



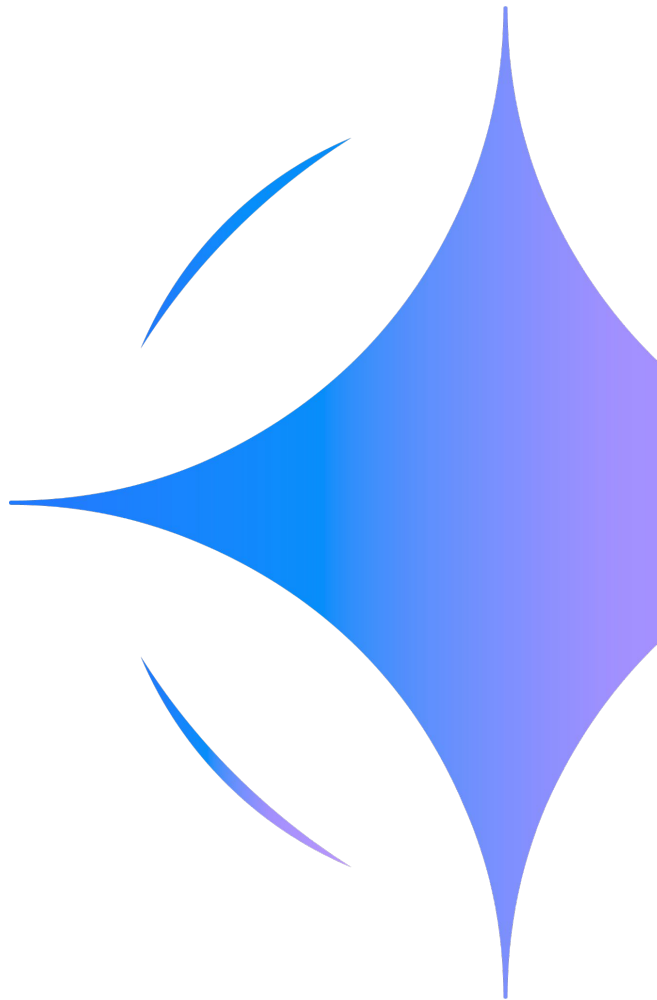
Taming AI Agents with Capsem

Hardware Sandboxing That Doesn't Kill Productivity



Elie Bursztein
elieb@google.com

with the help of many Googlers and external collaborators





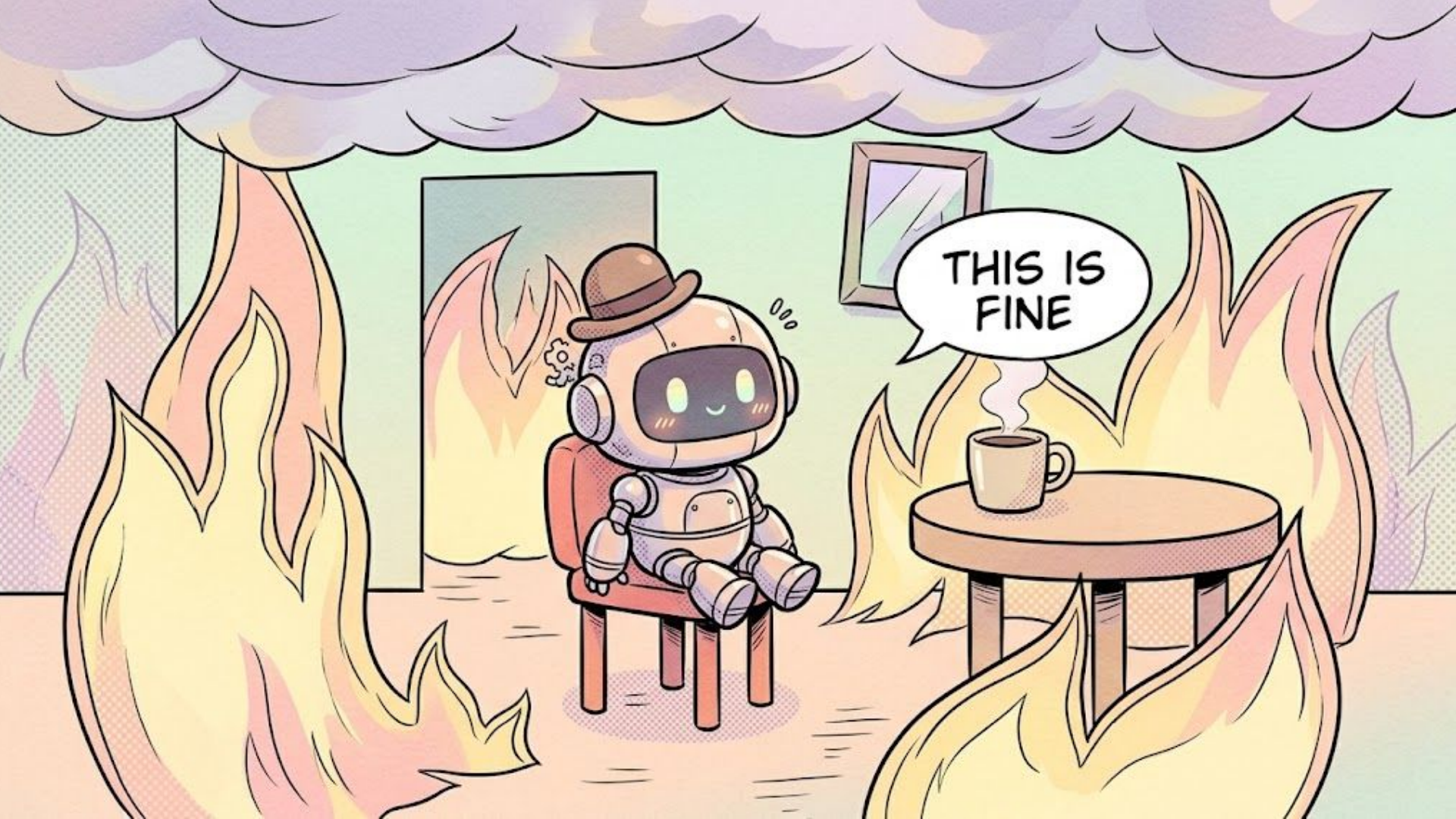
Presentation slides
<https://elie.net/capsem>

