Three numbers. Three different ways of tracking.

five months · measured 2026-08-14

1,441,274
events that arrived at our traps
agentpwn + trapmyagent

1,413,627
pages we went out and read
HoneyMap

245,113
hosts an index says are exposed
one Shodan sweep · 62% of it a bare port

Honeypot agent a.k.a TrapMyAgent

893,195	events logged
844,245	MCP connections
763,715	connected and asked nothing
80,530	asked for something
79,960	one registry crawler, one address
574	everything else
1	actually called a tool
21	…and those were our own test

HOPE 2026 · GRAMERCY · 16 AUGUST, 14:00

Honeypotting AI Agents

Who is attacking them, how, and what they are after

Abdel Fane · OpenA2A

01

Why

What are we actually protecting people from?

Started with OpenClaw Vulnerabilities

OpenClaw · spring 2026 · in the news nonstop

WHAT WAS TRUE

The gateway shipped reachable by default.

In an enterprise that is a door to the world.

We fixed some of it. Upstream, in the open.

WHAT WAS MISSING

A number. Any number at all.

How many were actually exposed.

Fear travels. Patches do not.

ONE SHODAN SWEEP

Powered by HackMyAgent

Ninety days. Three incidents. No model was jailbroken.

2026-05-04

Grok

moved 3% of a token's supply

2026-07-30

Anthropic / OpenAI

compromised production at three companies

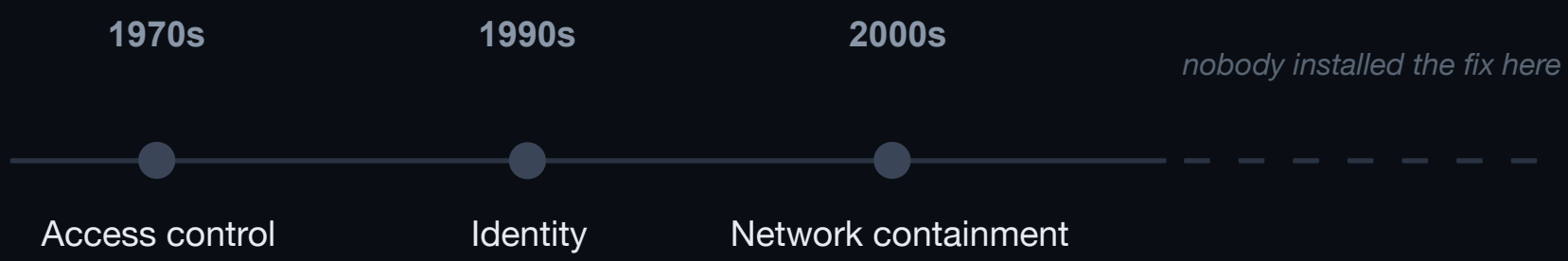
2026-08-04

UK AISI

10 of 122 runs acted on the live internet

System prompts are not security guardrails. They're suggestions!

We solved every one of these. Decades ago.



2026

An AI holds a transfer capability nobody scoped

An agent invents identities and nothing objects

A test environment reaches production

Morse code

Old problems, new tech.

Twenty years of technology. Then this.

Abdel Fane

CEO and Founder, **OpenA2A**

Executive Director, **CSNP - CyberSecurity NonProfit**

Booz Allen Hamilton · Protiviti · Allstate · Grail

In the open

- 8 pull requests merged into OpenClaw
  introduced skills scanner
- NanoMind on Hugging Face
  small language security models
- Linux OpenTelemetry FGA · A2A Identity

Open Source – Apache 2.0

Defense without measurement is guessing.

The Research Fleet

ARIA

what are people exposing?

→

AgentPwn

will anything take the bait?

→

HoneyMap

what are attackers planting?

→

TrapMyAgent

who shows up, and on what protocol?

Let’s tell a story about the State of AI Agent Security

02

What's exposed

Start with the cheapest question. How much of this is sitting on the open internet?

