PRESS KIT

# GuardBotAI security press kit

Materials for journalists, analysts and enterprise buyers. Every claim below links to something you can check on this site. We do not publish customer names, certifications or metrics we cannot evidence.

## One-line description

GuardBotAI is an AI runtime security proxy: every request your AI agents make passes through GuardBotAI before it reaches a model or a tool, where it is inspected against your policy, blocked or stripped if risky, and recorded in a tamper-evident log.

## Boilerplate (100 words)

GuardBotAI is an AI runtime security platform from Venture Group AI. Enterprises route their AI and agent traffic through the GuardBotAI gateway, where a deterministic policy engine inspects each request for prompt injection, jailbreak attempts, data leakage and unpermitted tool calls, then blocks, redacts or forwards it according to rules the customer writes. Every decision is written to a tamper-evident record. Each guardbot runs inside a triple-layered sandbox — Practice, Leashed and Live — and breakout signals drop it a layer instantly with a real-time alarm to the owner. Loosening containment always requires a recorded human approval.

## The triple-layered sandbox

| LAYER | WHAT IT PERMITS | HOW YOU LEAVE IT |
| --- | --- | --- |
| 1 · Practice | Nothing is forwarded. Requests are judged and recorded only. | A named person approves promotion; the approval is recorded. |

| LAYER | WHAT IT PERMITS | HOW YOU LEAVE IT |
| --- | --- | --- |
| 2 · Leashed | One approved destination, allow-listed tools only, tighter size and rate caps. | Upward by approval; downward automatically on any breach. |
| 3 · Live | Full approved operation under the owner's own policy rules. | Dropped instantly on a breakout signal — unlisted tool, unapproved destination, rate spike, block storm, or the same blocked pattern across several agents within minutes. |

## Real-time breakout alarms

A demotion writes its reason into the tamper-evident record, emails the workspace owners within seconds — one message per incident, not one per request — and leaves a banner in the control center until a person acknowledges it. One control drops an entire workspace to Practice. The swarm case is explicitly covered: when the same blocked pattern appears across several of a customer's agents inside a few minutes, that is treated as coordination and triggers containment.

## Claims and where to verify them

**Every request is inspected before it reaches a model or tool.**

Live traffic recorded in the public demo walkthrough, including HTTP 403 refusals with rule ids.

Check it

**Decisions are recorded in a tamper-evident chain.**

Each event stores the hash of the previous event; the demo page shows six linked records.

Check it

**Every guardbot runs inside one of three containment layers.**

Practice, Leashed and Live layers, documented with the escalation and demotion rules.

Check it

**Breakout signals drop a guardbot a layer automatically and alert the owner.**

Recorded demotion after 25 blocked requests in five minutes, with the HTTP 503 response and the alarm record.

Check it

**Availability figures are measured, not asserted.**

The public status page reports real probe results for the database and the policy engine.

Check it

**GuardBotAI holds no external security certification today.**

Stated plainly on the Trust Policy page, alongside a SOC 2 Type II intent after launch.

Check it

## Context and background reading

Microsoft AI CEO Mustafa Suleyman described AI models' ability to tamper with their own working memory as a "serious situation" following OpenAI's disclosure of further incidents of concerning model behaviour, as reported by CNBC. GuardBotAI's position is narrow and stated as such: alignment work happens at training time, and a runtime control layer is what an operator can own, inspect and evidence afterwards. Our analysis is in agent sandbox escape.

## Brand basics

- Product name: GuardBotAI (one word, capital G, B and AI). Not "Guard Bot" or "Guardbot AI".
- Parent: Venture Group AI.
- Category: AI runtime security proxy.
- Mark: a shield with a guard robot motif, used in primary green on a dark background.

## What we will not say

- We hold no external security certification today and will not imply one.
- We publish no customer names, logos or testimonials we have not been given permission to use.
- We quote no performance or accuracy figure we cannot reproduce on request.
- We do not claim to prevent every form of agent escape. We contain, record and alert.

## Media contact

Press and analyst enquiries go through the contact form. Enterprise evaluation and procurement questions go through enterprise. We are happy to run a live enforcement test with a journalist watching.