

AI SAFETY MONITORING

COMPETITIVE LANDSCAPE REPORT

October 2025

EXECUTIVE SUMMARY

The AI safety monitoring market has exploded from \$890M in 2024 to a projected \$5.8-6.6B by 2029-2034, driven by catastrophic AI failures (\$67.4B in hallucination losses alone in 2024) and mounting regulatory pressure including €35M fines under the EU AI Act. RAIDS enters this validated, high-growth market with a differentiated positioning at the intersection of real-time behavioral detection and business risk mitigation.

Market Validation

Three major acquisitions in 2025 alone validate this market:

- Lakera sold to Check Point for \$300M (10x return on \$30M raised)
- CalypsoAI to F5 for \$180M (4.2x return on \$43M raised)
- WhyLabs acquired by Apple

These exits signal that established cybersecurity giants and tech leaders see AI safety as mission-critical infrastructure requiring immediate integration.

The Opportunity Gap

While 72% of companies deploy AI, only 9% feel prepared for associated risks. The competitive landscape remains fragmented across three overlapping segments:

- **AI Observability:** Arize AI (\$176M raised, \$2.5B valuation), Fiddler (\$68M), Evidently; focused on model performance metrics for data scientists
- **AI Security:** Lakera (now Check Point), Protect AI (\$60M), HiddenLayer (\$58M); preventing attacks on known AI systems
- **AI Governance:** Credo AI, ModelOp (\$51M), Holistic AI; ensuring compliance through policy frameworks

No clear leader exists in real-time rogue AI detection - the intersection of behavioral monitoring, business risk focus, and unauthorized deployment identification. This is RAIDS' blue ocean opportunity.

RAIDS' Differentiation

RAIDS distinguishes itself through:

- **Business-focused positioning:** Targeting CEOs, CFOs, and business leaders rather than data scientists
- **Real-time intervention:** Active protection during operation v. passive post-hoc analysis
- **Rogue AI detection:** Identifying unauthorized agents, compromised systems, and malicious deployments
- **Vendor-neutral architecture:** Works across AWS, Azure, GCP, and all major LLM providers

MARKET OVERVIEW

Three Converging Forces

1. Regulatory Tsunami

The EU AI Act entered enforcement August 2024 with fines up to €35M or 7% of global revenue. The U.S. added 59 federal AI regulations in 2024 alone (up from 29 in 2023), with 500+ state-level AI bills emerging in Q1 2025. This fragmented regulatory landscape creates complexity that favors platforms offering automated compliance.

2. Security Incidents Skyrocketing

The AI Incident Database recorded 233 incidents in 2024, a 56.4% jump from 2023. Key incidents driving enterprise awareness:

- ChatGPT fabricating a Norwegian man as a child murderer
- Multiple lawyers sanctioned for submitting AI-hallucinated case citations
- Air Canada's chatbot honoring fabricated bereavement policies
- Australian government paying \$440K for Deloitte report citing fabricated sources

Total hallucination-driven losses in 2024: \$67.4B

3. Enterprise AI Adoption Outpacing Governance

72% of companies use AI but only 9% feel prepared. Shadow AI proliferation reveals enterprises discover 31 AI tools in use versus 5 approved (6x gap), with 59.9% of AI transactions blocked by enterprise security. 95% of corporate GenAI projects fail due to inadequate governance.

Market Size and Growth

Market Projections:

- 2024: \$890M (current market)
- 2029-2034: \$5.8-6.6B projected
- CAGR: 45-49% (explosive growth)

Geographic Distribution:

- North America: 31-40% market share (U.S. dominates at 86%)
- Europe: 28% share with strongest regulatory drivers
- Asia-Pacific: Fastest regional growth at 36-42% CAGR

Industry Adoption:

- Financial Services: 34.6% segment share, 78% implementing generative AI
- Healthcare: 33.2% CAGR (fastest-growing), 570 FDA-approved AI-enabled medical devices
- IT and Telecommunications: 27-30% market share

MARKET SEGMENTS AND CATEGORIES

The AI safety monitoring market has evolved into a complex ecosystem of overlapping capabilities and specialized focus areas. Unlike traditional software categories with clear boundaries, AI safety tools operate across a spectrum from development-focused technical monitoring to deployment-focused business protection.

The market divides into four primary segments, each with distinct buyer personas, use cases, and competitive dynamics. RAIDS' positioning at the intersection of all four segments, plus pioneering a new category (Rogue AI Detection), creates a unique strategic advantage.

Segment 1: AI Observability Platforms

Focus: Model performance monitoring, drift detection, data quality, and technical metrics for ML engineers and data scientists.

