// VARDA PRESENTS

# SILA

## Alignment that holds.

A safety layer inside the model, while it runs.

Zero Trust for AI

// THE PROBLEM

# AI can be talked out of its own rules.

**01**

It's called a jailbreak, and the techniques are public.

**02**

Most guardrails sit outside the model.

**03**

Usually a second model, screening what goes in and out.

// THE ANSWER

# SILA works inside the model, while it runs.

### One dial

Up, down, or off.

### No retraining

The model file is unchanged.

### No second model

No cloud. Runs on your own hardware.

// RECORDED TEST · SAME MODEL, SAME PROMPTS

# Optimized jailbreaks: 38% → 0%

150 attacks and 50 normal requests. SILA off, then on.

| | SILA OFF | | | SILA ON | |
|---|---|---|---|---|---|
| **Plain harmful requests** |  | 1/50 | → |  | 0/50 |
| **Disguised attacks** |  | 11/50 | → |  | 0/50 |
| **Optimized jailbreaks** |  | 19/50 | → |  | 0/50 |
| **Normal requests not refused**<br>refusal-phrase check |  | 50/50 | → |  | 50/50 |

Qwen2.5-7B-Instruct on one local GPU · attacks judged by the public HarmBench classifier · normal requests by refusal-phrase check · Varda, October 2026

// HEAD TO HEAD · LLAMA GUARD 3

# Guard-level blocking.
# Without the guard's hardware.

| | Llama Guard 3 | SILA |
|---|---|---|
| Blocking these attacks | **Baseline** | **Matched** |
| Extra GPU memory | **5.6 GB** | **None** |
| Extra model calls per request | **2** | **None** |

Same test as before · Qwen2.5-7B-Instruct · Varda, October 2026

// ALSO · HALLUCINATIONS · PRE-REGISTERED

# Catches and stops wrong answers.

| Wrong answers stopped, at 5% of right answers lost | Mistral-7B | Qwen3-8B |
|---|---|---|
| Standard confidence method | **27.4%** | **39.8%** |
| SILA's loop | **43.2%** | **57.2%** |

**Most stops happen before the model says a word.**

TriviaQA closed-book · 2,500 fresh questions per model · scored once · Varda, October 2026

// WHERE THIS GOES

# Rules held from the inside.
# Proof a buyer can check.

### The owner's dial

Every model an organization runs, held to its rules from the inside.

### Re-run any claim

On our hardware, with your own prompts. You judge the outputs.

// BUILT FOR WHERE IT MATTERS

# Designed for the open models you host yourself.

**RECORDED SAFETY RESULTS**

Qwen · llama.cpp

**BUILT FOR**

Defense · Intelligence · Critical infrastructure · Banking · Medicine

✓ Your own hardware ✓ Fully offline ✓ No cloud ✓ No retraining

// WHY VARDA

# AI has to hold to its rules. We show our work.

### SILA

Holds from inside the model.

### Proof ledger

Proven, partial, and failed. All listed.

### Evidence

Every result hashed. A buyer can check.

// VARDA

# Invest in the layer that holds.

## Chasing Amazing!

varda@vardasila.com

360.447.8845

vardasila.com