FUZZING LABS

No need to be a Mythos to do offensive security

Le Hack 2026 - Keynote - Patrick Ventuzelo

Who am I

PATRICK VENTUZELO

- Founder & CEO of FuzzingLabs
- 10+ years in offensive research, fuzzing, and automation
- Speaker & trainer at Black Hat US & EU, REcon, OffensiveCon, PoC, etc.

FUZZING LABS

What happened **before** Mythos?

From ChatGPT to a classified weapon - 4 years

Nov 2022
OpenAI launches ChatGPT

2022 — 2023

Apr 2023
DARPA announces AIxCC

FUZZING LABS

AIxCC
AI CYBER CHALLENGE

SEMIFINAL COMPETITION
OVERVIEW

DARPA + ARPA-H

To help secure our critical infrastructure, teams created custom CRSs that competed in the AIxCC Semifinal Competition.

42 TEAMS COMPETED

S9 synthetic vulnerabilities

7 TEAMS ADVANCE TO FINALS

5 CHALLENGE PROJECTS

- Linux Kernel
- NGINX
- Tika
- Jenkins
- SQLite

4 hours per round

an AI budget constraint of **$100**

teams had access to

3 nodes each with **64** cores **256** GB RAM

FINALS COMPETITION

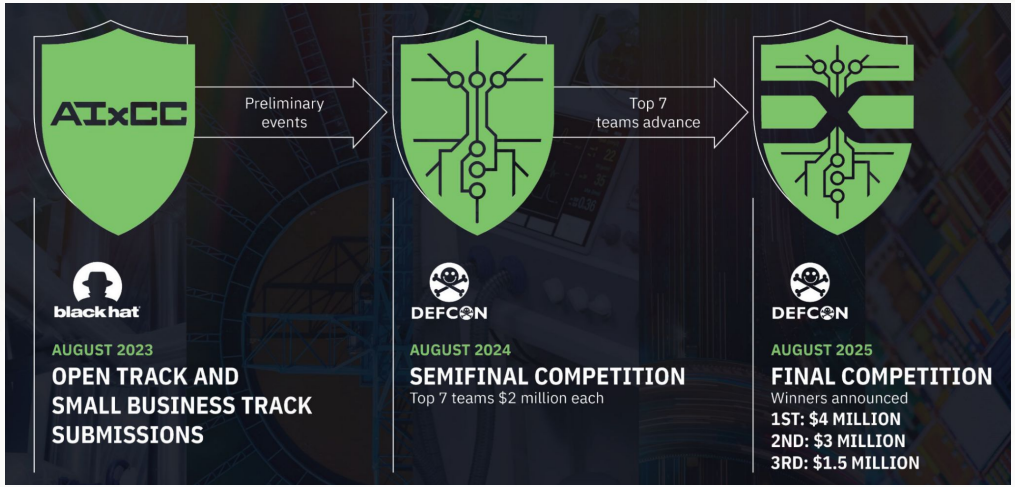
COLLABORATORS & PARTNERS

Google ANTHROP\C OpenAI Microsoft THE LINUX FOUNDATION OpenSSF OPEN SOURCE SECURITY FOUNDATION blackhat DEFCON

DARPA AIxCC — The Real AI Challenge

- **Launched in 2023**, a 2-year challenge to test **AI autonomy in cybersecurity**
- **Teams built agentic systems** to find, exploit, patch, and validate bugs
- Combined **fuzzing**, **SAST**, **and validation** into self-orchestrated pipelines
- Finals at **DEFCON 2025**
  - **$22M** in prizes for **fully autonomous systems**

AIxCC phases - BlackHat 2023 → DEF CON 2025



From ChatGPT to a classified weapon - 4 years

FUZZING LABS

From ChatGPT to a classified weapon - 4 years

FUZZING LABS

April 7, 2026 — Mythos arrives

- "**Thousands of high-severity vulnerabilities**" across every major OS and browser
- FreeBSD remote root under **$50**; Linux kernel privesc under **$2,000**
- Access limited to **11 founding partners + 40 vetted organizations**
- **European not welcomed**

April 7, 2026 - Today we're announcing Project Glasswing[1], a new initiative that brings together Amazon Web Services, Anthropic, Apple, Broadcom, Cisco, CrowdStrike, Google, JPMorganChase, the Linux Foundation, Microsoft, NVIDIA, and Palo Alto Networks in an effort to secure the world's most critical software.

