SILA Public Overview / Leave-Behind v2

Website-safe public overview

Cover

SILA

Safety below the surface.

Representation-layer steering for frontier AI systems.

Varda LLC - Public Overview

The problem worth solving

Today's AI safety lives on the surface - and the surface keeps getting talked around.

Frontier systems are guarded by impressive layers: system prompts, refusal training, constitutional methods, classifiers, monitoring, and output filters. They matter.

Yet every new jailbreak tells the same story: a route found through the model's own behavior, beneath the layer built to stop it.

The surface was never the real line of defense. It was just the visible one.

A different layer

SILA works where behavior is actually formed: inside the model, at inference, in the representation path itself.

Not another wall around the model. A runtime safety control designed to change the model's footing.

Most safety asks the model to behave. SILA shapes the place behavior comes from.

What SILA does

A runtime, representation-layer safety control - built to:

- read intent as it forms, during inference;
- steer the model's internal state toward a chosen safety constitution;
- make jailbreaks measurably harder to land;
- keep helpful answers helpful when the request is benign;
- strengthen the safety stack you already trust, from underneath.

It composes. SILA does not replace your safeguards - it gives them a floor they never had.

A dial, not a wall

SILA is reversible and continuously tunable: full on, full off, or anywhere between, set live without retraining.

Swap the value set and the same mechanism can enforce generic refusal, a specific constitution, or a brand's own voice.

One control. Your rules.

Honesty as a standard

We do not sell the word "unbreakable." We sell the measurement.

SILA's value is proven the only way that counts: the same model, same prompts, run with the layer off and on, and the difference shown plainly.

Across that on/off structure we track attack-success delta, refusal lift on harmful prompts, benign-refusal rate, and helpfulness and capability preservation - so safety is never bought with a model that will not talk.

The full benchmark detail comes in the private briefing, under NDA.

Where others market, we measure.

The invitation

SILA is ready for private technical review by safety, security, and frontier-model teams.

One conversation, under NDA: the live demonstration, the methodology, and the numbers - in the room.

Request a private SILA technical briefing.

Janice Bryant

Chief Executive Officer

Varda LLC

janice.b@vardasila.com

360-507-8745

www.vardasila.com