AI Safety Landscape

A documented view of public safety frameworks, model-level disclosures and institutional approaches to advanced AI risk.

Published: Aug 9, 2026 · Evidence date: Aug 9, 2026

Publisher: AI Governance World LLC · aigovernanceworld.com

Executive summary

The launch landscape combines public risk and management frameworks with model cards, system cards, developer policies and verified incident records. These evidence layers answer di f erent questions. A public framework describes an approach to risk management; a developer disclosure states what its publisher reports; an independent evaluation tests a model under a documented method; and an enforcement record states what an authority found or required.

The corpus currently contains developer-authored model documentation but no reviewed independent model evaluation linked to the launch model records. That absence is displayed explicitly and is not filled with a proxy rating. No model or company receives an aggregate safety score from first-party documentation alone.

Methodology

Eligible evidence includes o f icial Tier 1 frameworks and enforcement records, academic Tier 2 evaluations with attributable methods, and developer-authored Tier 3 model cards, system cards and safety reports. Media may provide discovery context but is never primary evidence.

Developer claims are identified through their publisher and source type and remain attached to the model or company record as first-party disclosure. Independent evaluations are stored in the model-evaluation table with evaluator, evaluation date, methodology, results, findings and sources. The two evidence classes are reported separately and are never merged into an unsupported safety rating.

Public safety and risk frameworks

Frameworks are compared by lifecycle coverage, governance, risk identification, measurement, treatment, monitoring, incident handling and continual improvement. NIST AI RMF and ISO/IEC 42001 serve di f erent functions: the corpus records their o f icial scope and conformity-assessment characteristics rather than treating them as interchangeable certifications.

OECD and UNESCO instruments add normative and policy context. Their presence in the corpus does not establish adoption or implementation by a particular company or country without separate evidence.

Developer model and system disclosures

Model records preserve disclosed capabilities, evaluation approaches, limitations, mitigations, access conditions and update dates without filling undocumented fields. The publisher is shown so readers can identify the evidence as first-party.

A developer statement that testing occurred is a verifiable disclosure claim. It is not, by itself, independent confirmation of the result, the completeness of the test set or the e f ectiveness of mitigations.

Assurance limits

A published framework is evidence of a governance commitment, not proof of implementation or model safety. A model card is evidence of disclosure, not certification. An independent evaluation is bounded by its method and tested version, while a regulator finding is bounded by its jurisdiction and legal action.

For these reasons the Safety Landscape publishes evidence counts and provenance categories, not an aggregate safety league table.

Four distinct evidence layers

The report separates public frameworks, developer disclosures, independent evaluations and incidents or enforcement. A record can contribute to more than one analytical question, but its provenance does not change. For example, a system card remains a developer claim even when it describes extensive testing.
This structure prevents disclosure volume from being mistaken for independently demonstrated safety and prevents an isolated incident from being generalized into an unsupported model-wide conclusion.

Independent evaluation register

An independent evaluation requires an evaluator distinct from the model developer, an evaluation date, a documented methodology, attributable findings and an eligible source. Records that do not identify these elements remain outside the reviewed independent-evaluation count.
The launch chart reports the current count directly from the model-evaluation registry. Zero means that no linked record has completed verification; it does not mean that no external research exists.

Incidents, response and enforcement

Incident records preserve the event date, severity, status, impact summary, response and primary source. Enforcement actions remain attributable to the responsible authority and are linked through the timeline. These records may inform later evaluation, but the platform does not infer technical causation or model-wide risk beyond the official evidence.

Charts and data

Safety evidence inventory

Frameworks, model disclosures and incidents are different record classes.

Public framework records: 4
Model disclosure records: 5
Confirmed incident records: 1
The corpus contains 4 public framework records, 5 model disclosure records and 1 confirmed incident record.

Developer disclosure and reviewed evaluation

Developer-authored documentation is kept separate from verified model evaluations.

Developer-authored model sources: 5
Reviewed first-party evaluations: 0
Reviewed independent evaluations: 0
5 model records use developer-authored primary sources; 0 first-party and 0 independent model evaluations have completed review.

Documented model access modes

Access mode is reported from the verified model record and is not a safety rating.

API: 3
Open weights: 2

Verified model records grouped by documented access mode.

References

1. Artificial Intelligence Risk Management Framework (AI RMF 1.0). Source: National Institute of Standards and Technology. Jan 26, 2023. https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-ai-rmf-10

2. ISO/IEC 42001:2023 — Artificial intelligence management system. Source: International Organization for Standardization. Dec 1, 2023. https://www.iso.org/standard/42001.html

3. OECD AI Principles. Source: Organisation for Economic Co-operation and Development. May 3, 2024. https://oecd.ai/en/ai-principles

4. Recommendation on the Ethics of Artificial Intelligence. Source: UNESCO. Nov 23, 2021. https://www.unesco.org/en/legal-affairs/recommendation-ethics-artificial-intelligence

5. NHTSA Announces Consent Order with Cruise After Company Failed to Fully Report Crash Involving Pedestrian. Source: National Highway Traffic Safety Administration. Sep 30, 2024. https://www.nhtsa.gov/press-releases/consent-order-cruise-crash-reporting

6. GPT-5.6 System Card. Source: OpenAI. Jul 9, 2026. https://deploymentsafety.openai.com/

7. Anthropic model system cards. Source: Anthropic. Jun 1, 2026. https://www.anthropic.com/system-cards

8. Gemini 3.5 Flash Model Card. Source: Google DeepMind. May 19, 2026. https://deepmind.google/models/model-cards/gemini-3-5-flash/

9. Llama model documentation. Source: Meta. Aug 9, 2026. https://ai.meta.com/llama/get-started/

10. NVIDIA Nemotron 3.5 Nano model card. Source: NVIDIA. Dec 15, 2025. https://build.nvidia.com/nvidia/nemotron-3.5-nano-30b-a3b/modelcard