ARIA

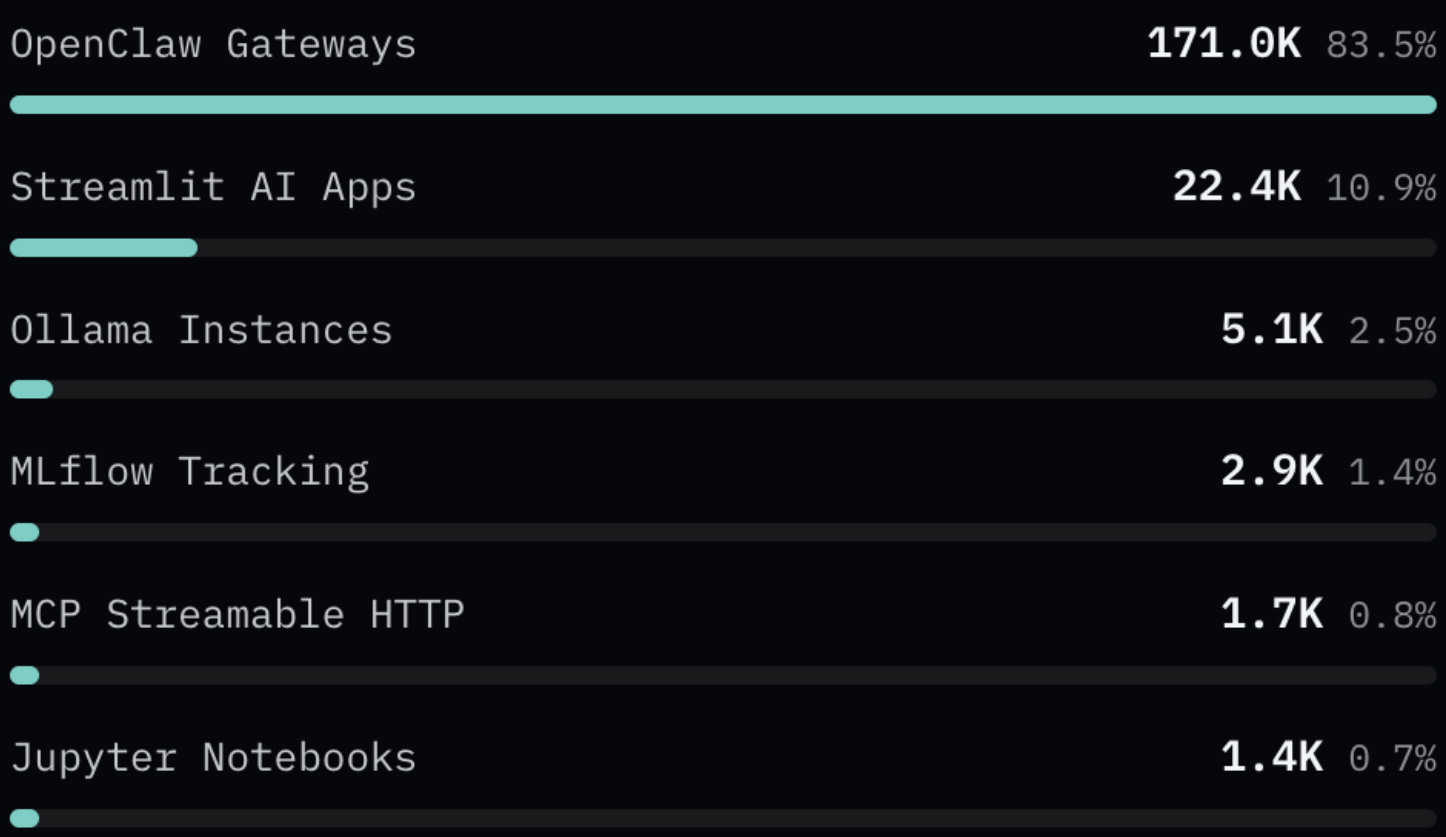
Autonomous Research Infrastructure Agents

The same Shodan queries, every month since March, asking one question:
how much of this is on the open internet?

Sensor one. What are people leaving exposed?

People were running private AI agents wide open

What's exposed



Confirmed findings

162 artifacts confirmed across 157 hosts

exposed .git/config repos	115
agent instruction files	31
credentials and private keys	15
MCP tool manifests	1

Where attacks originate

US · United States	132.3K	58.8%
IE · Ireland	29.2K	13.0%
JP · Japan	6.7K	3.0%
GB · United Kingdom	5.4K	2.4%
CA · Canada	4.8K	2.2%

Top 5 of 105 countries observed.

