ØDIN PROMPT TOOLKIT

Prompt Toolkit Datasheet

Powered by Mozilla

Build AI Control Validation Into Your Stack

A multi-language SDK that detects unsafe model inputs and matches them against ØDIN's researcher-validated exploit intelligence, running entirely in your environment so no prompts ever leave it.

Why ØDIN Prompt Toolkit?

Built for teams shipping GenAI to production

As LLMs move into customer-facing apps and agentic workflows, every prompt becomes untrusted input. The Prompt Toolkit lets developers score that input for risk and check it against known exploits, locally, with no new infrastructure and no data leaving your environment.

What You Get:

- → Drop-in libraries for Rust, Python, and TypeScript
- → Local ONNX inference — no prompts leave your environment
- → SusFactor risk scoring for unsafe inputs
- → LSH signatures for privacy-preserving similarity matching
- → Optional ØDIN Exploit Intelligence feed matching

*Distributed as a developer SDK; package-registry releases rolling out.

Key Benefits

Catch Unsafe Inputs Early

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

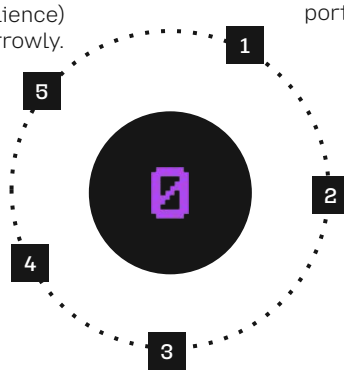
Keep Sensitive Prompts In-House

Detection runs locally on bundled ONNX models; matching compares 256-bit fingerprints, not raw text, so prompt data never leaves your environment.

Match Against Real-World Exploits

Fingerprint any prompt and match it against ØDIN's exploit intelligence: control-failure and untrusted-input patterns from our research and bug-bounty community. Every match carries severity.

Detection Lifecycle



Sync

Fetch the latest exploit signatures from the ØDIN portal and cache them locally.

Sign

Generate a 256-bit LSH signature and SusFactor score for each incoming prompt — locally.

Query

Find similar cached signatures via band-indexed lookup, then confirm with Hamming distance.

Act

Gate, log, or route using the score, plus each match's severity and security boundary.

Compliance Mapping

Findings tagged to OWASP LLM Top 10 and MITRE ATLAS today. NIST AI RMF and ISO 42001 mapping in external review. EU AI Act Art. 15 (cyber-resilience) scoped narrowly.

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, Consistently

Identical signatures across Rust, Python, and TypeScript: fingerprint a prompt once, match it anywhere.

Defense-in-Depth, By Design

A detection signal, not a firewall — pair with exploit-feed matching to catch suspicious and known-bad inputs.

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.

The ØDIN Prompt Toolkit validates controls where your app receives input: in-process, in your language, on your hardware. Score each prompt, check it against known exploits, and decide before it reaches the model, without exposing content.

Core Capabilities

SusFactor Classifier

Local 0–1 suspicion scoring for control failures and untrusted-input attacks (~50–200ms per prompt).

LSH Signatures

256-bit SimHash fingerprints that preserve semantic similarity for dedup and matching.

Exploit Intelligence

Token-based sync of ØDIN signatures, with full and incremental updates.

Cross-Language SDK

Rust, Python, and TypeScript with byte-identical, reproducible behavior.

ØDIN

Secure People, Secure World.

We unite security experts to safeguard humanity.

HTTPS://ØDIN.AI

ØDIN@MOZILLA.COM