Agente IA de Evaluación de Seguridad en Plataformas de IA (AESA-IA): Así se crea el Protocolo de Gobernanza y Prevención de Fuga de Datos Confidenciales

Introducción y Alcance del Agente de Evaluación (AESA-IA)

El rápido despliegue de soluciones de Inteligencia Artificial (IA) de terceros, incluyendo herramientas de uso común como COPILOT, GEMINI, GROK, ZAPIER, PERPLEXITY, CLAUDE, CHATGPT, MANUS, y plataformas colaborativas como CANVA y MIRO, ha generado brecha crítica entre la adopción tecnológica y los marcos de seguridad y gobernanza corporativos.[1] Esta situación, conocida como *Shadow AI*, expone a las organizaciones a riesgos inaceptables de fuga de datos confidenciales (DLP), exfiltración, y litigios por incumplimiento normativo.

El **Agente de Evaluación de Seguridad en Plataformas de IA (AESA-IA)** se establece como el artefacto normativo y la política operacional diseñada para cerrar esta brecha. Su objetivo principal es proporcionar criterios inequívocos para identificar soluciones de IA seguras, formalizar la debida diligencia de terceros (TPRM) y entregar un *checklist* accionable a los colaboradores, asegurando que la información confidencial (IC) solo se comparta con plataformas que cumplan con los más altos estándares de gobernanza y control de datos.

Parte I: Marco de Gobernanza de la IA y Gestión del

Riesgo Empresarial

La seguridad de la IA no es una mera función técnica; es un imperativo de gobernanza que debe anclarse en principios estratégicos y regulatorios.

1.1. Fundamentos Estratégicos y Definiciones Operacionales

1.1.1. El Imperativo de la Gobernanza de la IA: Contrarrestando el "Shadow AI"

La adopción de IA, particularmente la IA generativa, requiere que las organizaciones evolucionen de enfoques informales o *ad hoc* a una **Gobernanza Formal**.[1] La falta de supervisión permite que los empleados utilicen herramientas no sancionadas, introduciendo datos sensibles sin control, una vulnerabilidad que puede anular defensas cibernéticas maduras.[2]

Para mitigar la proliferación de *Shadow AI*, la organización debe formalizar el uso de Herramientas de Inteligencia Artificial (HIA). Esto implica establecer un Comité de Seguridad de la Información e Inteligencia Artificial que se haga responsable de la selección, validación y aprobación de todas las HIA corporativas.[3] Al formalizar la selección, se obliga a la empresa a mantener un inventario y un control riguroso sobre las soluciones de IA empleadas, previniendo que herramientas de consumo sin garantías contractuales accedan a los datos internos.[4] La eficacia de esta estrategia radica en establecer el proceso de selección formal como la primera línea de defensa.

1.1.2. Definición y Clasificación de la Información Confidencial (IC)

La política DLP es ineficaz sin una clasificación de datos clara.[5] Se considera Información Confidencial (IC) a todo dato que, de ser expuesto, causaría un daño material a la organización o violaría las regulaciones de privacidad. Esto incluye Datos de Identificación Personal (PII), secretos comerciales, propiedad intelectual (IP), información financiera y

credenciales. La directriz es clara: el colaborador debe saber exactamente qué tipo de información **nunca** debe ingresar en una plataforma de IA sin la aprobación explícita del Comité de Seguridad.

Desde una perspectiva técnica, la seguridad de los datos exige medidas proactivas, como el cifrado de datos confidenciales tanto en tránsito como en reposo, especialmente cuando se gestionan conjuntos de datos utilizados para el entrenamiento de IA.[6]

1.2. Adopción de Marcos de Referencia Internacionales y Locales

La estructura del AESA-IA se alinea con marcos internacionales para garantizar la solidez y la coherencia regulatoria.