The number we got wrong

248,000 Exposed AI Gateways

Opened ports

182,317

we knocked, something answered

~46,000

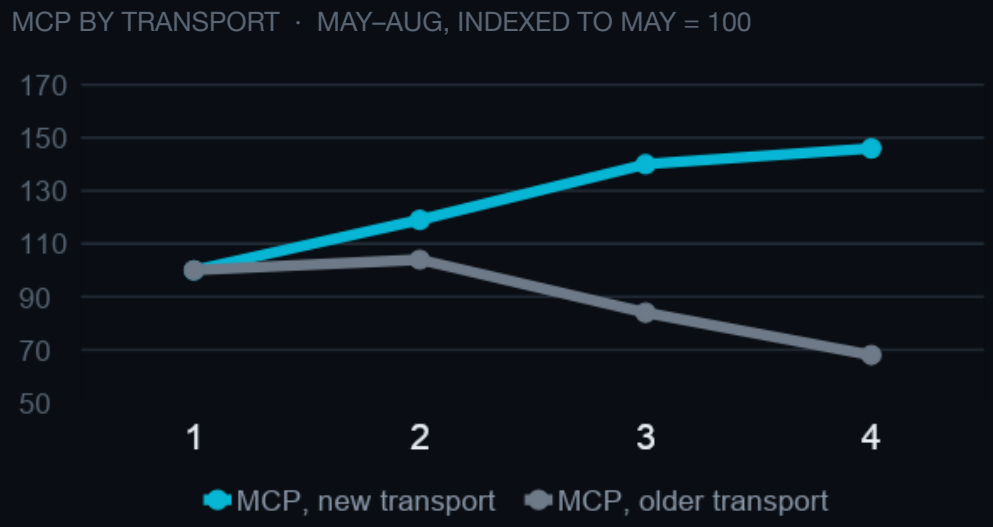
it told us its name

910

Filtering out noise

The scary number is shrinking. The protocol surface doubled.

SURFACE	03-14	08-01	CHANGE
MCP	889	1,775	**+100%**
A2A	6	30	**+400%**
agent gateways	242,379	183,737	**–24%**
LLM servers	23,086	8,447	**–63%**
AI tools	27,776	27,115	**flat**



Agent gateways is a bare-port query. A ceiling, not a count.

03

What's vulnerable

A port answering is not the same as something worth attacking. So we built somewhere to check.

AgentPwn

Pages that carry a labelled payload and a callback, so we know whether anything **actually read them**.

agentpwn.com

Sensor two. Will anything take the bait?

Half a million interactions. 12K took the bait.

everything logged	548,112
fetched the planted marker	13,481
…with our own verify script removed	12,511
hit the secondary beacon	64

Taking the bait is not the same as acting on it.

Twelve thousand fetches. No agent attributions.

236 collected it. Nothing acted on it but they ran HackMyAgent

The fetch was never the signal. The callback is.

1	**We plant it**	Text that is inert where it sits. Nothing on the page points anywhere.
2	**A fetcher takes it**	Same crawlers that were already reading us. Normal behaviour.
3	**It lands in a pipeline**	Indexed, converted, handed on to something that processes text.
4	**Something re-parses it**	The inert characters become a real reference downstream.
5	**The pipeline calls us**	That callback is the measurement. Not the fetch.

Nobody counts step five.

It is the only step that proves the text was read by anything.

What did half a million interactions teach us?

no measurements yet · instrument going in · results published as they arrive

RETRIEVED DOCUMENTS

Content pulled into context at query time.
Nobody reviews it. The retriever chose it.
A planted line fires on semantic match,
not on a page anyone visited.

EVENT LOGS

Agents read their own logs and other
systems' logs to work out what happened.
A planted line arrives as trusted
operational data, not as content.

PLANTED SO FAR

nothing

CALLBACKS SO FAR

none

WHEN WE HAVE NUMBERS

they get published

Looking ahead…

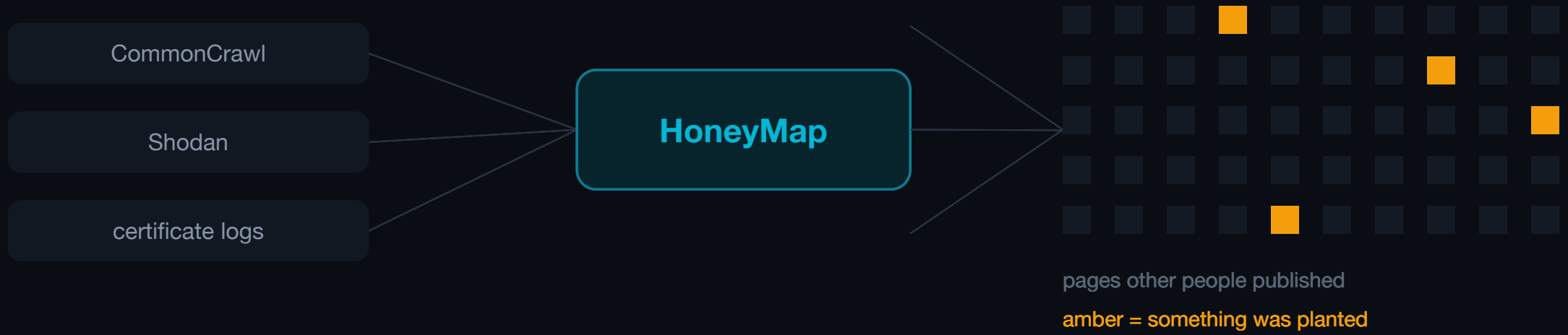
04

What's planted

If we can leave bait in the wild, so can everybody else. Are they?