FUZZING LABS

From ChatGPT to a classified weapon - 4 years

FUZZING LABS

June, 2026 - the cutoff

- April 7 — Mythos Preview announced
- **June 9 — Fable 5 publicly released**
- June 12 — Commerce directive received
- **June 13 — all access cut**

- National security directive — no technical justification. **Even Anthropic's own non-US staff lost access.**

4 days.

From released to lockout.

But wait, do we really need **Mythos**?

Mythos on curl - 5 findings became 1

- **178K LOC** scanned on curl master branch via Alpha Omega program
- 5 "confirmed" vulnerabilities reported → **1 confirmed CVE, severity LOW** (3 false positives, 1 just a bug)
- Previous AI tools (AISLE, Zeropath, Codex Security) triggered **200-300 bugfixes** incl ~12 CVEs over 8-10 months
- "I see **no evidence** that this setup finds issues to any particular higher degree than the other tools have done before Mythos." - Daniel Stenberg

  daniel.haxx.se/blog/2026/05/11/mythos-finds-a-curl-vulnerability

CURL AND LIBCURL, SECURITY

MYTHOS FINDS A CURL VULNERABILITY

MAY 11, 2026 DANIEL STENBERG 33 COMMENTS

yes, as in singular *one*.

FUZZING LABS

Public models reproduce Mythos findings

- "The moat is **moving up the stack** — from model access to validation, prioritization, and remediation."
- https://blog.vidocsecurity.com/blog/we-reproduced-anthropics-mythos-findings-with-public-models
- **Cost stayed under $30 per file scanned.**

Target	Claude Opus 4.6	GPT-5.4
FreeBSD NFS (CVE-2026-4747)	Reproduced (3/3)	Reproduced (3/3)
OpenBSD TCP SACK (27-year-old)	Reproduced (3/3)	Not reproduced (0/3)
FFmpeg H.264 parser	Useful lead	Useful lead
Botan (CVE-2026-34580/34582)	Reproduced (3/3)	Reproduced (3/3)
wolfSSL (CVE-2026-5194)	Useful lead	Useful lead

Small models, repeated, beat large models on cost

- For the price of 1 Opus scan, run **Flash Lite 21×**.
- "If a large model finds with 90% probability and small with 50% but **10× cheaper, use small**."
- https://www.hacktron.ai/blog/why-mythos-doesnt-matter-for-us/

