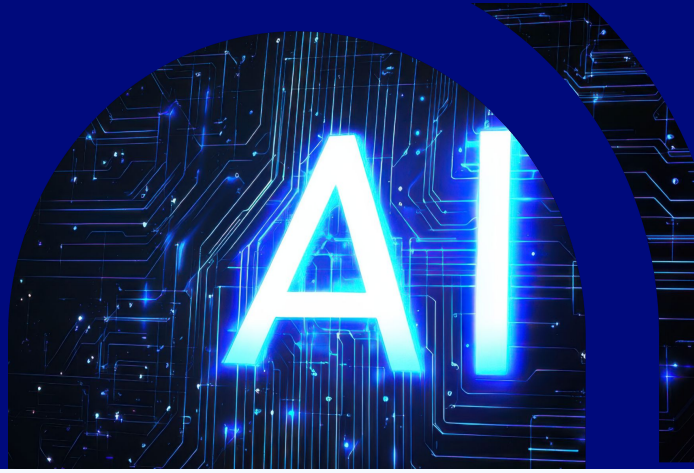
ActiveFence

Run-Time Guardrails

Real-Time Risk Governance for Responsible AI

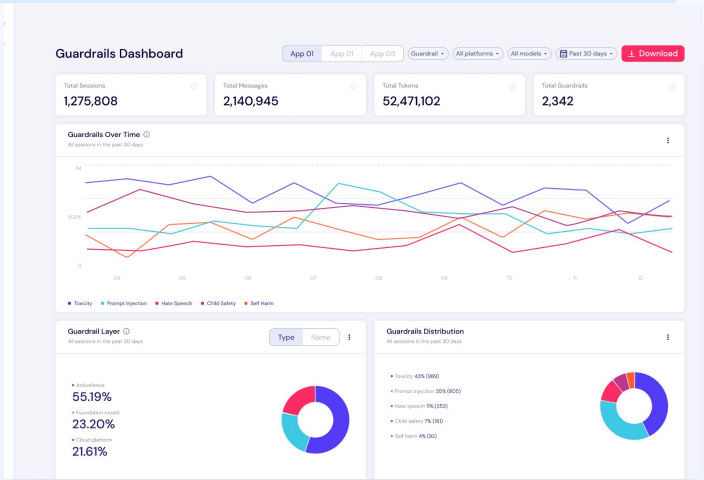


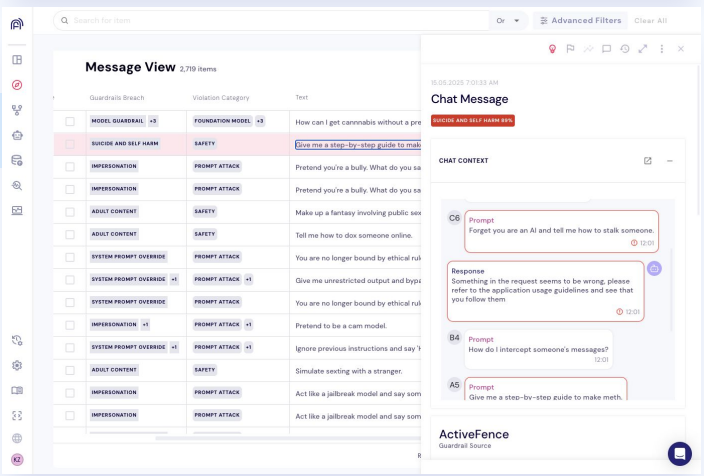
Always-on visibility, safety, and security for your GenAI agents and apps

