# Panorama de la seguridad de la IA

Una visión documentada de los marcos públicos de seguridad, las divulgaciones a nivel de modelo y los enfoques institucionales ante el riesgo de la IA avanzada.

Publicado: 9 ago 2026 · Fecha de la evidencia: 9 ago 2026

Editor: AI Governance World LLC · aigovernanceworld.com

# Resumen ejecutivo

El panorama inicial combina marcos públicos de gestión de riesgos con fichas de modelos, fichas de sistemas, políticas de desarrolladores y registros de incidentes verificados. Estas capas de evidencia responden a preguntas diferentes. Un marco público describe un enfoque de gestión de riesgos; una divulgación de un desarrollador indica lo que declara su editor; una evaluación independiente prueba un modelo conforme a un método documentado; y un registro de ejecución indica lo que una autoridad constató o exigió.
El corpus contiene actualmente documentación de modelos redactada por los desarrolladores, pero ninguna evaluación independiente revisada que esté vinculada a los registros de modelos iniciales. Esta ausencia se muestra de forma explícita y no se sustituye por una calificación indirecta. Ningún modelo ni empresa recibe una puntuación agregada de seguridad basada únicamente en documentación de primera parte.

# Metodología

La evidencia admisible comprende marcos oficiales y registros de ejecución de nivel 1, evaluaciones académicas de nivel 2 con métodos atribuibles, y fichas de modelos, fichas de sistemas e informes de seguridad de nivel 3 redactados por los desarrolladores. Los medios pueden aportar contexto para el descubrimiento, pero nunca constituyen evidencia primaria.
Las afirmaciones de los desarrolladores se identifican mediante su editor y tipo de fuente, y permanecen vinculadas al registro del modelo o de la empresa como divulgaciones de primera parte. Las evaluaciones independientes se guardan en el registro de evaluaciones de modelos con el evaluador, la fecha de evaluación, la metodología, los resultados, los hallazgos y las fuentes. Las dos clases de evidencia se presentan por separado y nunca se fusionan en una calificación de seguridad sin respaldo.

## Marcos públicos de seguridad y riesgo

Los marcos se comparan por su cobertura del ciclo de vida, gobernanza, identificación, medición y tratamiento de riesgos, seguimiento, gestión de incidentes y mejora continua. El NIST AI RMF y la norma ISO/IEC 42001 cumplen funciones diferentes: el corpus registra su alcance oficial y sus características de evaluación de la conformidad, en lugar de tratarlos como certificaciones intercambiables.
Los instrumentos de la OCDE y la UNESCO aportan contexto normativo y de políticas públicas. Su presencia en el corpus no demuestra la adopción o aplicación por una empresa o país concreto sin evidencia separada.

## Divulgaciones de los desarrolladores sobre modelos y sistemas

Los registros de modelos conservan las capacidades declaradas, los enfoques de evaluación, las limitaciones, las medidas de mitigación, las condiciones de acceso y las fechas de actualización, sin completar campos no documentados. Se muestra el editor para que los lectores puedan reconocer la evidencia de primera parte.
La declaración de un desarrollador de que se realizaron pruebas es una afirmación verificable. Por sí sola, no constituye confirmación independiente del resultado, de la integridad del conjunto de pruebas ni de la eficacia de las medidas de mitigación.

## Límites del aseguramiento

Un marco publicado es evidencia de un compromiso de gobernanza, no prueba de su aplicación ni de la seguridad del modelo. Una ficha de modelo es evidencia de divulgación, no una certificación. Una evaluación

independiente está limitada por su método y la versión probada, mientras que un hallazgo regulatorio está limitado por su jurisdicción y actuación jurídica.
Por estas razones, el Panorama de seguridad publica recuentos de evidencia y categorías de procedencia, no una clasificación global de seguridad.

# Cuatro capas de evidencia distintas

El informe separa los marcos públicos, las divulgaciones de los desarrolladores, las evaluaciones independientes y los incidentes o medidas de ejecución. Un registro puede contribuir a más de una pregunta analítica, pero su procedencia no cambia. Por ejemplo, una ficha de sistema sigue siendo una afirmación del desarrollador aunque describa pruebas exhaustivas.
Esta estructura evita confundir el volumen de divulgaciones con una seguridad demostrada de manera independiente e impide generalizar un incidente aislado hasta convertirlo en una conclusión sin respaldo sobre todo un modelo.