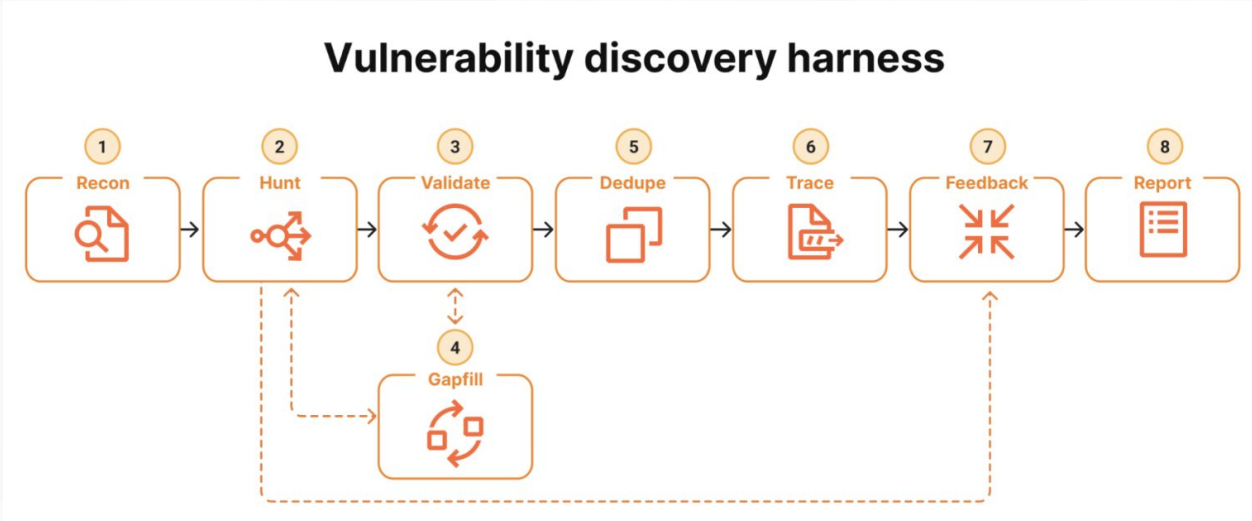
Benchmark Results

Model	Finding A	Finding B	Avg Findings	Avg Cost	Flash Factor
claude-opus-4-6	7/7	5/7	~260	~$79	21.3x
claude-sonnet-4-6	5/6	2/6	~1120	~$122	32.9x
claude-haiku-4-5	0/10	3/10	~180	~$7.3	1.9x
gpt-5.4	7/10	4/10	~30	~$12	3.2x
gpt-5.4-mini	7/10	6/10	~490	~$10	2.7x
gpt-5.4-nano	1/10	3/10	~78	~$2.8	0.75x
gemini-3.1-pro	6/8	6/8	~390	~$55	14.8x
gemini-3.1-flash	9/10	10/10	~264	~$3.7	-

FUZZING LABS

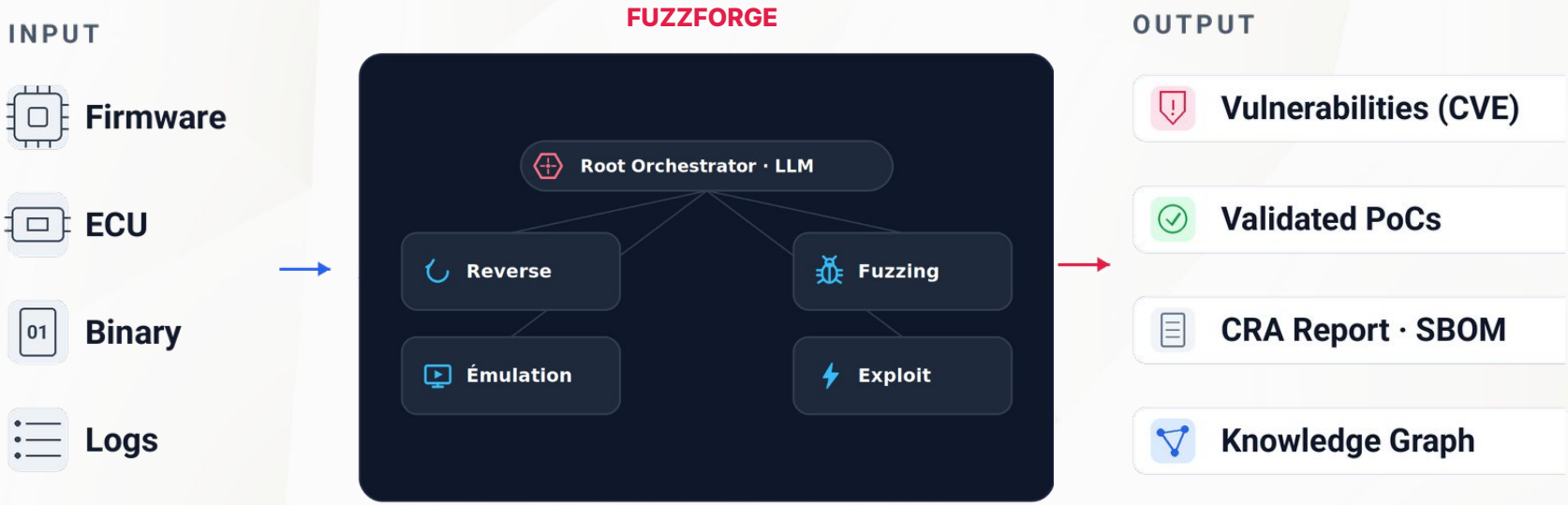
Cloudflare tested Mythos on 50+ repos

- Mythos pointed at **50+ Cloudflare repositories** through Project Glasswing
- What Mythos does uniquely well: **exploit chain construction, proof generation**
- "When we ran other frontier models through the same harness, they found a fair number of the **same underlying bugs**." — Cloudflare Security team
- "Patching faster **does not change the shape** of the pipeline that produces the patch."
- blog.cloudflare.com/cyber-frontier-models