1.2.1. Alineación con el NIST AI Risk Management Framework (AI RMF)

El NIST AI RMF proporciona un enfoque sistemático y sociotécnico para la gestión de riesgos en IA, enfatizando la rendición de cuentas, la transparencia y el comportamiento ético.[7] El AESA-IA opera bajo las cuatro funciones fundamentales del NIST AI RMF:

1. **Govern:** Establece los estándares, principios éticos y la estructura de responsabilidad que rigen la política (Parte I del informe).
2. **Map:** Identifica los riesgos específicos inherentes a las plataformas LLM y de integración utilizadas (Parte II del informe).
3. **Measure:** Define los requisitos técnicos y contractuales de los proveedores (Parte III: TPRM).[8]
4. **Manage:** Implementa los controles operacionales y las mitigaciones diarias a través del *checklist* del colaborador (Parte IV).

El riesgo de la IA no se limita a fallas técnicas, sino que abarca complejidades sociales, legales y éticas.[7] Por lo tanto, la implementación de controles puramente técnicos, como el cifrado, es insuficiente si no está respaldada por un marco ético y de rendición de cuentas (NIST Govern) que garantice que los sistemas operan dentro de los límites legales y corporativos.

1.2.2. Complemento de ISO/IEC 42001 y 27001

La debida diligencia de terceros debe exigir la adherencia a estándares de gestión. La norma ISO/IEC 42001 establece los requisitos para un Sistema de Gestión de Inteligencia Artificial (SGIA), el cual debe ser compatible e integrarse con el Sistema de Gestión de la Seguridad de la Información (SGSI) basado en ISO/IEC 27001.[9] Exigir el cumplimiento de estas normas asegura que el proveedor no solo gestione los riesgos generales de seguridad de la información, sino que también evalúe y gestione específicamente los riesgos únicos de la IA (como el sesgo o la explicabilidad).[10]

1.2.3. Cumplimiento Normativo (GDPR y Regional)

El uso responsable de la IA debe adherirse estrictamente a los marcos de protección de datos. El Reglamento General de Protección de Datos (GDPR) en Europa establece altos requisitos de transparencia, legalidad y consentimiento, afectando directamente a cómo las plataformas de IA manejan los datos personales.[11]

A nivel regional (por ejemplo, en Colombia), la legislación está evolucionando para alinearse con principios éticos y de seguridad de la IA, lo que requiere que las organizaciones implementen programas integrales de gestión de datos personales.[12] Cualquier evaluación de proveedor debe considerar el impacto de la Ley de IA de la UE y las normativas locales.[14] Un despliegue seguro de IA requiere que los sistemas sean seguros, éticos y legales, lo que implica que el TPRM debe validar el diseño ético y contractual antes de proceder con el análisis de la solidez operativa.

Parte II: Análisis Profundo de Amenazas Específicas en Plataformas de IA

El AESA-IA debe basarse en una comprensión granular de las amenazas únicas que los modelos de lenguaje grande (LLM) y las plataformas de orquestación introducen, siendo el denominador común el riesgo directo o indirecto de la fuga de Información Confidencial.

2.1. Riesgos de Seguridad Inherentes a los Modelos de Lenguaje Grande (LLM)

2.1.1. Inyección de Prompts y Manipulación de Entrada

La Inyección de Prompts, clasificada como la amenaza número 1 en el OWASP Top 10 para Aplicaciones LLM, representa un riesgo significativo de manipulación.[15] Un atacante puede utilizar avisos ingeniosamente diseñados para obligar al modelo a eludir sus filtros de seguridad, revelar información sensible, o incluso exponer el *prompt* original del sistema.[17]

La mitigación principal se centra en la **validación y desinfección de la entrada**. Esto incluye el diseño de *prompts* estructurados que separen claramente las instrucciones del sistema de la entrada del usuario, así como el monitoreo y filtrado de la salida. Las inyecciones a menudo emplean entradas largas y elaboradas o imitan la sintaxis del sistema para engañar al LLM, por lo que la detección de longitud excesiva o similitudes con ataques conocidos es una defensa clave.[17]

