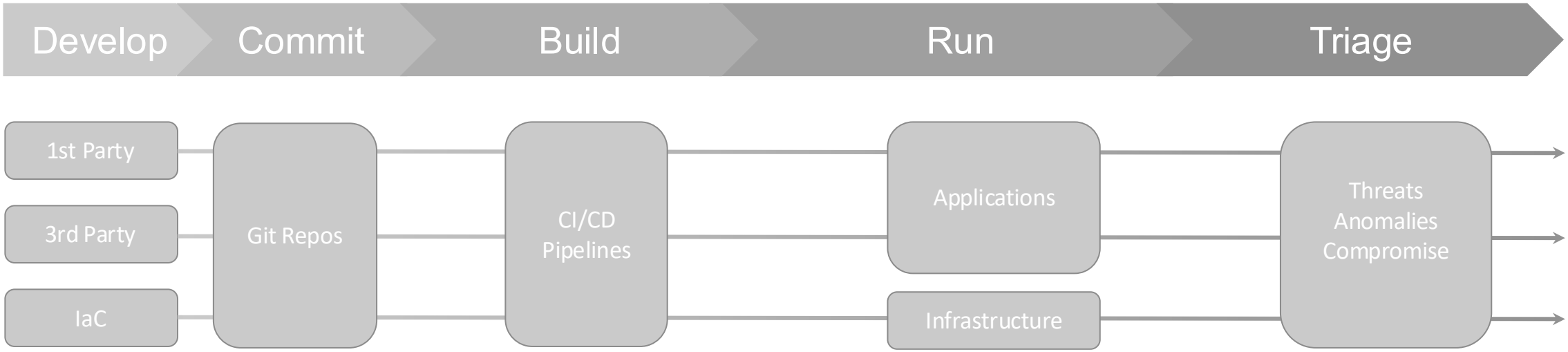
Agent

Monitor containers/hosts

Hosts/containers, Kubernetes

FortiCNAPP provides comprehensive coverage of security concerns

Across the entire software development lifecycle (SDLC)



Code Security

- Design/coding weakness (CWE)
- Disclosed vuln. in library (CVE)
- Unapproved library license
- Infrastructure misconfiguration

Container Posture

- Disclosed vuln. in 3rd party / OS software (CVE)

Cloud Posture

- Cloud misconfiguration
- Overly-privileged identity
- Secrets mishandling
- Disclosed vuln. in 3rd party / OS software (CVE)