Buyer Persona: ML engineers, data scientists, MLOps teams. Purchase authority typically sits in Engineering or Data Science departments.

Primary Use Cases: Detecting model drift, monitoring prediction accuracy, tracking data quality issues, debugging model performance degradation, A/B testing different model versions.

Arize AI

Funding: \$176M raised, \$2.5B valuation

Strengths: Market-leading observability platform with 2M+ monthly Phoenix downloads. Strong technical capabilities in drift detection, data quality monitoring, and explainability.

Limitations v. RAIDS: Focused on technical metrics rather than business impact. Passive analysis rather than active threat detection. No real-time intervention capabilities. Developer-focused rather than business-buyer-focused.

Fiddler AI

Funding: \$68M raised

Strengths: Pioneer in explainability and model transparency. Strong academic credentials and research backing. Good enterprise traction in regulated industries.

Limitations v. RAIDS: Explainability focus addresses 'why did the model do this' rather than 'should we allow this behavior.' No rogue AI detection or unauthorized deployment identification.

Evidently AI

Funding: \$1.15M raised

Strengths: Open-source first approach with strong developer community. Lightweight and easy to implement.

Limitations v. RAIDS: Limited funding restricts enterprise features. No real-time intervention. Self-service model not suitable for enterprise buyers seeking managed solutions.

Coralogix

Funding: \$350M raised, \$1B+ valuation

Strengths: Launched AI Center in March 2025. Strong infrastructure for log aggregation and analysis.

Limitations v. RAIDS: Late entrant to AI monitoring (2025 launch). Log-based observability rather than behavioral detection. No specialized rogue AI capabilities.

RAIDS Differentiation from Observability Platforms

While observability platforms monitor model performance metrics, RAIDS monitors for rogue behavior and unauthorized deployments. Observability answers 'how is my model performing?' RAIDS answers 'is my AI system exhibiting dangerous behavior or unauthorized activity?' A model can have excellent technical metrics (98% accuracy, low latency, no data drift) while simultaneously exhibiting dangerous rogue behavior. Observability tools would show green lights while RAIDS would trigger critical alerts.

Segment 2: AI Security Tools

Focus: Protecting AI systems from adversarial attacks, prompt injection, jailbreaks, data poisoning, and model theft.

Buyer Persona: Security teams, CISOs, application security engineers. Purchase authority typically sits in Security or InfoSec departments.

Primary Use Cases: Preventing prompt injection attacks, detecting jailbreak attempts, protecting against model extraction, securing API endpoints, filtering malicious inputs.

Lakera Guard (acquired by Check Point)

Exit: \$300M acquisition (10x return on \$30M raised)

Strengths: Leading prompt injection prevention. Strong technical execution on LLM security.

Limitations v. RAIDS: Filters malicious inputs to known AI systems. Does not detect unauthorized AI deployments or rogue behavior from legitimate systems. Now part of Check Point ecosystem.

Protect AI

Funding: \$60M raised

Strengths: Full lifecycle AI security covering training, deployment, and operations. Strong CISO-level positioning.

Limitations v. RAIDS: Securing known AI systems rather than detecting unknown rogue AI. Attack prevention focus rather than behavioral anomaly detection.

HiddenLayer

Funding: \$58M raised

Strengths: Model security specialists. Strong technical team from security research backgrounds.

Limitations v. RAIDS: Model-level security rather than behavioral monitoring. Requires access to model internals. Cannot monitor black-box third-party AI systems.

RAIDS Differentiation from Security Tools

Security tools protect known AI systems from external attacks. RAIDS detects internal threats from the AI itself, including rogue behavior, unauthorized deployments, and dangerous decision-making patterns. Security tools answer 'are bad actors attacking my AI?' RAIDS answers 'is my AI acting badly?' Both are necessary but address fundamentally different threat models.

Segment 3: AI Governance Platforms

Focus: Policy enforcement, compliance documentation, audit trails, risk assessment frameworks, and regulatory alignment.

Buyer Persona: Compliance officers, legal teams, risk managers, CROs. Purchase authority typically sits in Legal, Compliance, or Risk Management departments.

Primary Use Cases: EU AI Act compliance, model inventory management, risk assessment documentation, policy creation and enforcement, audit preparation.

Credo AI

Funding: \$21M raised

Strengths: Comprehensive governance framework. Strong regulatory expertise. Good enterprise positioning.

Limitations v. RAIDS: Policy documentation rather than real-time protection. Audit trail creation rather than active threat detection.

ModelOp

Funding: \$51M raised

Strengths: ModelOps platform with governance features. Enterprise sales motion. Integration with ML platforms.

Limitations v. RAIDS: Model management platform with governance add-ons rather than purpose-built safety system. No rogue AI detection.