2.1.2. Riesgos a la Integridad y Confidencialidad del Modelo

Además de la inyección, existen amenazas que comprometen la base del modelo. Los ataques de **Inversión de Modelos** buscan deducir los datos de entrenamiento subyacentes, lo que podría exponer inadvertidamente datos de clientes o secretos corporativos utilizados en la fase de entrenamiento.[16] El **Envenenamiento del Modelo** compromete la integridad al manipular los datos de entrenamiento para inyectar sesgos o vulnerabilidades.[19]

2.1.3. Alucinaciones y Pérdida de Providencia de Datos

La generación de información falsa o engañosa por parte de los LLM (alucinaciones) introduce un riesgo operativo que, en el contexto de decisiones críticas (ej. *underwriting* o manejo de reclamaciones), puede tener graves consecuencias.[2] Esto subraya la necesidad de

que los colaboradores utilicen su criterio y conozcan las limitaciones de la IA.[4] Un problema relacionado es la **pérdida de proveniencia de datos**, donde los resultados generados por IA no pueden ser trazados a una fuente validada y original, socavando la confianza y el cumplimiento regulatorio.[2]

2.2. El Riesgo Crítico: Fuga de Datos por Entrenamiento del Modelo (DLP)

El vector más peligroso de fuga de datos en plataformas de IA generativa es la utilización del *input* del usuario para mejorar los modelos fundacionales.

2.2.1. Exfiltración por Monetización y Uso de Plataformas No Sancionadas

El uso descontrolado de plataformas de IA, especialmente las gratuitas o de prueba, implica un riesgo inherente de exfiltración, ya que estas plataformas a menudo monetizan los datos del usuario o los utilizan para la mejora de modelos sin garantías contractuales adecuadas.[2] La validación de la política de retención y eliminación de datos es, por lo tanto, una medida fundamental de DLP.[5]

2.2.2. El Conflicto de las Políticas de Uso de Datos

Existe una diferencia crucial en el uso de los datos entre los distintos proveedores:

- **Zona de Baja Confidencialidad (Riesgo Alto):** Las versiones de consumo de IA, como las "Apps con Gemini", pueden utilizar las conversaciones para mejorar los modelos, incluyendo la posibilidad de que el dato sea analizado por revisores humanos.[20] Aunque el usuario puede desactivar la "Actividad en las Apps con Gemini", el proveedor podría continuar el procesamiento para crear datos *anonimizados* utilizados en la mejora del servicio.[20]
- **Zona de Alta Confidencialidad (Riesgo Mitigado):** Plataformas diseñadas para el entorno empresarial, como Microsoft 365 Copilot, ofrecen garantías contractuales de que los datos de interacción de los usuarios **no** se utilizarán para entrenar los modelos

fundacionales.[21]

Esta distinción en las políticas de entrenamiento es el principal factor de control de riesgo para la IC. Un proveedor puede ser técnicamente robusto (ej. cumplimiento SOC 2), pero si su diseño contractual permite usar datos del cliente para mejorar el modelo, el riesgo de pérdida de control y, por lo tanto, de fuga, persiste a un nivel inaceptablemente alto.

2.3. Riesgos Asociados a Plataformas de Integración y Colaboración

2.3.1. Integraciones Inseguras y Escalada de Privilegios

Plataformas de automatización como ZAPIER, MAKE, y N8N actúan como "tuberías" de datos, conectando múltiples sistemas internos a través de APIs.[16] El riesgo clave aquí no es el entrenamiento del modelo, sino la seguridad del entorno operativo y la gestión de credenciales. La integración insegura puede llevar al *Authentication Drift*, donde las *plugins* de IA se conectan a sistemas corporativos sin una gestión central de credenciales.[2] Peor aún, puede ocurrir una **escalada de privilegios *cross-domain***, permitiendo que el flujo automatizado acceda a datos que exceden los permisos del usuario que inició la conexión.[2]

Para estos sistemas, la defensa debe enfocarse en la **seguridad del entorno operativo** (evidenciada por SOC 2) y en la implementación estricta del Principio del Mínimo Privilegio (PoLP).[5]