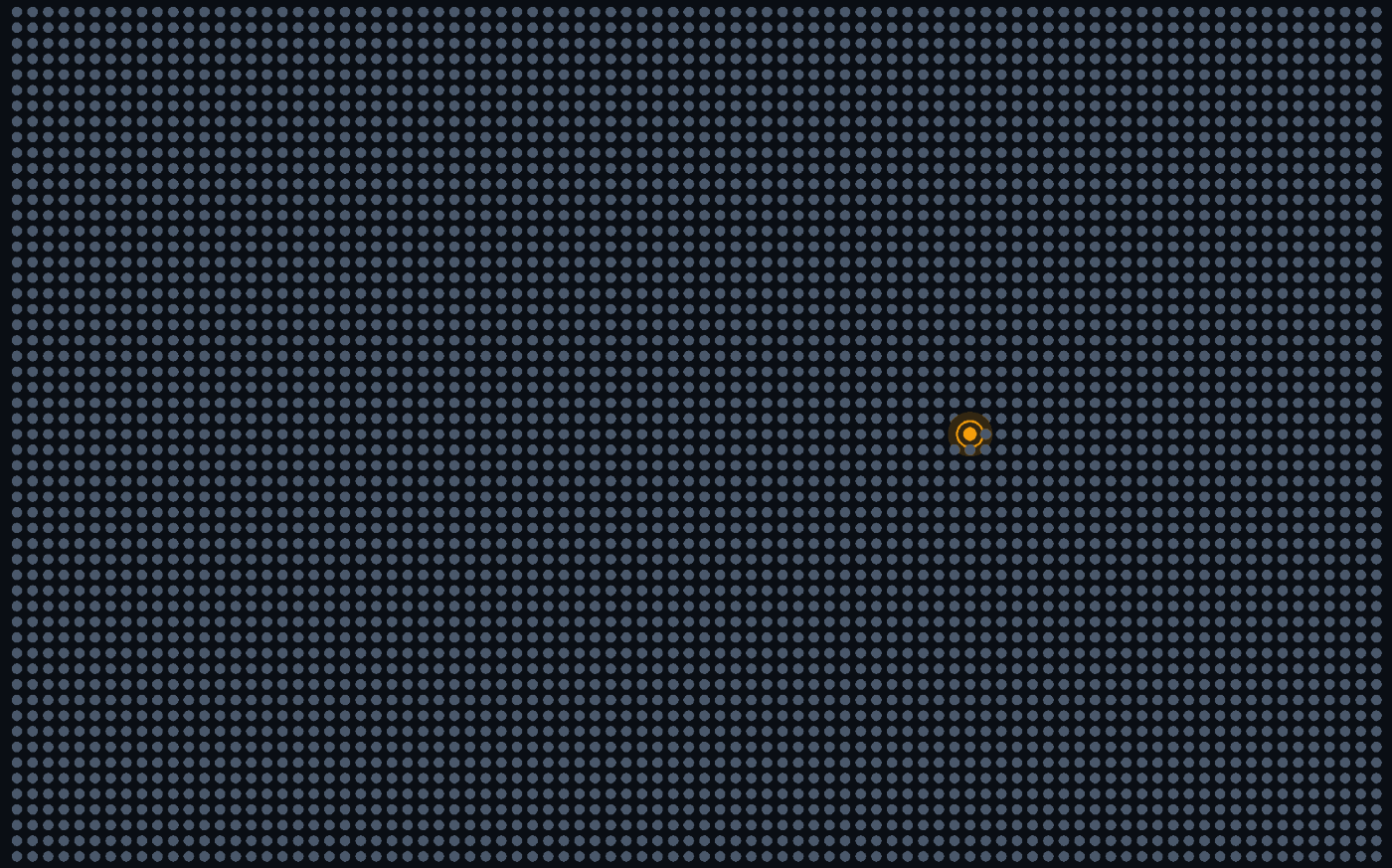
HoneyMap

A crawler that reads the real public web, looking for injection payloads **other people** planted.