# Registro de evaluaciones independientes

Una evaluación independiente requiere un evaluador distinto del desarrollador del modelo, una fecha de evaluación, una metodología documentada, hallazgos atribuibles y una fuente admisible. Los registros que no identifican estos elementos quedan fuera del recuento de evaluaciones independientes revisadas.
El gráfico inicial muestra el recuento actual directamente desde el registro de evaluaciones de modelos. Cero significa que ningún registro vinculado ha completado la verificación; no significa que no exista investigación externa.

# Incidentes, respuesta y ejecución

Los registros de incidentes conservan la fecha del suceso, su gravedad, estado, resumen del impacto, respuesta y fuente primaria. Las medidas de ejecución siguen atribuidas a la autoridad responsable y están vinculadas mediante la cronología. Estos registros pueden fundamentar una evaluación posterior, pero la plataforma no deduce causalidad técnica ni riesgo para todo el modelo más allá de la evidencia oficial.

# Gráficos y datos

## Inventario de evidencia de seguridad

Los marcos, las divulgaciones sobre modelos y los incidentes son clases de registros diferentes.

Registros de marcos públicos: 4
Registros de divulgaciones sobre modelos: 5
Registros de incidentes confirmados: 1
El corpus contiene 4 registros de marcos públicos, 5 registros de divulgaciones sobre modelos y 1 registro de incidente confirmado.

## Divulgación del desarrollador y evaluación revisada

La documentación redactada por los desarrolladores se mantiene separada de las evaluaciones de modelos verificadas.

Fuentes de modelos redactadas por desarrolladores: 5
Evaluaciones de primera parte revisadas: 0

Evaluaciones independientes revisadas: 0

5 registros de modelos utilizan fuentes primarias redactadas por los desarrolladores; 0 evaluaciones de primera parte y 0 evaluaciones independientes han completado la revisión.

## Modos documentados de acceso a los modelos

El modo de acceso procede del registro de modelo verificado y no es una calificación de seguridad.

API: 3

Pesos abiertos: 2

Registros de modelos verificados agrupados por el modo de acceso documentado.

# Referencias


1. Marco de Gestión de Riesgos de Inteligencia Artificial (AI RMF 1.0). Fuente: Instituto Nacional de Estándares y Tecnología. 26 ene 2023. Título original de la fuente: Artificial Intelligence Risk Management Framework (AI RMF 1.0). https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-ai-rmf-10

2. ISO/IEC 42001:2023 — Sistema de gestión de la inteligencia artificial. Fuente: Organización Internacional de Normalización. 1 dic 2023. Título original de la fuente: ISO/IEC 42001:2023 — Artificial intelligence management system. https://www.iso.org/standard/42001.html

3. Principios de la OCDE sobre inteligencia artificial. Fuente: Organización para la Cooperación y el Desarrollo Económicos. 3 may 2024. Título original de la fuente: OECD AI Principles. https://oecd.ai/en/ai-principles

4. Recomendación sobre la Ética de la Inteligencia Artificial. Fuente: UNESCO. 23 nov 2021. Título original de la fuente: Recommendation on the Ethics of Artificial Intelligence. https://www.unesco.org/en/legal-a f airs/recommendation-ethics-artificial-intelligence

5. La NHTSA anuncia una orden de consentimiento con Cruise tras no informar plenamente de un accidente con un peatón. Fuente: Administración Nacional de Seguridad del Tráfico en las Carreteras. 30 sept 2024. Título original de la fuente: NHTSA Announces Consent Order with Cruise A f er Company Failed to Fully Report Crash Involving Pedestrian. https://www.nhtsa.gov/press-releases/consent-order-cruise-crash-reporting

6. Ficha del sistema GPT-5.6. Fuente: OpenAI. 9 jul 2026. Título original de la fuente: GPT-5.6 System Card. https://deploymentsafety.openai.com/

7. Fichas de sistema de los modelos de Anthropic. Fuente: Anthropic. 1 jun 2026. Título original de la fuente: Anthropic model system cards. https://www.anthropic.com/system-cards

8. Ficha del modelo Gemini 3.5 Flash. Fuente: Google DeepMind. 19 may 2026. Título original de la fuente: Gemini 3.5 Flash Model Card. https://deepmind.google/models/model-cards/gemini-3-5-flash/

9. Documentación del modelo Llama. Fuente: Meta. 9 ago 2026. Título original de la fuente: Llama model documentation. https://ai.meta.com/llama/get-started/

10. Ficha del modelo NVIDIA Nemotron 3.5 Nano. Fuente: NVIDIA. 15 dic 2025. Título original de la fuente: NVIDIA Nemotron 3.5 Nano model card. https://build.nvidia.com/nvidia/nemotron-3.5-nano-30b-a3b/modelcard