2.3.2. Riesgo en Plataformas de Contenido (Miro, Canva)

Las herramientas de colaboración y diseño (Miro, Canva) presentan un riesgo de almacenamiento no controlado. La IC (como tokens de autenticación, números de tarjeta o PII) puede ser copiada inadvertidamente en tableros o diseños, resultando en una dispersión y fuga de datos que complica el cumplimiento.[23] La mitigación en estos entornos se basa en la capacidad de la plataforma Enterprise de activar el **Descubrimiento de Datos Sensibles** (*Data Discovery*) a nivel empresarial, escaneando periódicamente los tableros para clasificar y proteger automáticamente la información de alto riesgo.[23]

La defensa contra el riesgo en las tuberías de datos (Zapier, Make) requiere una estrategia de **segregación de red y aislamiento lógico**.[25] Si una integración es comprometida, la mitigación de daños depende directamente de la capacidad de la infraestructura para aislar el tráfico y el acceso a los datos críticos, por ejemplo, mediante el uso de VLANs.[25] Por lo tanto, el TPRM debe preguntar sobre las capacidades de segregación y aislamiento de la infraestructura de nube del proveedor.[26]

Parte III: Protocolo de Debida Diligencia para Plataformas de IA de Terceros (TPRM - Measure Function del NIST)

El Protocolo TPRM es la función *Measure* del AESA-IA, que traduce los riesgos identificados en requisitos técnicos y contractuales medibles antes de la aprobación de cualquier plataforma.

3.1. Requisitos Mínimos de Seguridad y Conformidad

3.1.1. Verificación de Certificaciones de Seguridad y Auditoría

La aprobación de un proveedor de IA se requiere sobre una base de evidencia de auditoría independiente. La certificación mínima aceptable incluye la **SOC 2 Tipo II** (que cubre los criterios de seguridad, confidencialidad y privacidad) y/o la **ISO/IEC 27001** (Sistemas de Gestión de la Seguridad de la Información).[14]

Se debe otorgar una preferencia significativa a los proveedores que demuestren adhesión a la **ISO/IEC 42001**, que certifica específicamente la gestión de los riesgos asociados a los sistemas de IA.[10] Las plataformas que actúan como procesadores de datos, como Firebase, ya establecen el listón con el cumplimiento de ISO 27001 y SOC 1.[26]

3.1.2. Gobernanza de Seguridad del Proveedor

La estructura de gestión de riesgos del proveedor debe ser formal. El TPRM exige la verificación de que el proveedor cuenta con un Director de Seguridad de la Información (CISO) o rol equivalente, un equipo de seguridad dedicado y políticas y procedimientos documentados de ciberseguridad, incluyendo formación regular para empleados.[28]

3.2. Controles Técnicos Obligatorios y Estrategias DLP

3.2.1. Gestión de Identidad, Acceso y Privilegios

La plataforma debe integrarse con los sistemas corporativos para el control de acceso, requiriendo el uso de Autenticación Única (SSO) y Autenticación Multifactor (MFA).[6] Esto es vital para prevenir el *Authentication Drift*. Se exige la aplicación rigurosa del **Principio del Mínimo Privilegio (PoLP)** en todas las capas del servicio: en las cuentas de usuario y, fundamentalmente, en las configuraciones de API e integraciones de terceros (ej. Zapier, Make).[5]

3.2.2. Cifrado, Retención y Segregación de Datos

El cifrado es obligatorio para todos los datos confidenciales en reposo y en tránsito.[5] Además, para entornos multi-inquilino, se debe validar la capacidad de la plataforma para garantizar la **segregación y aislamiento de datos** a nivel de cliente. La política de retención y eliminación de datos debe ser definida y divulgada, garantizando que la organización retiene el control sobre el ciclo de vida de su IC.[5]

3.2.3. Control de Fuga Específico de LLM

Para los LLM, se debe exigir que el proveedor implemente:

1. **Validación de Prompts:** Mecanismos de seguridad a nivel de API o interfaz para prevenir la Inyección de Prompts.[15]
2. **Visibilidad de Auditoría:** La capacidad de los administradores corporativos para auditar y rastrear las interacciones del usuario con la IA, especialmente en entornos integrados (ej. Copilot Chats, donde los administradores pueden ver los datos almacenados de interacción [29]).