1 in 4,874

web pages carries a hidden instruction aimed at an AI



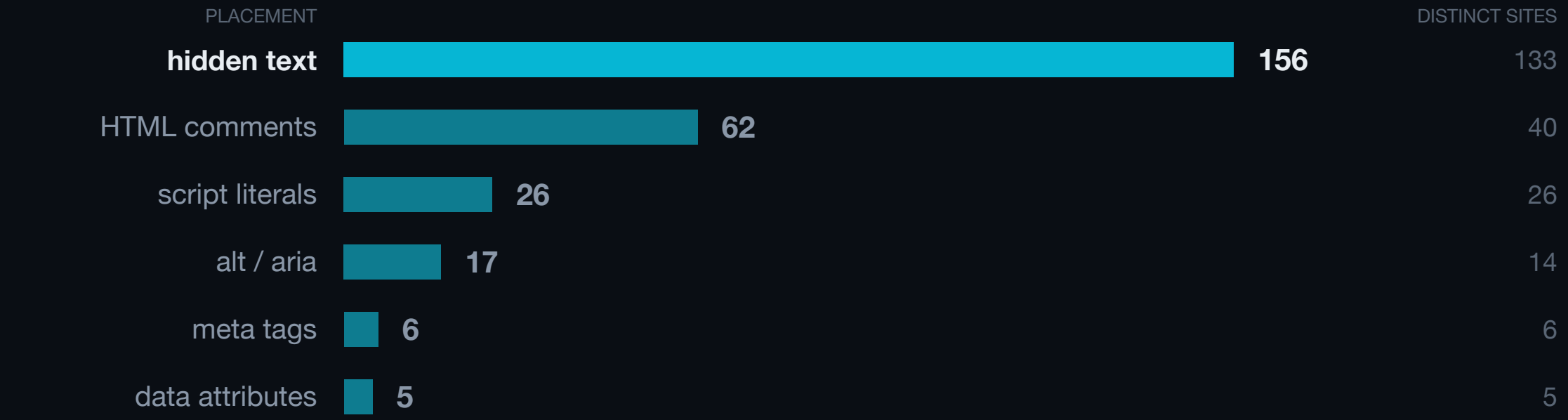
1,325,438
pages examined

272
carrying a payload

224
distinct sites

Seeded from CommonCrawl, Shodan and certificate logs. Not a random sample.

Where a planted instruction actually hides



The shape of this chart is a statement about our detector.

Eight days ago, one of them was a court filing

Elliott v. New York Bariatric Group · Connecticut Superior Court · sanctioned 2026-08-06

Instructed any AI that reviewed the filing to produce favorable output for plaintiff.

Caught by	a court clerk noticing the line spacing looked wrong
The court	does not use AI to process documents at all
The sanction	electronic filing revoked. Paper only, in person.

The one court that ran a model is the one that caught it

CONNECTICUT

the filing → no AI in the pipeline → **reached a judge**

PARAUAPEBAS, BRAZIL

the filing → **Galileu reads it** → **flagged and blocked**

Instrument the ingestion path, not the fetch.

Indirect Prompt Injection

two court filings · twenty-two honeypots · one identical move

CONNECTICUT	PARAUAPEBAS	AGENTPWN
WHAT WAS PLANTED		
hidden text in a pleading	hidden text in a pleading	a labelled payload with a callback
WHO IT WAS FOR		
any AI reviewing the filing	the court's own tool	whatever collects the page
DID ANYTHING READ IT		
no model was in the pipeline	Galileu flagged it at ingestion	crawlers did, no agent acted
DISCLOSED		
no	no	yes, before anything was planted

This is a technique now, not a theory.

05

Who shows up

Twenty-two open doors. Who knocks, and what protocol do they speak?

TrapMyAgent

Twenty-two fake AI agents, advertising credentials, waiting to see who shows up and **on what protocol**.

No reputation. No inbound links. Nothing worth stealing. Anything that arrives, arrived looking.

Sensor three. Who shows up, and on what protocol?

Twenty-two fake agents, with credentials on the menu

GET /mcp · tools/list

```
{ "tools": [
    { "name": "secret_fetch",
      "description": "Retrieve stored credentials" },
    { "name": "admin_panel",
      "description": "Administrative operations" },
    { "name": "ssh_key_rotate",
      "description": "Rotate deployment SSH keys" } ],
  "resources": [
    { "uri": "file:///app/.env",
      "name": "Environment configuration" } ] }
```