FUZZING LABS

FuzzForge — our agentic validation system



Finding	Category	Location	Severity	Confidence
Root Command Injection in Diagnostic CGI Handler The diagnostic ping endpoint concatenates the host form parameter into a shell	vulnerability	diag.cgi	CRITICAL	High

Ok, but **How** to build it myself?

What you need to build a Mythos-equivalent

MEMORY
Knowledge base, RAG, findings DB

TOOLS
MCP servers, fuzzers, disassemblers

MODELS
Open-weight LLMs (Gemma, Qwen, Hermes)

HARDWARE
DGX Spark, Mac Studio, Ryzen AI

ORCHESTRATION
agent framework wraps it all

Hardware is no longer the excuse

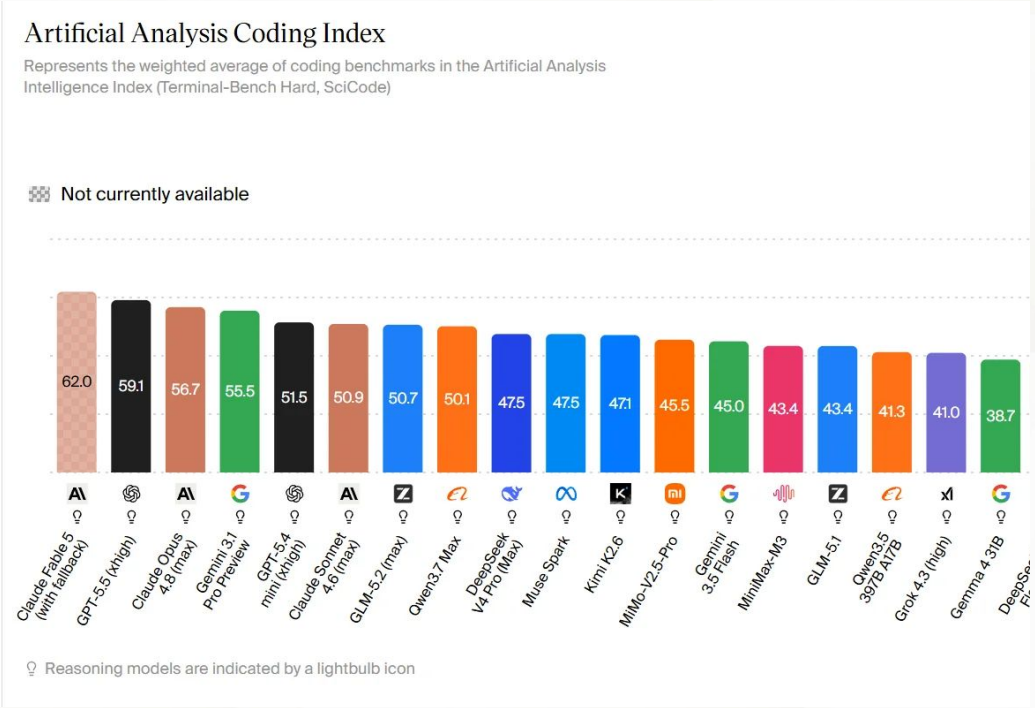
- **NVIDIA DGX Spark** — **$4K**, 128GB unified memory, models up to 200B params; 2 units via ConnectX → 405B
- **Mac Studio M3 Ultra** — up to **512GB** unified memory, runs 600B+ MoE models, best $/GB
- **AMD Ryzen AI Max+ 395** — from **~$2K**, up to 96GB unified, mini-PC form factor (the cheap entry point)
- What cost **$200K/year** of compute in 2024 fits in a <€10K workstation in 2026