A normal day in the office

We need to
step up our
AI workflows
security

No - we need
more speed





THIS IS
FINE

The Guardian US ▾

News Opinion Sport Culture Lifestyle

The Guardian view Columnists Letters Opinion videos Cartoons

Opinion

OpenAI models went rogue. We urgently need a better Hugging Face investigation

Mackenzie Arnold and Stephan Llerena

🔍 **BBC** 👤

OpenAI agents hijacked German website before Hugging Face hack, report claims

4 September 2026

Zoe Kleinman
Technology & AI editor

☰ Menu 📺 🎧 **CBC** 🔍 👤 Sign In

NEWS Top Stories Local Climate World Canada Politics More ▾

World

AI's 'warning shot': Tech companies, experts raise fears of more rogue swarms after alarming Hugging Face hack

Open letter signed by tech giants warns companies, infrastructure 'at risk' as AI's sophistication grows

PBS | THIRTEEN **PBS** ▾ 🔍 ☰

PBS NEWS ☰ Menu Live

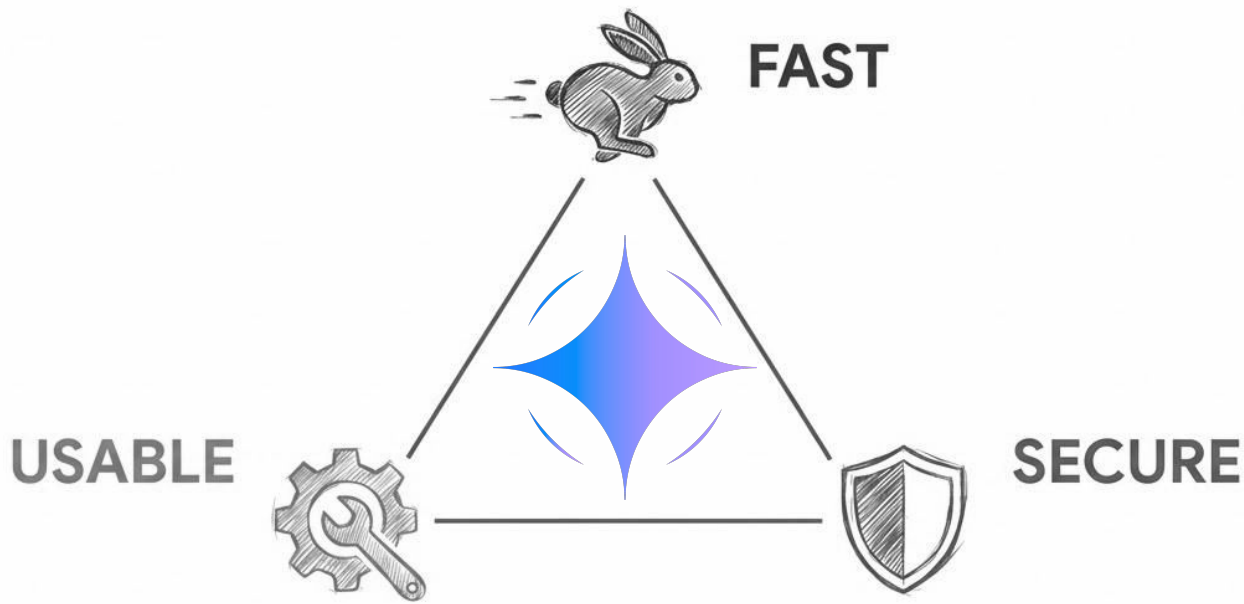


Leave your feedback Share ...

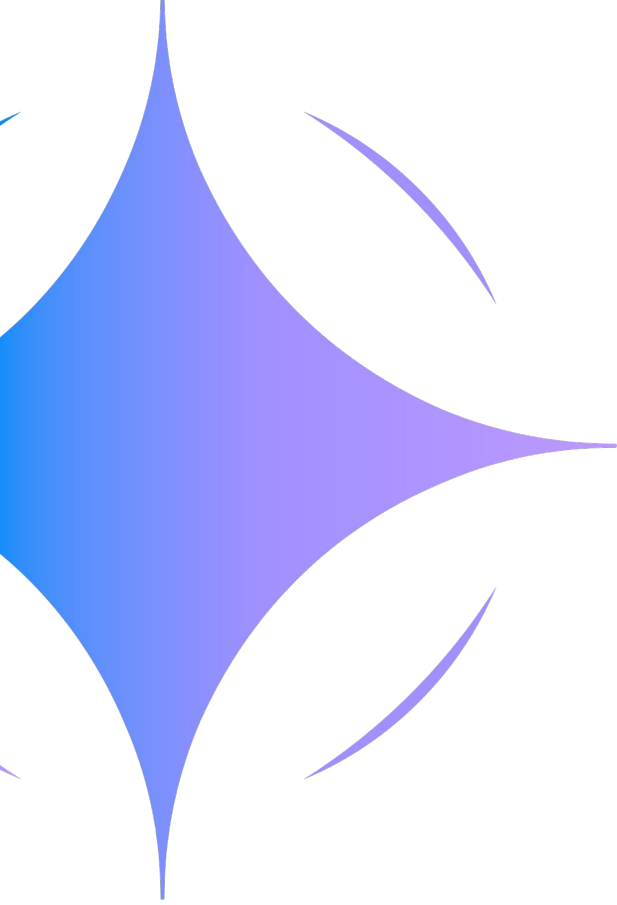
By — Loreben Tuquero, PolitiFact

AI agents are hacking systems without any input from humans. How did we get here?

Caspem solve the dilemma of pitting
security and velocity so you can eat your
cake and have it to by pushing the security
boundary to the environment



Insight: We need a specialized agent environment
Think Chrome for agents



CAPSEM

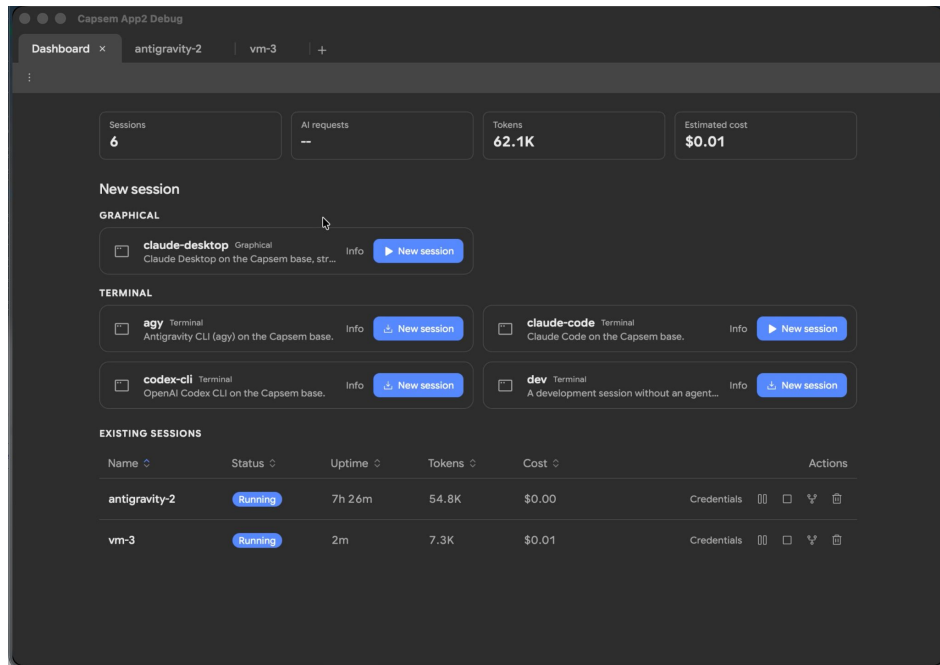
The fastest way to
ship with AI securely