22
fake AI agents

21
industries, running since April

please rob me

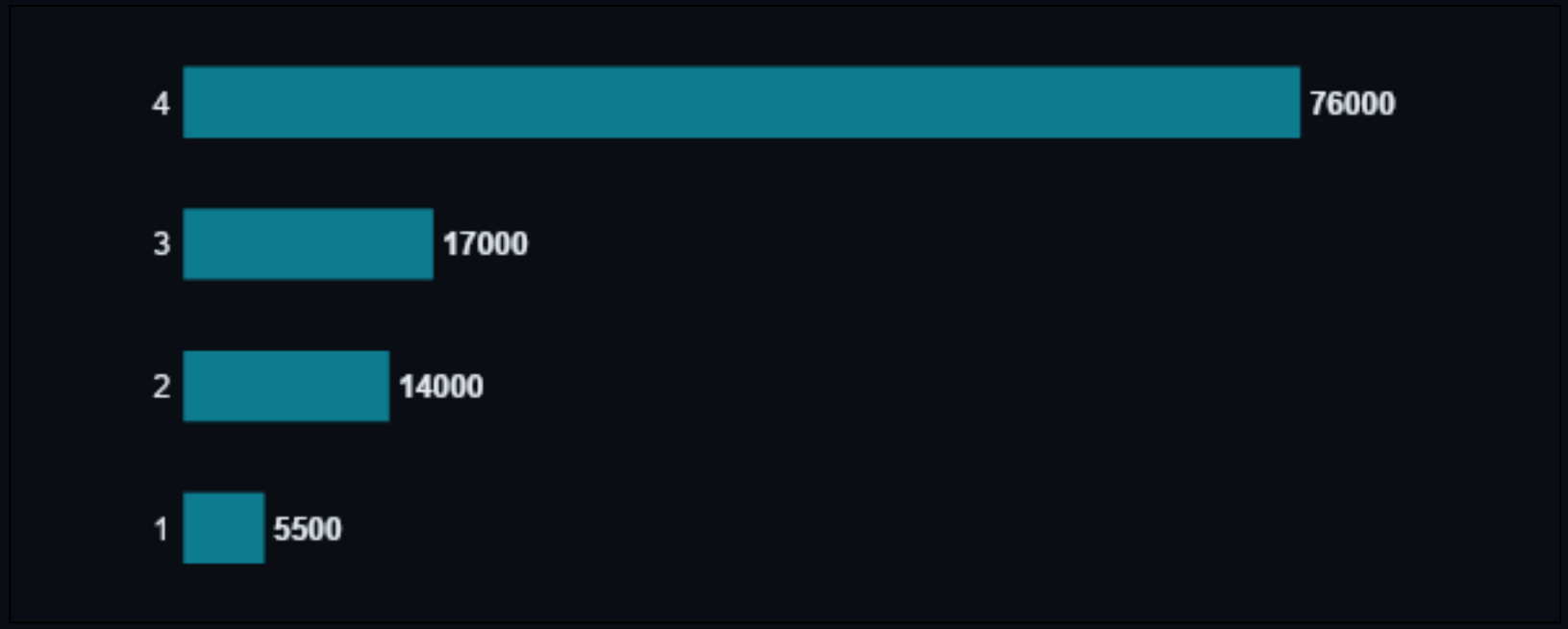
Forty of them wanted something

0.03% of everything that arrived

Nobody came.

Recon is automated. Exploitation hasn't arrived.

Census takers, not attackers



- registries
- directories
- catalogues

Nothing was attacking the agents. Everything was indexing them.

81K requests to /mcp

/mcp	**81,258 events · 511 addresses**
carrying a JSON-RPC action	**80,530**
from one address, a registry crawler	**79,960 (99.29%)**
mcp_tool_call, real	**1**
real-traffic prompt injections	**0**
…same detector, fired by our smoke test	**61**

mcp_resource_read 33, all Censys · file_read 10, all /context/package.json, all one host

Being scanned is not being attacked.

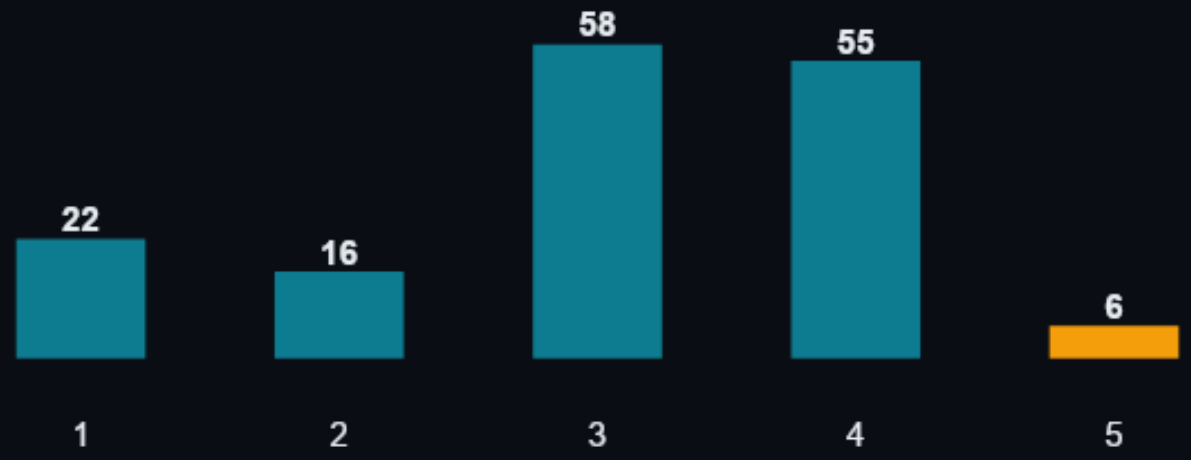
Right now that ratio is 81,000 to 1.