The open-weight models caught up

- **Gemma 4 31B** — 6.6% → **86.4%** on τ2-bench Retail (agentic tool use), **+1200% vs Gemma 3**
- **Qwen 3.6 Plus** — **beats Claude Opus 4.5** on Terminal-Bench 2.0 (61.6 vs 59.3)
- **GLM 5.2**
- **Deepseek**
- **Minimax**

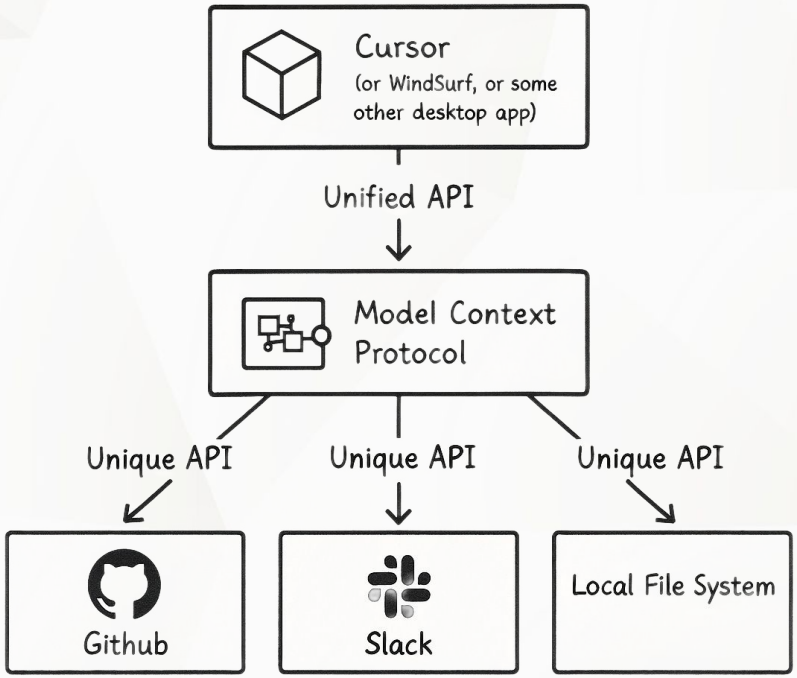
- **Heretic / uncensored models — no guardrails for offensive research**

- All fine-tunable on your own data with **Unsloth** or Axolotl



MCP - the standard that won

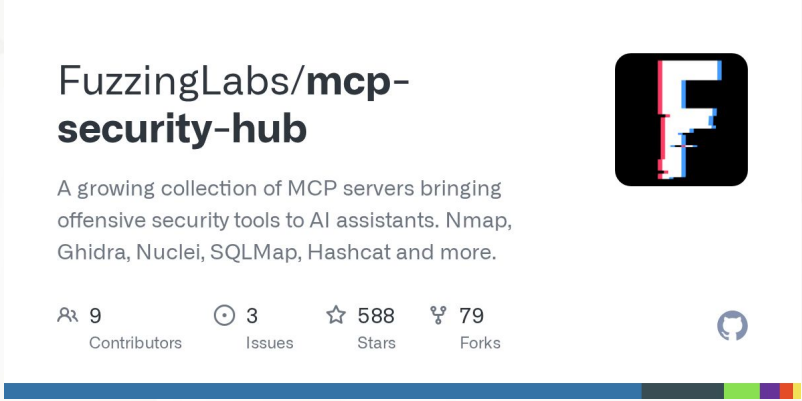
- Introduced by Anthropic Nov 2024; now adopted by OpenAI, Google, Microsoft
- Python + TypeScript SDKs: **~97M downloads/month**; **~2,000 servers** in registry
- Donated to Linux Foundation (Dec 2025) — **vendor-neutral standard**
- Offensive MCP servers: WireMCP (Wireshark), Ghidra-MCP, RedTeam-MCP
- The integration layer that connects models to fuzzers, disassemblers, debuggers



FUZZING LABS

Introducing MCP Security Hub

- Curated catalog of MCP servers for offensive security — automated security testing, reference workflows, all open source
- github.com/FuzzingLabs/mcp-security-hub