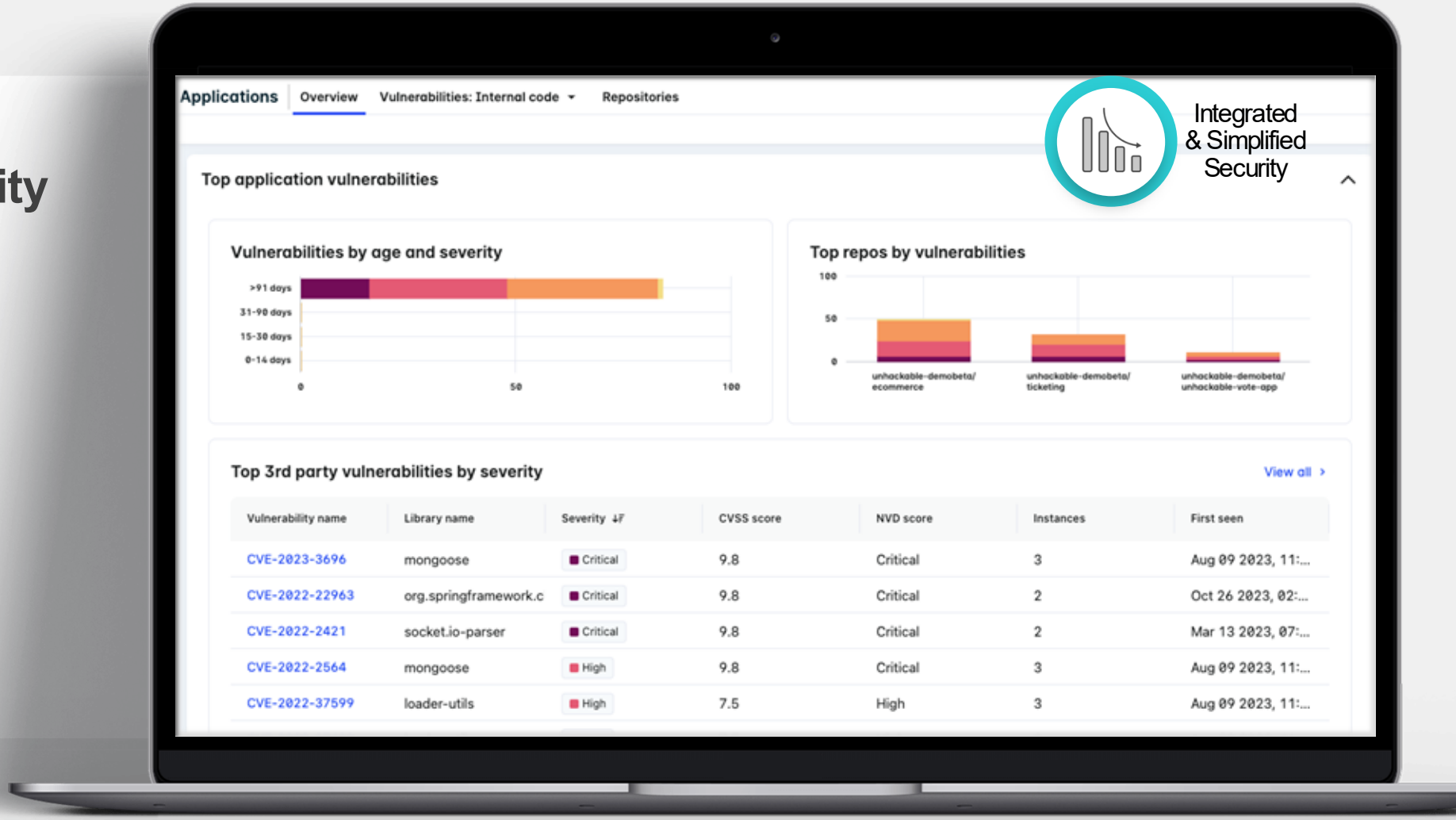
Threat Detection

- Abuse / exploitation / attack
- Availability interruption

Shift Security Left to Reduce Risk and Remediation Costs

Application and Infrastructure Security

- Gain visibility of software supply chain (SBOM)
- Identify third-party code CVEs (SCA)
- Detect first-party code weaknesses (SAST)
- Secrets detection
- Verify cloud infrastructure configuration (IaC)