Holistic AI

Funding: \$10M+ raised

Strengths: EU AI Act specialist. Strong regulatory expertise. Good European market positioning.

Limitations v. RAIDS: Compliance assessment tool rather than operational monitoring system. Manual review processes rather than automated detection.

Arthur AI

Funding: \$42M raised

Strengths: Model monitoring with governance overlay. Good technical capabilities. Enterprise customer base.

Limitations v. RAIDS: Performance monitoring extended into governance rather than purpose-built safety platform. No rogue AI detection capabilities.

RAIDS Differentiation from Governance Platforms

Governance platforms create documentation and policies. RAIDS enforces safety in production. Governance platforms answer 'what policies should we have?' RAIDS answers 'are those policies being violated right now?' The relationship is complementary: governance platforms define acceptable behavior, RAIDS monitors for violations.

Segment 4: Rogue AI Detection (RAIDS' Blue Ocean)

Focus: Identifying and mitigating unauthorized AI deployments, detecting behavioral anomalies in production systems, and preventing AI-driven business risks before they materialize.

Buyer Persona: CEOs, CFOs, CROs, board members; business leaders concerned with AI-driven business risk rather than technical metrics.

Primary Use Cases: Detecting shadow AI deployments (the 6x gap between 5 approved and 31 discovered tools), identifying when approved AI systems exhibit dangerous behavior, preventing AI hallucinations from causing business harm, alerting on anomalous AI decision patterns, monitoring third-party AI tools for rogue behavior.

Why This is a Blue Ocean Opportunity

While the three established segments above have clear leaders and intense competition, rogue AI detection remains dramatically underserved. The market gap is validated by:

- **The shadow AI epidemic:** Enterprises discover 31 AI tools in use versus 5 approved (6x gap). None of the established segments address unauthorized deployment detection.
- **Business risk focus:** Existing tools serve technical teams. No solution addresses C-suite concerns about AI-driven business disasters.
- **Black-box monitoring:** RAIDS works without model access, enabling monitoring of third-party AI tools and SaaS AI features that organizations cannot inspect internally.
- **Real-time intervention:** Existing segments provide post-hoc analysis or passive logging. RAIDS detects and alerts on dangerous behavior as it happens.

RAIDS as Category Pioneer

RAIDS defines the rogue AI detection category through three unique capabilities that no competitor combines:

1. Behavioral Baseline Monitoring: RAIDS establishes normal behavior patterns for AI systems and detects deviations that indicate rogue activity; whether from unauthorized systems, compromised approved systems, or legitimate systems exceeding their intended scope.

2. Shadow AI Discovery: RAIDS identifies unauthorized AI deployments across the enterprise by monitoring network traffic, API calls, and system behaviors that indicate AI usage not in the approved inventory.

3. Business Risk Translation: Rather than technical metrics, RAIDS translates AI behavior into business impact; 'this hallucination could cost \$X in customer harm' rather than 'model confidence dropped 3%.'

As the category pioneer, RAIDS gains first-mover advantages in defining market education, establishing best practices, and shaping buyer expectations.

COMPETITIVE LANDSCAPE ANALYSIS

Using a Gartner-style Magic Quadrant methodology, we categorize AI safety and monitoring solutions based on their execution capability (ability to deliver solutions today) and strategic vision (understanding of market direction and innovation). This analysis evaluates 24 companies across the AI safety landscape.

Key Finding: RAIDS emerges as the only pure-play rogue AI detection platform in the Leaders quadrant, while NeMo and Bedrock represent platforms with safety features as components rather than core focus.

Leaders Quadrant

RAIDS (Rogue AI Detection System)

Execution Score: 8.2 | Vision Score: 8.4

Strategic Positioning: The only dedicated rogue AI behavior detection platform with real-time anomaly detection, interpretability, and proactive mitigation. Purpose-built for identifying unauthorized AI deployments, behavioral deviations in approved systems, and AI-driven business risks.

Core Capabilities:

- Patent-pending dual autoencoder architecture achieving 93% rogue behavior detection with 7% false positive rate
- Real-time behavioral baseline monitoring with sub-100ms detection latency
- Black-box monitoring requiring no model access; works with any AI implementation
- Shadow AI discovery identifying unauthorized deployments across enterprise
- Business-focused alerting translating technical anomalies to business impact

Target Market: CEOs, CFOs, CROs, and business leaders in regulated industries (financial services, healthcare, legal) and AI-dependent enterprises.

Why RAIDS Leads: RAIDS defines and owns the rogue AI detection category. While competitors monitor AI performance or secure AI systems, RAIDS uniquely identifies when AI goes rogue; whether through unauthorized deployment, behavioral deviation, or dangerous decision patterns.

