01 · THE PROBLEM

The attack that gets through is not in the prompt. It's in the conversation. SPAN · FLAG

And the defences do not hold where the attack lives. On HarmBench, multi-turn human jailbreaks exceed a **70% attack success rate against defences that report single-digit rates on single prompts**. Every turn stays inside the safety distribution. The conversation is the attack.

IBM, Cost of a Data Breach 2026 · Li et al., Scale AI, arXiv:2408.15221

1 in 4
organisations suffered an AI-driven breach in the past year

+56%
year-over-year rise in AI-driven attacks

$11.5M
average cost of a breach in the United States, a record high

02 · THE APPROACH

One API. One layer. In front of any model.

A model-agnostic layer that reads the mechanism of a manipulation attempt and the intent behind the request, not its surface words, and flags the input before it reaches the model it defends. Reading mechanism rather than wording is what it is built on, so rewording, encoding or a language swap does not reset the read. The engine is self-hosted (open weights running on NVIDIA GPUs) and ships as managed SaaS, private cloud, on-premises, or fully air-gapped.

03 · WHY IT WORKS

Mechanism, not signature.
Phrasing changes. Mechanisms persist. A novel attack is scored by the mechanism it uses, not matched against a list of known strings.

Fail-closed by default.
On any error the service returns an error, never a false 'safe'. No code path answers PASS on failure. Security-product discipline.

Prompt and conversation.
Scored as a single input and across a whole conversation, so an attack assembled over several turns is caught and located. Risk is assessed across the whole conversation, not one message at a time.

04 · THE PROOF, IN SHORT

99.66% separation, attack vs benign · **0.00%** false positives, ordinary traffic · **35+** languages, out of the box

Measured against the live API, not a laboratory harness. **The full measurement, the properties that always hold, and our known limits are overleaf.**

05 · HOW IT WORKS

01 · INPUT
`POST /v1/analyze`
One endpoint takes a prompt, a conversation transcript, a retrieved document, or tool output.

→

02 · DETECTION
`evidence spans · turn findings`
Returns a risk score, the quoted evidence, the turn at which the assessment changed, and a plain-language reason.

→

03 · ACTION
`PASS · LOG · FLAG · BLOCK`
A recommendation, not an enforcement point. How you act on it stays inside your application.

Bring your own attacks.

A free, time-boxed, revocable evaluation key to the live API, and 30 minutes with whoever on your team wants to try to break it. The interesting result is what happens to content we have never seen.

Request a key →

06 · MEASURED ON THE LIVE API

Produced against the deployed endpoint, not a laboratory harness.

The same endpoint, authentication, rate limiting and output sanitization an evaluator receives. No result was taken from a bypass path.

99.66%
Separation, attack vs benign
ROC AUC, measured on the live API

0.00%
False positives, ordinary traffic
ordinary production-shaped traffic

3.78%
False positives, adversarial-adjacent
deliberately hard, benign look-alikes

95.35%
Attack detection
scored strictly, unsure counted as a miss

35+
Languages out of the box
no per-language configuration

0
Errors answered PASS
fail-closed, every request

Read the scope, not just the number:

- The two false-positive rates measure different populations, so they are reported separately, never blended.
- 95.35% is the floor, not the ceiling: where we were unsure, we counted the prompt as a failure rather than give ourselves the benefit of the doubt.
- Known limits, disputed items, and what has not yet been measured are published in full.

07 · PROPERTIES THAT ALWAYS HOLD

A verdict cannot be talked back down.

A conversation that reaches a blocking verdict stays there. Arguing with the assessment, or apologising and asking again, does not undo it.

The request shape does not change the answer.

The same text submitted through different request shapes returns the same recommended action. There is no submission path that is quieter than the others.

Language is not a way around it.

More than 35 languages are covered out of the box, with no per-language configuration and no separate model to deploy. Translating an attack does not make it invisible.

Failure is closed, never silent.

Backend failure, an over-budget conversation or an internal error returns an explicit error status. There is no code path that answers PASS on error.

Evidence carries two indices.

A response reports the turn at which the assessment changed and the turn the quoted text came from, separately. A payload planted early and triggered later reports both.

Internal identifiers never reach the wire.

Responses carry public-vocabulary labels only, enforced by a fail-closed output gate that refuses to emit an internal identifier. Verified across every request in this measurement.

08 · REWRITING DOES NOT RESET THE READ

6
Tested languages
en · fr · es · zh · ar · ru
More to be validated soon

98.52%
overall accuracy
across the multilingual test set

100%
encoded attacks caught
base64 · Morse · ROT13

0
false positives
on the benign set tested

The same attack written in another language, or carried in another encoding, is scored on the mechanism it uses rather than the characters it arrives in. Adversarial recall held at 100% in English and Chinese, and at 95% or above in French, Spanish, Arabic and Russian. The engine supports more than 35 languages out of the box; the six above are those tested so far. **Same engine and policy as above, scored directly rather than through the API**, so it is reported as its own run and not folded into the figures above.

Verifying this independently

- **Bring your own attacks.** The corpus above is ours; the interesting result is content we have never seen.
- **Errors are distinguishable from verdicts.** An error never carries a verdict, so a failure cannot read as benign.
- **Every verdict is inspectable.** Quoted evidence, the turn the assessment changed, and a plain-language rationale.
- **Keys are per-evaluator, time-boxed, revoked on request.** Full API reference and measured results accompany this brief.