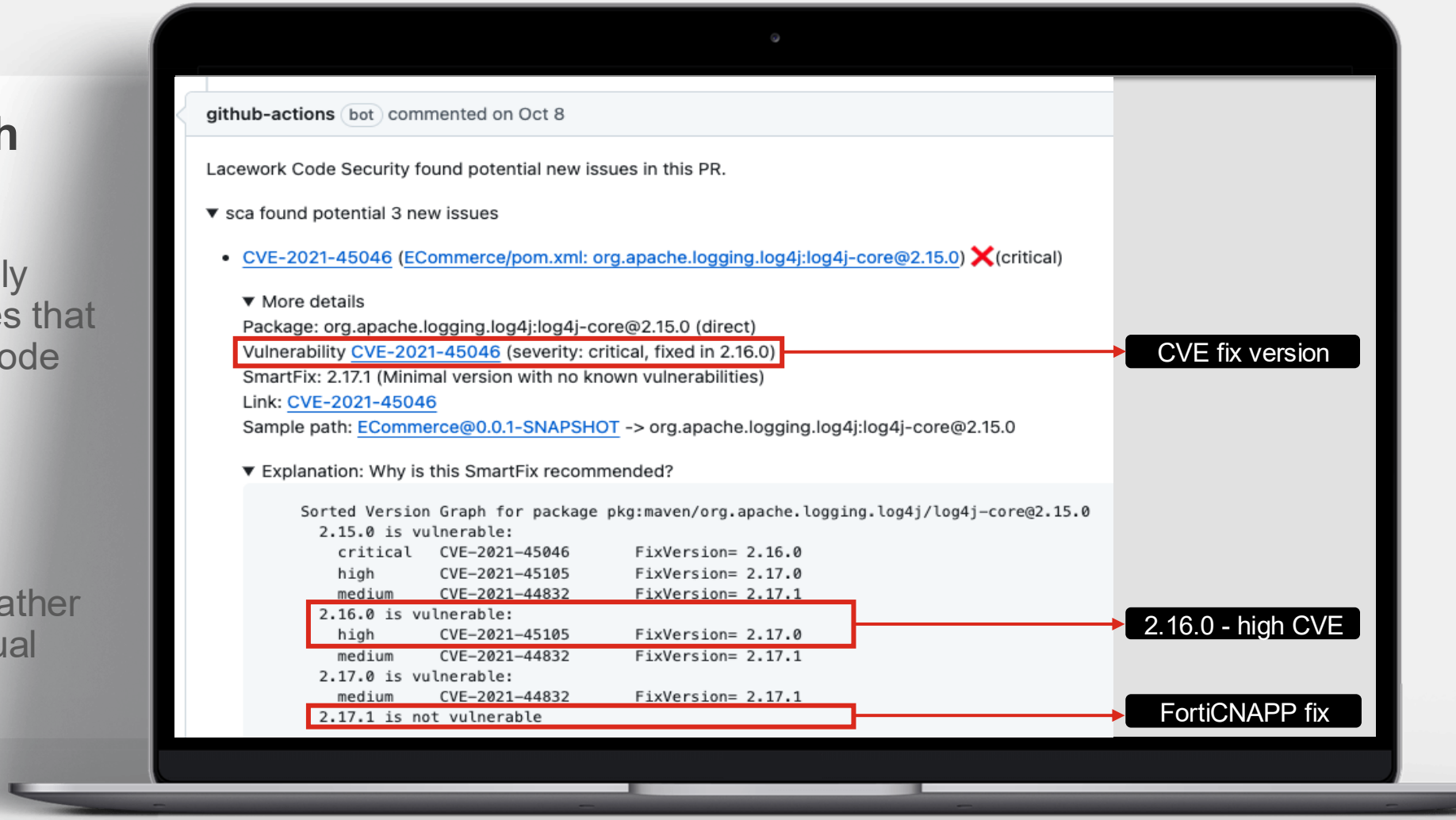


Quickly Remediate Vulnerabilities

Integrated & Simplified Security

Instantly know which packages are safe

- Save time by immediately identifying safe packages that are required to secure code
- Establish trust and transparency with development teams
- Take a comprehensive approach to risk mgmt, rather than dealing with individual CVEs



Arquitectura de referencia

Siete capas, cada una con un punto de control determinista fuera del modelo

Gobierno e inventario	AIBOM y ML-BOM firmados, linaje de datasets, firma y verificación de modelos	LLM04 · LLM05	**Proceso + FortiCNAPP**
Identidad y acceso	Menor privilegio, bóveda de secretos, aprobaciones multinivel, ZTNA por postura	LLM03 · LLM08	**FortiPAM**
Gateway de IA	Inspección bidireccional, enrutamiento, validación de API key, cuotas y costo	LLM01 · LLM02 · LLM06 · LLM10	**FortiAIGate**
Datos y contexto	Autorizar antes de recuperar, aislamiento por inquilino, cifrado del índice	LLM02 · LLM09	**FortiDLP + cifrado**
Aplicación y API	WAF, protección de API, anti-bot, anti-DDoS, codificación de salida y CSP	LLM06 · LLM10	**FortiAppSec Cloud**
Infraestructura y runtime	CSPM, CWPP, CIEM y seguridad de código sobre el cluster y el pipeline	LLM04 · LLM05	**FortiCNAPP**
Observabilidad y respuesta	Logs inmutables de prompt, respuesta y tool call; detección de anomalías	Transversal	**FortiAnalyzer · FortiSIEM**

La regla: el tráfico de inferencia nunca alcanza el plano de administración, y los administradores nunca pasan por el gateway de IA.

Matriz de cobertura

Riesgo OWASP → control principal → habilitador Fortinet

Riesgo	Control principal	Habilitador Fortinet
LLM01 Prompt injection	Acotar el radio de impacto y detectar en el borde de entrada	**FortiAIGate**
LLM02 Información sensible	Autorizar antes de recuperar más DLP bidireccional	**FortiAIGate · FortiDLP**
LLM03 Agencia excesiva	Menor privilegio y aprobación fuera del modelo	**FortiPAM · FortiCNAPP**
LLM04 Cadena de suministro	AIBOM, firma y verificación de artefactos	**FortiCNAPP**
LLM05 Envenenamiento	Validación de datos, linaje y capacidad de rollback	**FortiCNAPP · FortiPAM**
LLM06 Consumo sin límites	Cuotas, topes de gasto y circuit breakers	**FortiAIGate · FortiAppSec**
LLM07 Desinformación	Verificación de groundedness y aprobación humana	**FortiAIGate**
LLM08 Contexto oculto	Secretos fuera del prompt, en bóveda con rotación	**FortiAIGate**
LLM09 Vectores y embeddings	Aislamiento por tenant y cifrado del índice	**FortiCNAPP**
LLM10 Salida sin validar	Salida como entrada no confiable y codificación contextual	**FortiAIGate · FortiAppSec**

Plan de adopción en 90 días

Priorizado por reducción de riesgo, no por facilidad de despliegue

Días 0 – 30

Visibilidad

- Inventariar aplicaciones, agentes y proveedores LLM en uso, incluida la IA en la sombra
- Desplegar FortiDLP para ver qué datos salen hoy hacia herramientas de GenAI
- Levantar el AIBOM: modelos, datasets, adaptadores y servidores MCP
- Aplicar la prueba de la trifecta letal a cada caso de uso agéntico

Días 31 – 60

Control

- Poner FortiAIGate como punto de paso obligatorio y retirar las llaves de API de las aplicaciones
- Activar Input y Output Guard en modo Alert y calibrar umbrales con tráfico real
- Migrar las credenciales de la plataforma de IA a la bóveda de FortiPAM y activar rotación
- Recortar herramientas y permisos de cada agente al mínimo que la tarea exige

Días 61 – 90

Endurecimiento

- Pasar los guards de Alert a Alert & Deny o Redact donde el falso positivo sea tolerable
- Topes de gasto, rate limiting y circuit breakers por aplicación y por sesión
- Aislamiento por inquilino y cifrado del índice vectorial; borrar embeddings con su origen
- Red teaming adaptativo y ejercicio de respuesta ante una fuga de contexto

Regla práctica: si un control solo reduce la probabilidad de que engañen al modelo, es un complemento. Si acota lo que pasa cuando ya lo engañaron, es la defensa.


FORTINET