NVIDIA NeMo

Execution Score: 8.0 | Vision Score: 8.0

Strategic Positioning: AI development and model management framework with built-in governance and compliance tools. Part of NVIDIA's broader AI infrastructure play.

Relationship to RAIDS: Complementary rather than competitive. NeMo helps build AI systems, RAIDS monitors them in production. Organizations using NeMo for development would still need RAIDS for real-time behavioral monitoring.

Amazon Bedrock

Execution Score: 8.4 | Vision Score: 8.0

Strategic Positioning: Fully managed service for foundation models with built-in safety guardrails. Part of AWS's AI services portfolio.

Relationship to RAIDS: Complementary; Bedrock provides AI infrastructure with basic safety rails, RAIDS provides behavioral monitoring across all AI systems. Organizations using Bedrock would add RAIDS for real-time rogue behavior detection beyond Bedrock's environment.

Challengers Quadrant

Challengers demonstrate strong execution but more limited strategic vision. Key players include Coralogix (7.4), Lakera Guard (7.0, now Check Point), Lasso Security (6.8), Holistic AI (7.0), Deepchecks (7.0), and ModelOp (7.0).

Visionaries Quadrant

Visionaries demonstrate strong strategic vision but currently have lower execution capability. Key players include WhyLabs (7.6, now Apple), Arize AI (7.0), Evidently AI (6.8), Credo AI (7.2), and AI21 Labs (7.2).

Niche Players Quadrant

Niche Players have focused market scope. Key players include ManageEngine (6.6), SentinelOne (6.0), Intenseye (5.0), Digital Watchdog (4.6), Surveily (4.2), and Chooch (4.2).

STRATEGIC POSITIONING RECOMMENDATIONS

Based on this competitive analysis, the following positioning strategies should be emphasized in media, events, podcasts, and PR efforts:

1. Own the Rogue AI Detection Category

Position RAIDS as the category pioneer and leader for rogue AI detection, distinct from crowded observability, security, and governance segments. Define the category around three threat types:

- **Malicious Rogue AI:** Unauthorized deployments and shadow AI (addressing the 31 discovered v. 5 approved tools gap)
- **Accidental Rogue AI:** Approved systems exceeding their intended scope or exhibiting dangerous behavior patterns
- **Compromised Rogue AI:** Legitimate systems subverted through attacks or manipulation

Category leadership creates pricing power, media attention, and first-mover advantage in defining market education.

2. Target Business Buyers Over Technical Buyers

Frame capabilities as business outcomes rather than technical metrics. Target CFOs, CROs, and business leaders with business impact positioning:

- Preventing financial losses (\$67.4B hallucination losses in 2024)
- Detecting unauthorized AI before compliance violations (€35M EU AI Act fines)
- Protecting reputation (233 AI incidents in 2024, 56.4% increase)

Business buyers have larger budgets, shorter evaluation cycles for clear risk mitigation, and influence over enterprise-wide adoption.

3. Emphasize Independence as Strategic Advantage

With Lakera now Check Point, WhyLabs now Apple, and CalypsoAI now F5, position RAIDS as the independent third-party validation that enterprises require.

Key Message: Would you trust a cloud provider to monitor AI systems running on their own infrastructure? Independence matters for objective AI safety monitoring.

4. Lead with Regulatory and Business Risk

Open media conversations with concrete business risks before discussing technical capabilities:

- €35M EU AI Act fines
- \$67.4B hallucination losses in 2024
- 233 AI incidents (56.4% increase)
- 500+ state AI bills in Q1 2025 alone

These headline figures create C-suite awareness and urgency.

5. Clarify Complementary v. Competitive Positioning

Most 'competitors' are actually complementary. Use this positioning to reduce perceived competition:

- **Observability platforms:** 'They monitor model performance, we monitor for rogue behavior; complementary capabilities'
- **Security tools:** 'They secure your approved AI systems, we detect unauthorized ones; both needed'
- **Governance platforms:** 'They define policies, we enforce them in real-time; complementary layers'

This framing positions RAIDS as an essential addition to existing tools rather than a replacement, accelerating adoption.

6. Highlight Real Customer Pain Points

Frame RAIDS around specific customer stories and pain points rather than abstract capabilities:

- 'Financial services firm discovers 31 AI tools in production, only 5 were approved; RAIDS identified all 26 unauthorized deployments'
- 'Healthcare provider's AI system began recommending off-label drug uses; RAIDS detected behavioral deviation before patient harm'
- 'Legal tech company avoided \$500K+ sanctions by detecting AI hallucinations in case citations before filing'

Real scenarios (even anonymized) resonate more than technical specifications.