3.3. Matriz de Riesgo Contractual: El Uso de Datos para Entrenamiento

El criterio de aceptación más crítico en el TPRM es el uso que el proveedor da a los datos de entrada del cliente.

Checklist de Aprobación de Seguridad en Plataformas de IA (TPRM - CISO/Comité)

Dominio de Seguridad (NIST RMF Measure)	Criterio de Aceptación Crítico	Evidencia Requerida (Soporte)
1. Gobernanza y Cumplimiento	¿El proveedor ha presentado una certificación SOC 2 Tipo II o ISO/IEC 27001/42001 vigente?	Informe de auditoría SOC 2/Certificado ISO válido.[10]
2. Confidencialidad de Datos	¿Existe una cláusula contractual que garantice que los datos de entrada **no** se utilizan para el entrenamiento futuro del modelo?	Cláusula contractual específica de uso de datos, firmada por el departamento Legal.[20]
3. Seguridad Técnica	¿La plataforma soporta la	Documentación técnica de IAM y configuración de

(IAM)	gestión de Identidad y Acceso Corporativa (SSO/MFA) y aplica PoLP?	privilegios mínimos.[6]
4. Arquitectura DLP	¿El proveedor ofrece segregación de datos a nivel de cliente (Data Isolation) y cifrado de datos en reposo y tránsito?	Política de retención, esquema de cifrado y evidencia de cumplimiento de segregación de datos.[25]
5. Mitigación de Riesgos LLM	Si es un LLM, ¿se implementan controles internos de validación/desinfección de entrada para mitigar la Inyección de Prompts?	Documentación de la arquitectura de seguridad LLM y cumplimiento con guías OWASP.[15]

La aplicación de esta matriz define las zonas de riesgo:

1. **Zona Verde (Aprobación Alta):** Plataformas que garantizan contractualmente que los datos de entrada de la empresa nunca se utilizan para entrenar o mejorar modelos fundacionales (ej. Microsoft 365 Copilot).[21]
2. **Zona Amarilla (Aprobación Condicional):** Plataformas que requieren configuraciones de *opt-out* explícitas o anonimización, donde el riesgo residual requiere monitoreo continuo y la restricción estricta de la IC (ej. Gemini Apps).[20]
3. **Zona Roja (No Aprobación):** Plataformas de consumo o prueba que utilizan los datos para mejorar servicios o donde el dato es visible para revisores humanos sin control de auditoría.[2]

Parte IV: AESA-IA Checklist Operacional para Colaboradores (El Artefacto)

Esta sección proporciona el artefacto final: el *Checklist* que los colaboradores deben aplicar antes de interactuar con cualquier plataforma de IA, transformando la política en práctica.

4.1. Directrices de Uso y Responsabilidades del Colaborador

4.1.1. Principio de Precaución y Clasificación de Datos

La regla operativa es que toda plataforma de IA debe considerarse **no segura** para la Información Confidencial (IC) hasta que haya sido formalmente aprobada por el Comité de Seguridad de la IA y figure en el inventario oficial de HIA Corporativas.[3] El desconocimiento del contenido de esta política no exime al colaborador de su cumplimiento.[3]

4.1.2. Supervisión Humana (Human-in-the-Loop - HITL)

Los colaboradores son el control de seguridad final. Siempre deben utilizar su criterio para interpretar y actuar sobre las recomendaciones de la IA.[4] La implementación del *Human-in-the-Loop* previene que el LLM ejecute acciones maliciosas o acceda a datos sin aprobación, mitigando el riesgo de ejecución inherente a los ataques de inyección de *prompts*.[17]

La defensa conductual (formación en clasificación y prohibición de *prompts*) es la capa de protección más fuerte, ya que las tecnologías DLP tienen dificultades para interceptar la transferencia manual de pequeños volúmenes de texto confidencial introducido en una interfaz de chat.

