Cómo proteger tu servidor MCP en la era de los agentes de IA

En Latinoamérica los agentes de IA ya están conectados a sistemas de producción. El problema es que casi nadie está protegiendo la pieza que los hace posibles. Te lo explicamos en lenguaje claro —y con el detalle técnico que tu equipo necesita

NEKT

La inteligencia artificial dejó de ser un experimento. En toda la región —banca, seguros, retail, gobierno— las organizaciones están conectando agentes de IA directamente a sus sistemas para consultar datos en tiempo real, tomar decisiones y ejecutar tareas que antes requerían manos humanas.

El motor silencioso detrás de esa revolución tiene nombre: el **Model Context Protocol (MCP)**. Y como toda pieza nueva y crítica, llegó mucho antes que los controles que deberían protegerla.

El eslabón que casi nadie está mirando

Un servidor MCP es la capa de integración que permite a un modelo de IA **descubrir y usar herramientas** —bases de datos, APIs, sistemas internos— mediante lenguaje natural. Es, en la práctica, un traductor universal entre el "cerebro" de la IA y todo aquello que puede accionar.

Ahí está el riesgo: cuando le das a un agente la capacidad de actuar sobre tus sistemas, también se la das a quien logre engañarlo. El MCP no solo entrega información; puede disparar acciones. Eso cambia por completo lo que significa "asegurar una API".

Por qué el tráfico MCP no se parece a nada que ya proteges

El tráfico de una API tradicional es predecible y de alcance acotado: rutas conocidas, formatos rígidos, comportamiento determinista. El tráfico MCP es lo contrario.

- Descubrimiento dinámico de herramientas: el agente decide en tiempo de ejecución qué puede invocar.

- Interacciones en lenguaje natural: la "carga útil" es texto humano, no parámetros validables.

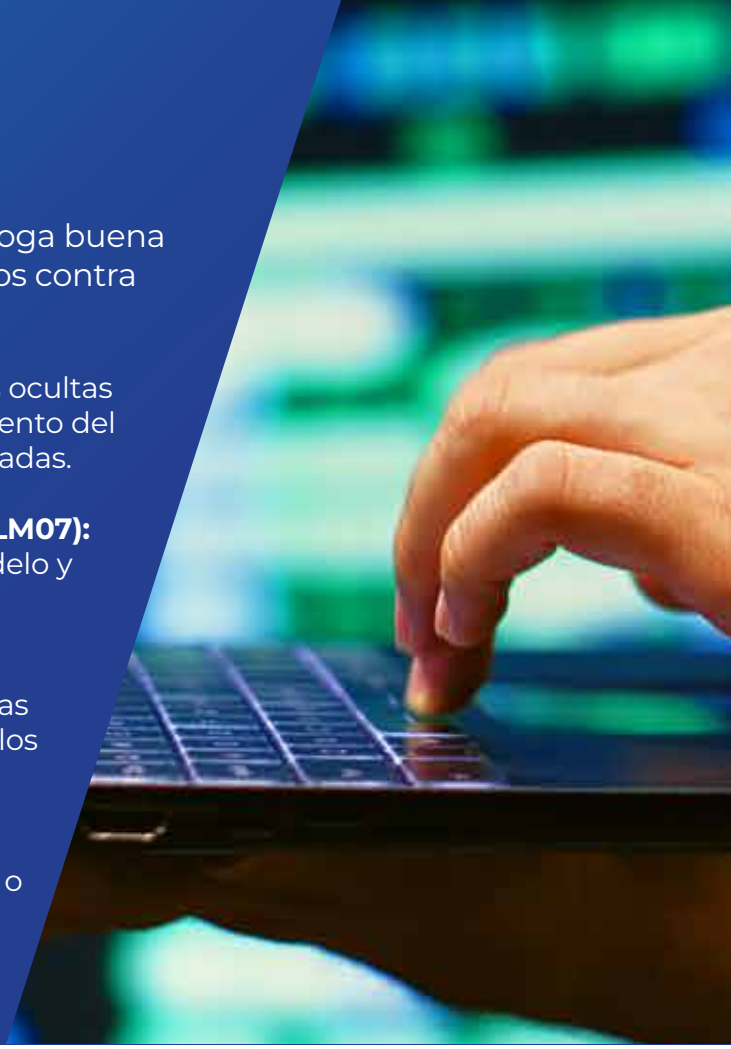
- Orquestación entre fronteras de confianza: una sola petición puede cruzar varios sistemas con distintos permisos.