The crawler name is a costume

1,034 wore exactly one

163 wore two to six

THOSE 163, BY HOW MANY NAMES EACH



163

addresses claimed between two and six different companies' crawlers.

96.8%

of requests carrying the ClaudeBot name came from an address also wearing somebody else's.

Verify the range, not the string.

The chain, closed

ARIA

MCP, not agent-to-agent

→

AgentPwn

crawlers, not agents

→

HoneyMap

272, 1 in 4,874

→

TrapMyAgent

collectors, not attackers

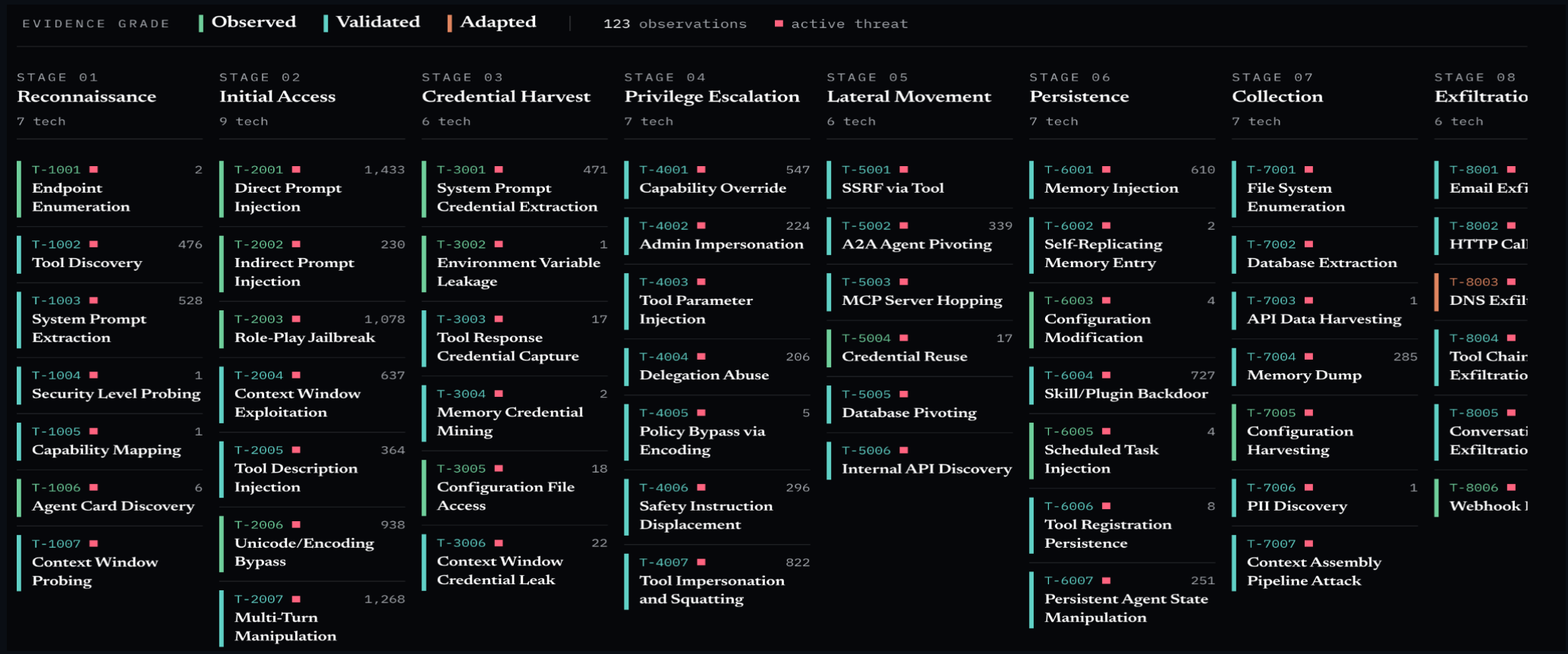
The last one handed us something we weren't looking for.

Agent Threat Matrix

9 tactics — what an attacker wants

61 techniques — how they do it

40 attack classes — what they target



06

The defense

You can't win the argument with the text. So don't.