4.2. Checklist Operacional de Seguridad (Antes, Durante y Después de la Interacción)

Riesgo Específico (LLM/Integración)	Impacto en la Confidencialidad (DLP)	Control de Colaborador (Acción	Control de la Plataforma (Requisito TPRM)

		Inmediata)	
Inyección de Prompts	Manipulación para revelar datos sensibles.[16]	Usar *prompts* estructurados; Evitar datos no autorizados; Limitar la longitud de la entrada.[18]	Validación y Desinfección de Entradas/Salidas; Monitoreo de Salida.[15]
Fuga de Datos por Entrenamiento	Datos ingresados retenidos/usados para mejora de modelo, expuestos a revisores humanos.[2]	Usar **exclusivamente** versiones Enterprise con Cláusula de No-Entrenamiento.[21]	Política contractual de No Uso de Datos de Usuario para Entrenamiento (Zona Verde).
Cross-Domain Privilege Escalation	Herramienta de IA o integración accede a sistemas internos con privilegios excesivos.[2]	Revisar y limitar los permisos de todas las integraciones (plugins/APIs); Asegurar PoLP.	Aplicación rigurosa del Principio del Mínimo Privilegio (PoLP).[5]
Alucinaciones y Providencia	Generación de información falsa; decisiones operativas basadas en datos no validados.[2]	No actuar sobre las recomendaciones sin validación humana experta.[4]	Mecanismos de trazabilidad de las fuentes de datos.

Fase 1: Preparación (Pre-Uso de la Plataforma)

1. **Verificación de Aprobación:** Consultar el inventario oficial de HIA Corporativas. Si la plataforma (ej. GROK, DeepSeek, Gamma) no está en la lista de Aprobación Alta, **la interacción se detiene.**
2. **Clasificación de Datos:** Determinar la clasificación de la información que se planea introducir. Si es IC (PII, IP, Credenciales), la entrada está **estrictamente prohibida** en cualquier plataforma que no esté en la Zona Verde.

3. **Configuración de Privacidad (Zona Amarilla):** Si se utiliza una plataforma de riesgo condicional (ej. Gemini, versiones API de Claude/ChatGPT), confirmar que las configuraciones de *opt-out* para la mejora del modelo (*Actividad en las Apps con Gemini*) estén desactivadas y que se opere bajo una licencia Enterprise verificada.[20]

Fase 2: Interacción Segura (Formulación del Prompt)

1. **Prohibición Absoluta de IC: Nunca** introducir PII, contraseñas, tokens de autenticación o secretos comerciales en el *prompt* ni en el texto de entrada.
2. **Estructura del Prompt:** Utilizar siempre un formato estructurado, separando claramente las instrucciones para el sistema del dato de entrada. Esto eleva la barrera contra los ataques de inyección.[18]
3. **Principio de la Mínima Data:** Limitar la entrada al mínimo absoluto de datos necesarios para que la IA complete la tarea. Minimizar la longitud de la entrada, ya que las inyecciones suelen utilizar entradas largas para eludir las defensas.[17]

Fase 3: Post-Interacción y Auditoría (Salida y Archivo)

1. **Validación de la Salida:** Siempre ejercer la supervisión humana (*HITL*) al revisar la salida. Verificar la precisión de la información generada (mitigación de alucinaciones) y confirmar que la salida no contenga información sensible del sistema (*prompt* secreto) que pudiera haber sido filtrada por un ataque de inyección.[4]
2. **Gestión de Plataformas de Colaboración:** Si se utilizan plataformas de contenido (Miro, Canva), el colaborador debe confirmar que la función de **Descubrimiento de Datos Sensibles** de la empresa (Enterprise Guard) está activa para escanear tableros o archivos en busca de IC.[23]
3. **Auditoría de Integraciones (Zapier/Make/n8n):** Cuando se configura un flujo de trabajo de automatización, el colaborador actúa como el "propietario de datos" de ese flujo. Debe documentar el flujo de datos y requerir autorización formal para la exportación de datos fuera de la red corporativa.[25] Revisar periódicamente los permisos de las integraciones para asegurar que se mantiene el Principio del Mínimo Privilegio (PoLP), mitigando la escalada de privilegios *cross-domain*.[2]

Conclusiones y Recomendaciones

La gestión segura de las Plataformas de IA de terceros requiere una aproximación que equilibre la velocidad de la innovación con una gobernanza formal y rigurosa. El AESA-IA, al integrar los requisitos del NIST AI RMF con los estándares de cumplimiento (ISO 42001, SOC 2), ofrece un marco sistemático para la evaluación de riesgos.