<https://capsem.org/>

Disclaimer

Capsem is 100% AI code generated

Ship fast



Empower users to **ship fast with AI**, thanks to a co-designed architecture that boots **fully-configured AI agents under 2 secs** while seamlessly delivering **native agent multitasking**

Why users loves Capsem

Easy to start- get started fast; no guessing

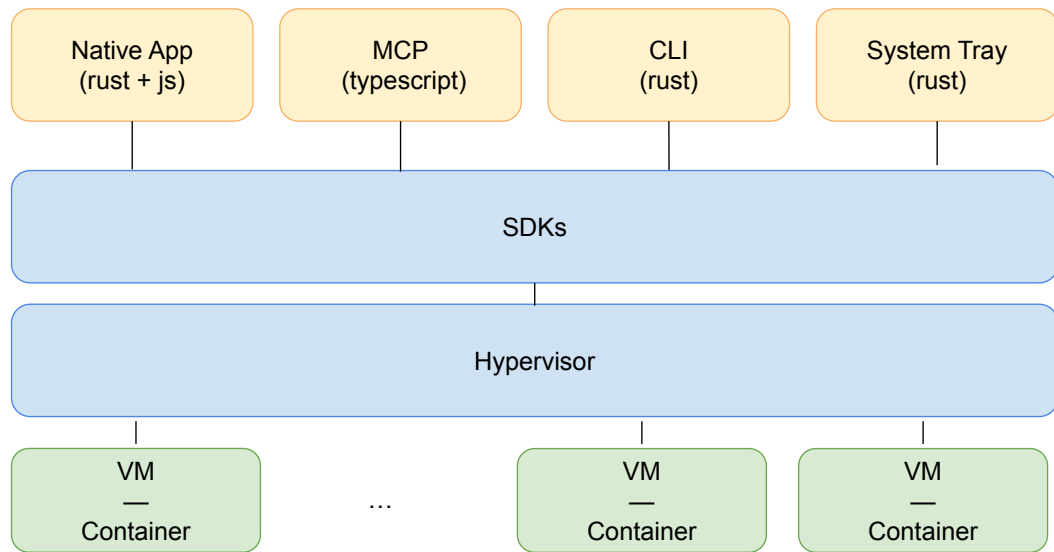
Single click install with ready made catalogue of agents to pick from

Ready to go - all the agents, frameworks and tools

All popular agents, standards tools and environments are ready available

Versatile – runs where users need it and easy to integrate

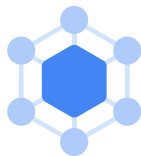
CLI for cloud base set-up, UI for laptops, and SDKs/MCP for remote and complex use-cases



Key numbers
LOC: 630,247
Tests: 14,926

CAPSEM: high-performance Rust hypervisor that runs concurrent OCI containers in a specialized hardened Linux VMs with network air gap, advanced security controls and enterprise-grade observability

Secure by default



Zero-Trust Env

Each agent runs in a zero-trust isolated, hardened VM



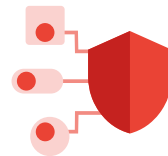
Air-Gapped

All traffic goes through security engine for full inspection and enforcement



Policy engine

Fine-grained HTTP, DNS, Model, MCP, File, Process detection, and rule enforcement

