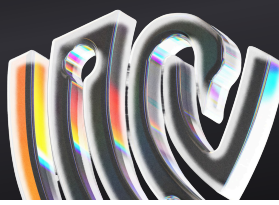
GRAY SWAN

Battle-tested AI Security for Enterprise AI



The Threat Model Has Changed

Your existing tools weren't built for this. The question use to be, "Who can get in?" Now it's "What can the model do?" AI agents introduce failure modes that your current stack wasn't designed for:

Behavioral Failures	It's not just attacks we have to worry about: agents can behave dangerously and create breaches through benign interactions with users or autonomously.
Dynamic System-Level Risk	Risk is determined by what the model-powered agents can do, read, write, and trigger. Models can adapt and evolve unpredictably in real time.
Threat Intel Velocity & Adaptability	AI attack techniques evolve faster than vulnerability management cycles can keep up with.

Gray Swan was built to address these challenges

Gray Swan was founded by the Carnegie Mellon researchers who discovered how LLMs get attacked and built the methods to defend them.

Their insight: AI systems must be secured using **purpose-built AI models trained on high-quality data that can evolve, anticipate, and counter** known and unknown threats. And that's what they did.

Research powers everything we do. It's the backbone of our platform and continuously strengthens our products. This is why leading AI labs came to us first and rely on us before shipping models. It's also why enterprises trust us to protect what they've deployed.

Trusted by AI makers to protect AI

OpenAI

Model cards:
o1, o3 mini, GPT 5

ANTHROP\C

Model cards:
Sonnet 4.6 & 4.5, Opus 4.7, 4.6 & 4.5, Haiku 4.5, Mythos Preview

Meta

Model card:
Muse Spark

Google DeepMind

amazon

AISI | AI SECURITY INSTITUTE

ByteDance

xAI

Trusted by enterprises to secure AI

Deloitte.

METR

Artificial Intelligence Underwriting Company

Our Solutions

Shade

AUTOMATED ADVERSARIAL TESTING

Find vulnerabilities before attackers do.

Shade is an AI attack agent that plugs into your CI/CD pipeline, running adversarial tests against AI models and agents. One-time assessments go stale the moment you ship a new version. Shade runs the actual attack strategies being used in the wild right now, adapted dynamically to your system.

Cygnal

RUNTIME MONITORING AND PROTECTION

Stop what guardrails miss.

Cygnal understands your agents' purpose, and enforces that in real time.

Standard guardrails are static, deterministic rules trying to regulate non-deterministic systems. Cygnal is powered by an adaptive AI model trained on Arena data to identify all suspicious out-of-policy actions.

Configured to your environment, violations are blocked, logged, timestamped, and exportable. Continuous threat intel from Arena keeps your defenses ahead of emerging threats.

Arena

WORLD'S LARGEST AI RED TEAMING NETWORK

Red-teaming battlefield for breaking AI models.

The Gray Swan Arena is the world's largest AI red teaming network with over 15,000 red teamers breaking frontier models, finding novel attack vectors, and surfacing dangerous vulnerabilities.

This uniquely comprehensive data powers Gray Swan's AI-based solutions, Shade and Cygnal.

A continuously improving virtuous cycle

- Arena data to train Shade and Cygnal
- Shade and Cygnal adversarially test each other
- Arena participants test Cygnal

Can Your AI Withstand the Pressure?

Contact Us
sales@grayswan.ai

Schedule a Demo

GRAY SWAN