Se establece que la principal defensa contra la fuga de datos confidenciales no es puramente técnica (como el cifrado, que es una capa de control esencial), sino **contractual y conductual**. La cláusula de **No Uso de Datos de Cliente para Entrenamiento del Modelo Fundacional** es el criterio de aceptación más crítico, definiendo la 'Zona Verde' de seguridad para los LLM. Para las plataformas de orquestación y flujo de datos (como Zapier o Make), el foco del riesgo se desplaza a la **seguridad del entorno operativo y la segregación lógica** de la red.

El AESA-IA concluye con la necesidad de que la organización:

1. **Formalice el Comité de Seguridad de IA:** Otorgándole la autoridad para mantener el inventario oficial de HIA Corporativas Aprobadas.
2. **Exija Garantías Contractuales:** Nunca aprobar una plataforma de LLM para IC sin una garantía contractual de No-Entrenamiento.
3. **Implemente Capacitación Comportamental:** Reforzar la formación del colaborador en la clasificación de la información (PII, IP, etc.) y en la prohibición de introducir IC en *prompts*, reconociendo que el control humano es la defensa más efectiva contra la inyección y la fuga de datos de bajo volumen.

Fuentes citadas

1. acceso: noviembre 11, 2025, https://www.ibm.com/mx-es/think/topics/ai-governance
2. Intersys: Awareness, then Action – containing re/insurance's shadow AI threat, acceso: noviembre 11, 2025, https://www.globalreinsurance.com/home/intersys-awareness-then-action-containing-re/insurances-shadow-ai-threat/1456878.article
3. Política de Desarrollo y uso responsable de la Inteligencia Artificial - Grupo Elecnor, acceso: noviembre 11, 2025, https://www.grupoelecnor.com/storage/media/files/shares/Responsabilidad_Corporativa/politica-de-desarrollo-y-uso-responsable-de-la-ia-01-2025-es.pdf
4. POLÍTICA IA - ISMS Forum Spain, acceso: noviembre 11, 2025, https://www.ismsforum.es/ficheros/descargas/politica-ia1700038318.pdf
5. ¿Qué es DLP (prevención de pérdida de datos)? - IBM, acceso: noviembre 11,