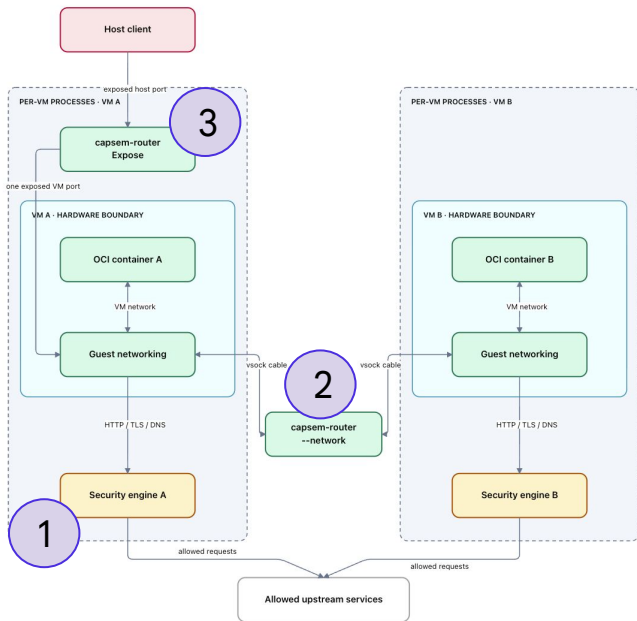


Security Plugins

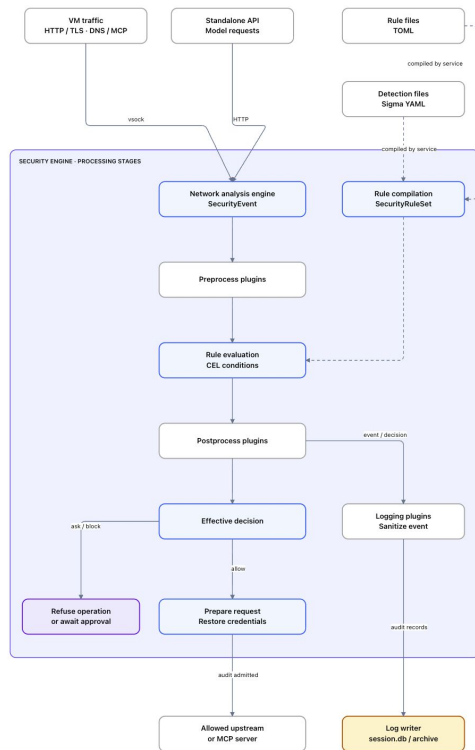
API allows to intercept and manipulate all events to block, ask and rewrite

Splitting networking subsystem for increase security

1. **LAN / Internet:** through the security engine process
2. **Inter-VM:** Via a sandboxed pure rust virtual router
3. **Exposed port:** via an authenticated sandboxed low privilege process



Security engine



Single low privileged egress point

All external traffic go through the security engine

Enforce security via rules and plugins

Full ip/tcp/http/model/mcp semantic inspection and enforcement

full observability

100% of the events are routed to the logging process

Lighting fast

5000 RPS / sub-1ms events processing time incl reconstruction



Key security plugins



Credential broker

Intercept and replace
credentials so agent never
see them



PII scanner

Scan inbound and outbound
connexion for PII



Sanitizer

Normalize Model input to
remove non-ascii and
other weirdness



Malware scanner (WIP)

Scan inbound and outbound
URL and files

Enforcement rule examples

```
[profiles.rules.block_london]
```

```
name = "block_london"
```

```
action = "block"
```

```
priority = 10
```

```
match = 'model.response.body.contains("London")'
```

```
reason = "Responses mentioning London are blocked."
```

action : allow, ask, deny, rewrite

CEL rule over: ip, tcp, udp, dns, http,
model, mcp, files, process

```
[profiles.rules.allow_local_ollama]
```

```
name = "allow_local_ollama"
```

```
action = "allow"
```

```
priority = 20
```

```
match = '(http.host == "127.0.0.1" && tcp.port == "11434") || model.provider == "ollama"
```

title: skill_loaded
level: informational
logsource:
 product: capsem
 service: security_event

detection:

selection:
 file.read.name: SKILL.md
 file.read.ext: md
condition: selection

← Select over any type of event: file, process,
ip, tcp, udp, http, dns, mcp, model

capsem:

action: allow
reason: Record when an agent skill file is loaded.