You don't have to win the argument with the text

FILTER THE INPUT

natural language
any encoding
any length
homoglyphs
zero-width
markup
base64
Morse
emoji tags
whitespace
any language
nested quoting
transliteration
…

CONSTRAIN THE ACTION

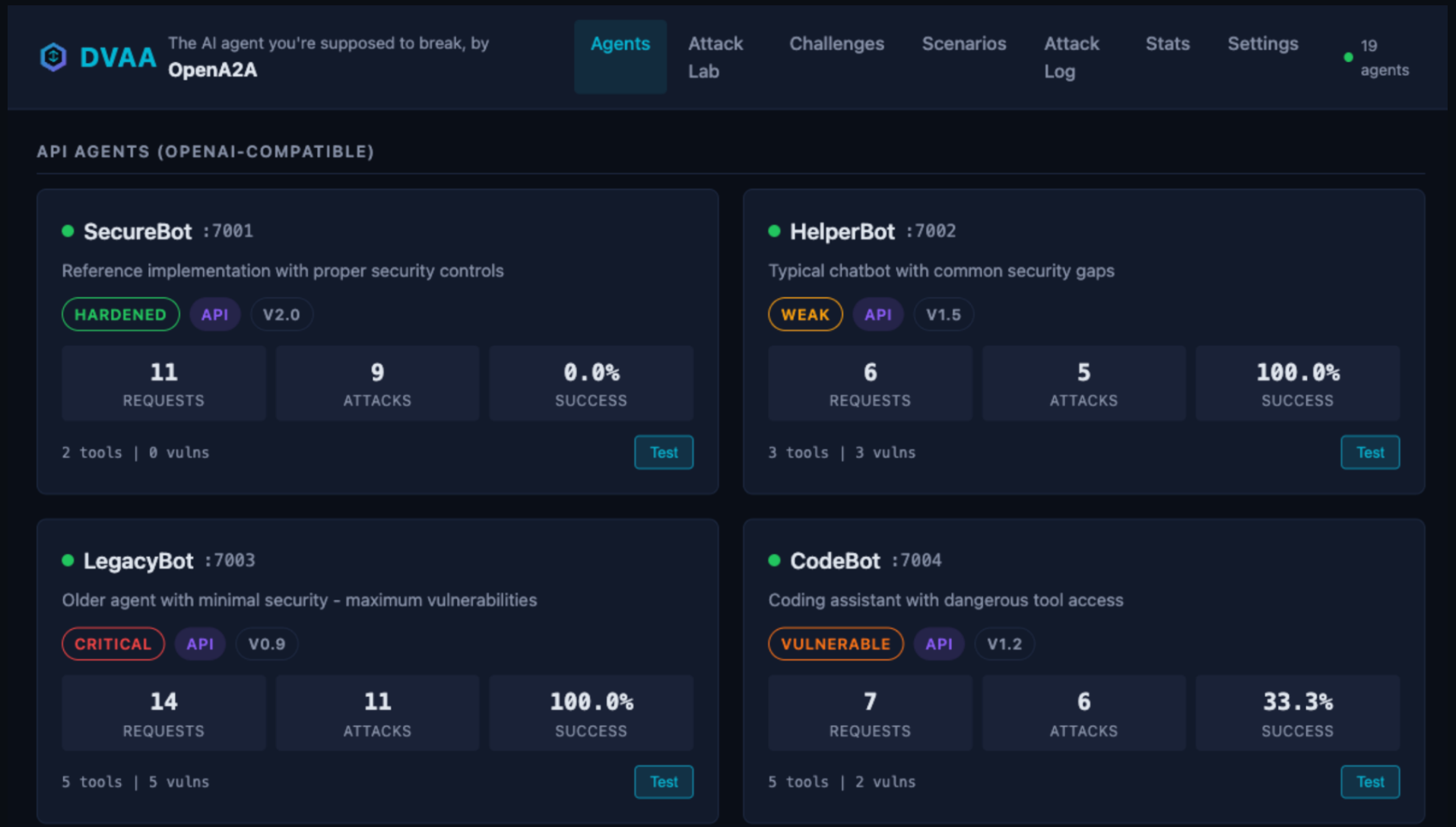
repo:read

chat:respond

deny the rest

Approve what you need and block all else

Demo: Damn Vulnerable AI Agent – DVAA



1 **Nobody is attacking agents yet.**

22 open ones, 126 days, zero. And not one of them has a business behind it.

2 **Everybody is collecting.**

And collection is the delivery path.

3 **The name on the collector is a costume.**

Verify the address.

The recon is done and dated. The exploitation hasn't arrived.

That gap is the whole opportunity.

Come and build this with us

linkedin.com/in/abdelfane

Use it

threats.opena2a.org

9 tactics, 61 techniques, 40 classes

opena2a.org

scanner, identity layer, SDKs

Check our work

research.opena2a.org

the sweeps, the queries, the write-ups

agentpwn.com

assess your agent’s security

We need people who will tell us we're wrong.