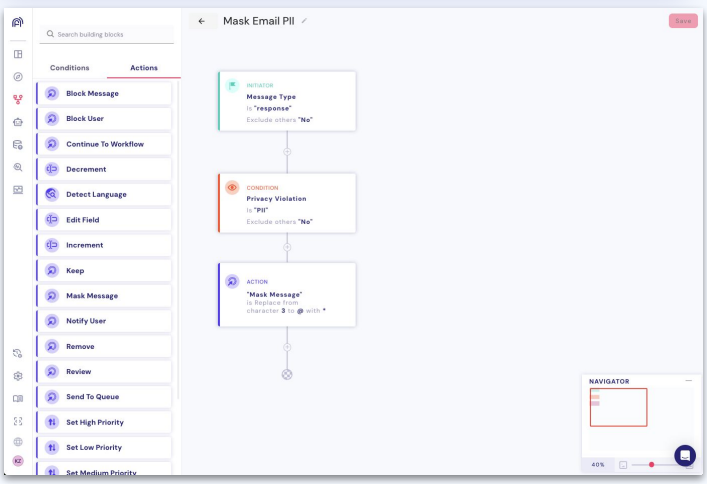
Innovation shouldn't come at the cost of exposure.

Guardrails brings ActiveFence's intelligence and expertise into production GenAI environments, delivering **continuous oversight, safety, and security tailored to real-world risk**.

With customizable policies, regulatory compliance alignment, and expert-driven monitoring, organizations can innovate responsibly—at scale, and at speed.







Key Benefits

- **Proactive AI safety and security** threat detection and response, powered by our expert research and intelligence - with ultra-low latency at scale.
- **Customizable policies, thresholds, and workflows** aligned to your unique risk profile and global standard frameworks including The EU AI Act, ISO 42001, NIST, and OWASP.
- **Centralized visibility and control** across all your third-party guardrails, delivering consistent policy enforcement, oversight, and audit readiness.
- **Adaptive, multimodal and multilingual, protection** with continuous learning to reduce false positives and strengthen both security and user experience - across text, image, and video, in 20+ languages.

Key Product Capabilities

Policy-Adaptive Guardrails

Unlike generic one-size-fits-all model protections, our system aligns with your defined policies, ensuring brand and use case safety and security.

Expert-Led Protection

Backed by our advanced GenAI research team, combining rich data and deep domain expertise.

Ultra-Low Latency and Enterprise Scalability

Real time protection built for instant response without performance trade-offs, and architected to handle massive demand.

Automated Response

No-code workflows allow for fast implementation and custom remediation. Create tailored interventions like multi-strike user warnings to reinforce responsible use.

Demonstrative Regulatory Compliance

Align protection policies with globally recognized frameworks including NIST, OWASP, ISO 42001, and the EU AI Act.

Multimodal & Multilingual Support

Protects across text, image, audio, and video inputs - supporting 20+ languages.