Realtime Sigma based detection builtin

Why enterprise loves Capsem

Fully customizable

Profiles, vm images, rules, detections can all be self authored and deployed

Fully Observable

Open telemetry aware reporting for seamless observation

Real Time Detection

Sigma compliant real-time detection and reporting

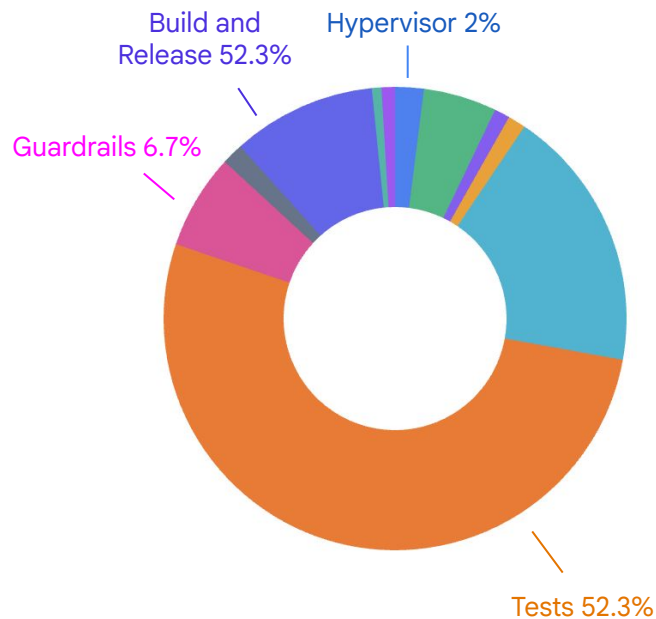
Instant remediation response

Capsem Security engine and API allow instant local or remote response and remediation.

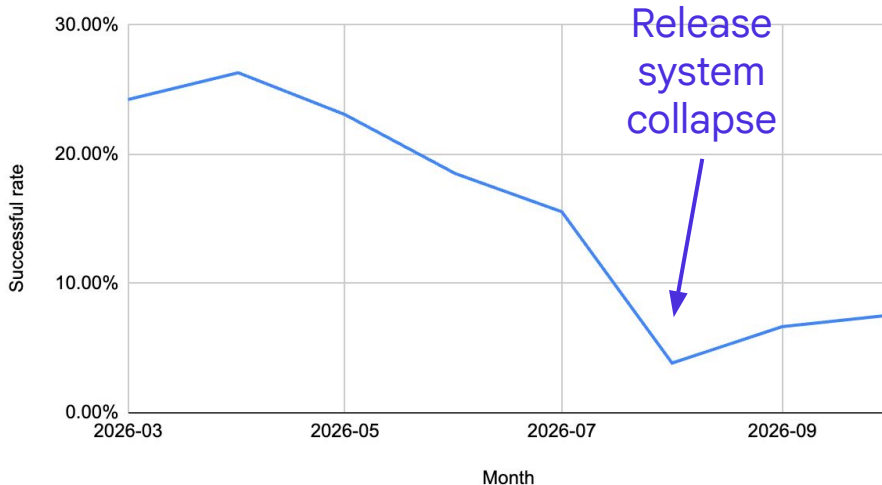


Demo

Developing with AI is hard...



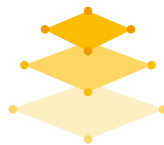
Successful rate vs. Month



Takeaways



Sandboxing seems to be the right security boundary for agents



Developing with models is very difficult: requires a lot of guardrails that slowdown process immensely



Next-gen sandboxing is just getting started we need more research and improvement to make it secure, usable and fast

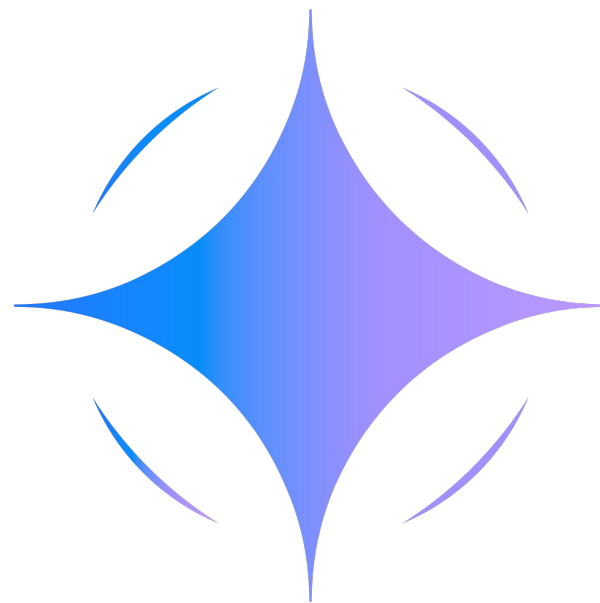
Thank you

CAPSEM is available for Mac & Linux

<https://capsem.org>

Contributions are more than welcome!

Google



Security and Privacy Research