ØDIN DEFENSE

ØDIN Defense Datasheet

Powered by Mozilla

Block Jailbreaks in Real Time

A fast, self-hostable classifier model that scores every prompt for adversarial intent, enabling your system to block, log or take action before it reaches your LLM.

Why ØDIN Defense?

Built for teams shipping GenAI to production

As LLMs move into customer-facing apps and agentic workflows, every prompt becomes untrusted input. ØDIN Defense scores that input for attack risk and blocks jailbreaks inline, with no new infrastructure and no data leaving your environment.

What You Get:

- → Real-time inline blocking of jailbreak attacks
- → Local ONNX inference, no prompts leave your environment
- → SusFactor Ø-1 suspicion score on every prompt
- → Runs CPU-only, small enough for commodity hardware
- → Trained on real bug-bounty attacks, retrained as threats evolve

*JailbreakBench: 87.9% of attacks blocked at a 9.0% benign false-positive rate; tune per path.

Key Benefits

Catch Attacks Early

SusFactor scores every prompt from Ø (benign) to 1 (suspicious) with a fine-tuned encoder, so you can gate, log, or reject requests at the application boundary.

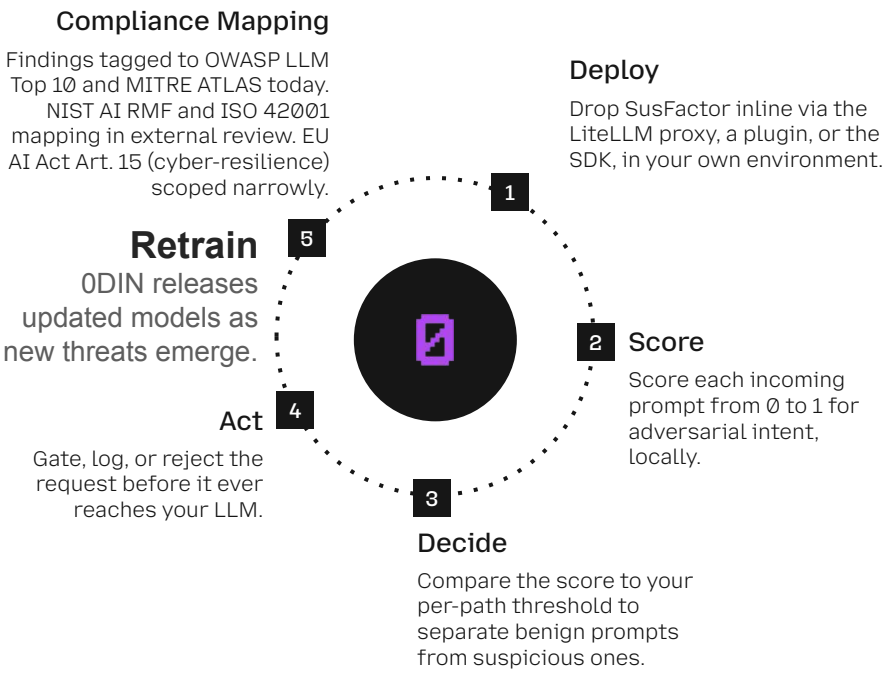
Keep Sensitive Prompts In-House

Detection runs entirely in-process on bundled ONNX models, with no network calls. Prompts are inspected inside your environment and never sent to ØDIN.

Trained on Real Attacker Data

SusFactor learns from ØDIN's bug-bounty corpus: thousands of real jailbreaks disclosed by 2,000+ researchers, so detection reflects how models are actually attacked.

Detection Lifecycle



Key Benefits

Tune Sensitivity to Your Risk

One continuous score per team. Set it strict for critical paths, looser for customer-facing flows.

Deploy Anywhere, Fast

Small enough to run CPU-only and fast enough to sit inline, via the LiteLLM proxy, AWS Bedrock, or the SDK.

Defense-in-Depth, By Design

Jailbreak-focused, with partial prompt-injection coverage. Scores prompts, not retrieved documents. Not a sole control.

As GenAI moves into production, traditional input validation can't keep pace with adversarial prompts that exploit model behavior rather than code. Every prompt your application receives is untrusted input.

ØDIN Defense scores each prompt for attack likelihood in-process, on your own hardware, then blocks jailbreak attacks before they reach the model, without exposing prompt content.

Core Capabilities

SusFactor Classifier

Local 0–1 suspicion scoring for adversarial intent (~70–240ms per prompt, CPU-only).

Benchmarked Detection

JailbreakBench: 12.1% attack success at 9.0% false positives, lowest FPR among effective detectors.

Real-World Threat Data

Trained on nearly 69,000 examples, including 3,426 real attacks from the ØDIN Threat Feed.

Deploy Your Way

Self-hosted download or hosted early-access trial.

ØDIN

Secure People, Secure World.

We unite security experts to safeguard humanity.

HTTPS://ØDIN.AI

ØDIN@MOZILLA.COM