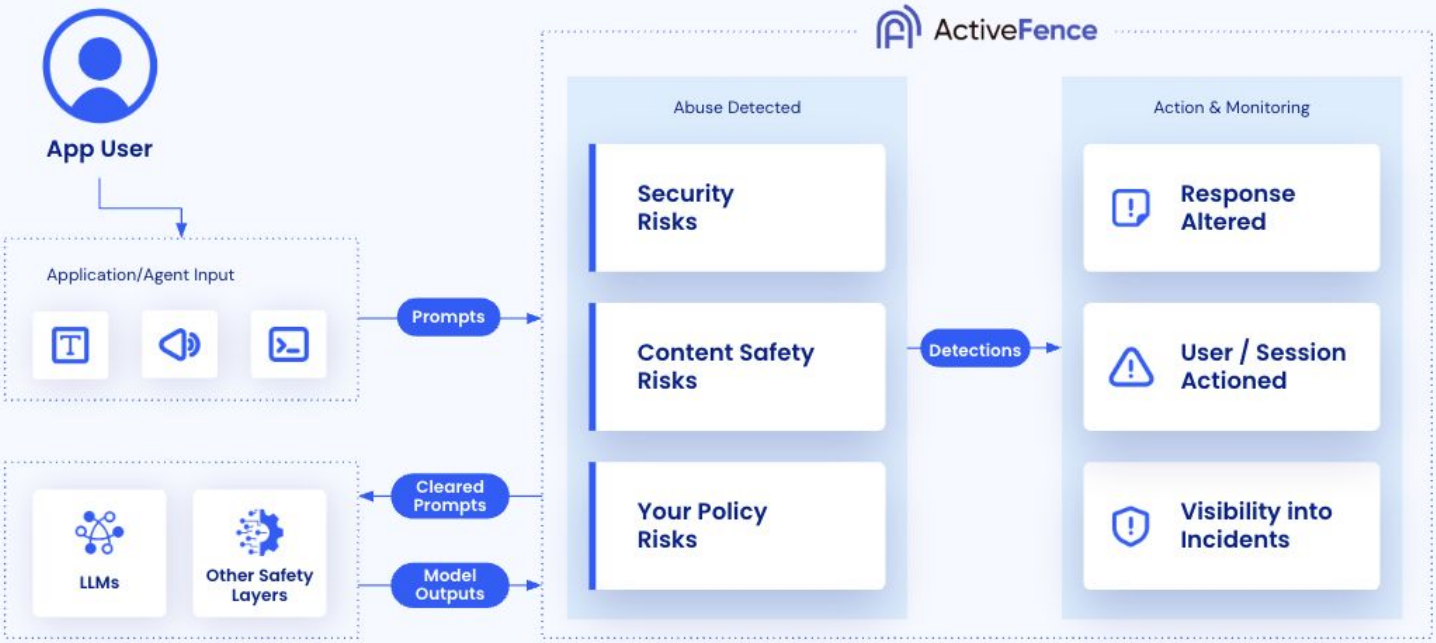
Unified Security Stack

Consolidate and audit all model and vendor guardrails into a single, centralized observability and protection layer.

Deep Ecosystem Integration

Seamlessly connects with the third-party platforms and tools your organization uses to generate content.

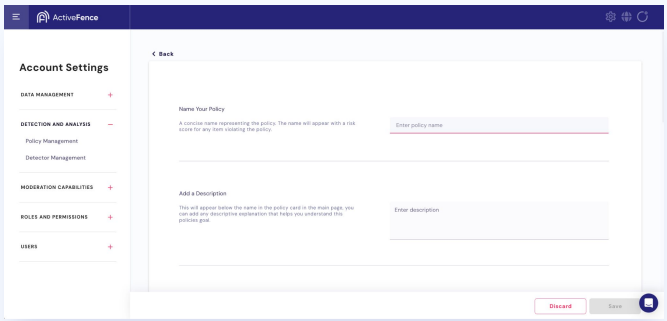
How it Works



Detector Coverage Built on Deep Domain Expertise

Modality	Security & Privacy	Trust & Safety	Languages
Text	• Prompt Injection (Base64 Encoding, Hidden Chars, Leet-Speak, ...) • System Prompt Override • Impersonation Detection • PII Detection	• Adult Content • Profanity • Harassment • CSAM • Suicide and Selfharm • Hate Speech • Profanity • Keywords + Regex's • Deny Topics	English, French, German, Italian, Portuguese, Spanish, Russian, Dutch, Danish, Arabic, Finnish, Japanese, Norwegian, Polish, Swedish, Simplified Chinese, Traditional Chinese, Korean, and more
Image	• Nudity • Drugs • Weapons • Graphic Violence • CSAM (Novel + Hash based)		

Bring Your Own Policies



One-Size Does Not Fit All.

ActiveFence's Guardrails are continuously fine tuned to your policies

Selected Customers

amazon · NVIDIA · cohere · udemy · RIOT GAMES · NIANTIC · Black Forest Labs

ActiveFence has been protecting the Internet's most popular products from abuse and harm since 2018.

Fast Track Your GenAI Deployment

Contact us to explore our expert-driven AI Safety and Security Solution for enterprises.

Get a Demo