Cada una de esas características amplía la superficie de ataque y exige controles que van más allá de la protección de API convencional.

Los nuevos vectores de ataque

El **OWASP Top 10 para Aplicaciones LLM** ya cataloga buena parte de estos riesgos. Estos son los que más vemos contra servidores MCP:

- **Inyección de prompts (OWASP LLM01):** instrucciones ocultas —incluso codificadas— que secuestran el comportamiento del agente para filtrar datos o ejecutar acciones no autorizadas.
- **Jailbreak y evasión de guardrails (OWASP LLM01 / LLM07):** frases diseñadas para saltarse las restricciones del modelo y desbloquear capacidades que deberían estar vetadas.
- **Abuso de herramientas y agencia excesiva (OWASP LLM06 / LLM08):** el agente encadena acciones legítimas para lograr un fin malicioso, o recibe más permisos de los que necesita.
- **Enumeración y scraping por bots (OWASP LLM10):** automatización que mapea tus herramientas expuestas o agota recursos con consumo desmedido.



Defensa por capas: la arquitectura que recomendamos

Ningún control aislado protege un servidor MCP. En NEKT Group integramos tecnología de vanguardia —incluyendo la plataforma de **Fastly**, con quien se desarrolló un prototipo de referencia para un editor financiero global— en una arquitectura de capas que se refuerzan entre sí:

1. **Reglas en el borde (edge):** el tráfico se filtra al entrar a la red: se bloquean tipos de contenido inesperados (como XML) y se restringen los métodos HTTP permitidos.
2. **Gestión de bots:** visibilidad del tráfico automatizado para bloquear clientes no deseados, huellas maliciosas y bots con mal comportamiento.
3. **WAF de nueva generación + rate limiting:** mitiga ataques OWASP y específicos de LLM —inyección de prompts codificada, frases de jailbreak— y aplica límites por IP o por API key que exceden lo normal.
4. **Autenticación flexible:** convivencia de modelos por IP, por token y por mTLS —sin falsos positivos— para que cada agente demuestre quién es.
5. **Anclaje de tráfico en un POP:** un punto de presencia actúa como ancla antes del origen: reduce la carga y acelera las conexiones al eliminar handshakes costosos.

Resultados medidos en 10 semanas de prueba

- Latencia promedio **< 50 ms**, contra un objetivo de 100 ms.
- Throughput de carga sostenido de **100 Mbps+**, según lo exigido.
- Payload máximo de **500 MB+** procesado sin timeouts.
- **3 modelos de autenticación** (IP, token y mTLS) conviviendo sin falsos positivos.

La otra mitad del problema: gobierno de la IA

La tecnología resuelve la mitad. La otra mitad es **gobernanza**: poder demostrar cómo tomas decisiones con IA y quién responde por ellas. En Latinoamérica, donde las leyes de protección de datos se endurecen año con año, esto ya no es opcional.

- **ISO/IEC 42001** — Sistema de gestión de IA (SGIA).
- **EU AI Act** — clasificación por riesgo y trazabilidad.
- **ISO/IEC 27001** — SGSI como base de seguridad.
- **NIST AI RMF** — gestión de riesgo del ciclo de vida de IA.

Un servidor MCP protegido y un marco de gobierno como **ISO/IEC 42001** o el **EU AI Act** son las dos caras de la misma moneda: sin controles técnicos no hay cumplimiento posible, y sin gobierno los controles no escalan.

NEKT y Fastly en ANDICOM 2026

Este artículo forma parte de las iniciativas de colaboración entre NEKT y Fastly para compartir estrategias que permitan a las organizaciones proteger sus aplicaciones, APIs y servicios digitales frente a las amenazas actuales.

ANDICOM
2 - 4 SEPTIEMBRE 2026

¿Asistirás a ANDICOM 2026?

Visítanos para conversar sobre cómo fortalecer la seguridad y el rendimiento de tus aplicaciones con una arquitectura preparada para los desafíos de la IA y las amenazas modernas.

NEKT | fastly

Protegiendo la información de nuestros clientes en México y Latinoamérica.