2025, https://www.ibm.com/mx-es/think/topics/data-loss-prevention
6. ¿Qué es Seguridad de IA? Protección de sistemas de IA - Microsoft, acceso: noviembre 11, 2025, https://www.microsoft.com/es-es/security/business/security-101/what-is-ai-security
7. NIST AI Risk Management Framework (AI RMF) - Palo Alto Networks, acceso: noviembre 11, 2025, https://www.paloaltonetworks.com/cyberpedia/nist-ai-risk-management-framework
8. Governing AI Risk with the NIST AI RMF - BigID, acceso: noviembre 11, 2025, https://bigid.com/blog/governing-ai-risk-with-the-nist-ai-rmf/
9. Definición Norma ISO 42001 - ORSYS, acceso: noviembre 11, 2025, https://www.orsys.fr/orsys-lemag/es/glosario/norma-iso-42001-%F0%9F%9F%A6/
10. ISO 42001 - Gestión de la Inteligencia Artificial - ISOTools, acceso: noviembre 11, 2025, https://www.isotools.us/normas/inteligencia-artificial/iso-42001/
11. ¿Es seguro ChatGPT GDPR ? | Alumio, acceso: noviembre 11, 2025, https://www.alumio.com/es/blog/is-chatgpt-gdpr-safe
12. Protección de datos en Colombia: sanciones, nuevas reglas de la SIC e impacto de la inteligencia artificial | News - Holland & Knight, acceso: noviembre 11, 2025, https://www.hklaw.com/es/news/intheheadlines/2025/08/data-protection-in-colombia-sanctions-new-sic-rules
13. "Por medio del cual se regula la inteligencia artificial en Colombia para garantizar su desarrollo ético y responsable y s - Minciencias, acceso: noviembre 11, 2025, https://minciencias.gov.co/sites/default/files/upload/noticias/pl_ia_finalizado.pdf
14. Cumplimiento del SOC 2 | Soluciones - OneTrust, acceso: noviembre 11, 2025, https://www.onetrust.com/es/solutions/soc-2-compliance/
15. OWASP Top 10 for LLM Applications 2025: Prompt Injection - Check Point, acceso: noviembre 11, 2025, https://www.checkpoint.com/es/cyber-hub/what-is-llm-security/prompt-injection/
16. Seguridad de los LLM: riesgos, mejores prácticas y soluciones | Proofpoint ES, acceso: noviembre 11, 2025, https://www.proofpoint.com/es/blog/dspm/llm-security-risks-best-practices-solutions
17. Cómo prevenir los ataques de inyección de prompt - IBM, acceso: noviembre 11, 2025, https://www.ibm.com/es-es/think/insights/prevent-prompt-injection
18. LLM Prompt Injection Prevention - OWASP Cheat Sheet Series, acceso: noviembre 11, 2025, https://cheatsheetseries.owasp.org/cheatsheets/LLM_Prompt_Injection_Prevention_Cheat_Sheet.html
19. Challenges of AI | Office of the Provost | Washington State University, acceso: noviembre 11, 2025, https://provost.wsu.edu/challenges-of-ai/
20. Centro de privacidad de las Apps con Gemini - Ayuda de ..., acceso: noviembre 11, 2025, https://support.google.com/gemini/answer/13594961?hl=es-419
21. Microsoft 365 Copilot Chat Privacy and Protections, acceso: noviembre 11, 2025, https://learn.microsoft.com/en-us/copilot/privacy-and-protections

22. Comparativa n8n vs Make vs Zapier: ¿Cuál es la Mejor Herramienta de Automatización para tu Empresa en 2025? - Juan Carlos Mejía, acceso: noviembre 11, 2025, https://www.juancmejia.com/transformacion-digital/comparativa-n8n-vs-make-vs-zapier-cual-es-la-mejor-herramienta-de-automatizacion-para-tu-empresa-en-2025/
23. Activar el descubrimiento de datos relacionados con la privacidad - Miro Help Center, acceso: noviembre 11, 2025, https://help.miro.com/hc/es/articles/15457957080082-Activar-el-descubrimiento-de-datos-relacionados-con-la-privacidad
24. Seguridad de IA en Canva, acceso: noviembre 11, 2025, https://www.canva.com/es_us/policies/ai-safety/
25. MANUAL DE POLÍTICAS Y LINEAMIENTOS DE SEGURIDAD DE LA INFORMACIÓN - DIAN, acceso: noviembre 11, 2025, https://www.dian.gov.co/atencionciudadano/LMDP/Informacion-Innovacion-y-Tecnologia/Seguridad-de-la-Informacion/Manuales/MN-IIT-0072.pdf
26. Privacidad y seguridad en Firebase - Google, acceso: noviembre 11, 2025, https://firebase.google.com/support/privacy?hl=es-419
27. Requisitos de Cumplimiento SOC 2 - Secureframe, acceso: noviembre 11, 2025, https://secureframe.com/es-es/hub/soc-2/requirements
28. THIRD-PARTY VENDOR CYBERSECURITY DUE DILIGENCE CHECKLIST - Baker Donelson, acceso: noviembre 11, 2025, https://www.bakerdonelson.com/webfiles/Publications/Third_Party%20Vendor%20Cybersecurity%20Due%20Diligence%20Checklist_09-25-25.pdf
29. Data, Privacy, and Security for Microsoft 365 Copilot, acceso: noviembre 11, 2025, https://learn.microsoft.com/en-us/copilot/microsoft-365/microsoft-365-copilot-privacy