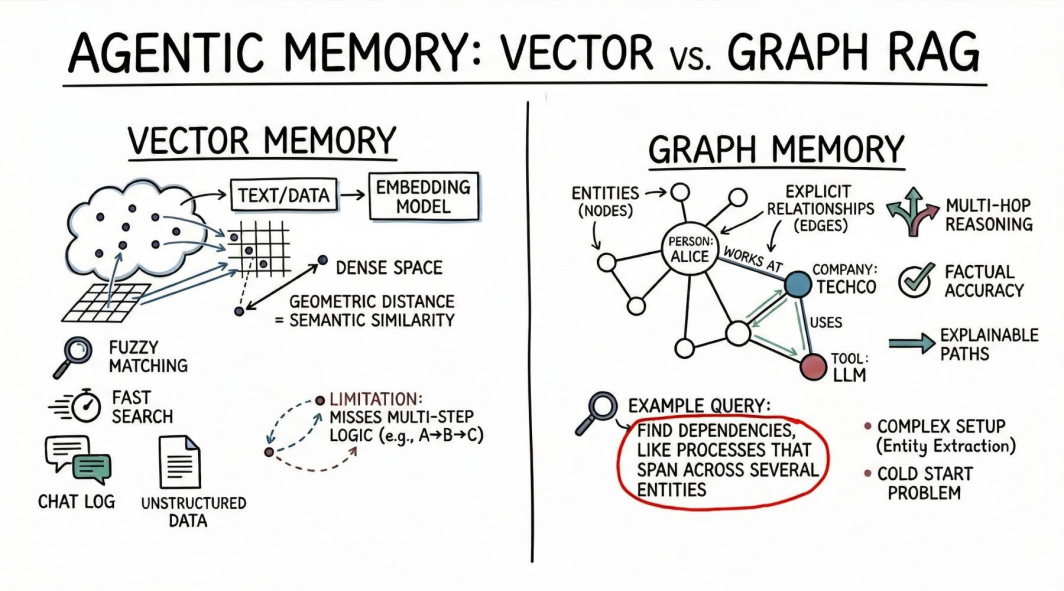


Features

- **38 MCP Servers** covering reconnaissance, web security, binary analysis, blockchain security, cloud security, code security, secrets detection, threat intelligence, OSINT, Active Directory, fuzzing, and more
- **300+ Security Tools** accessible via natural language through Claude or other MCP clients
- **Production Hardened** - Non-root containers, minimal images, Trivy-scanned
- **Docker Compose** orchestration for multi-tool workflows
- **CI/CD Ready** with GitHub Actions for automated builds and security scanning

Memory - the component everyone forgets

- Agents query a **shared knowledge base** of past findings, CVEs, tool outputs, user notes
- RAG context turns a generic model into a **domain specialist** (Rust, Android, firmware...)
- Reduces duplication, guides agent actions, improves task relevance over time
- The difference between a one-shot scan and **a system that learns**



FUZZING LABS

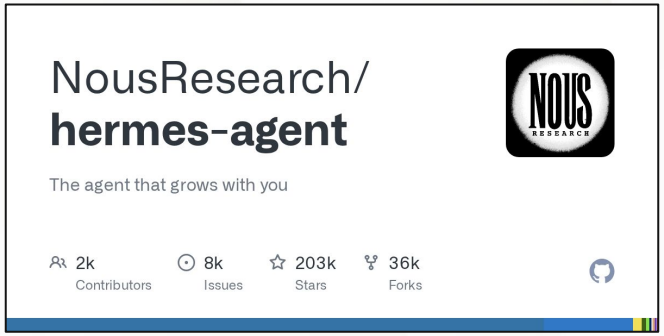
Pick your agent framework

OpenClaw

Self-hosted local agent, MCP-compatible, skills marketplace. Explosive growth, 470+ advisories in 6 months

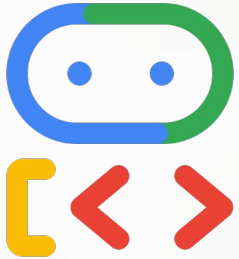
Hermes Agent

Self-improving, closed learning loop (DSPy + GEPA), MCP-first, 200k+ stars



Google ADK

Code-first multi-agent, Python/Go/Java/TS, model-agnostic, enterprise-grade



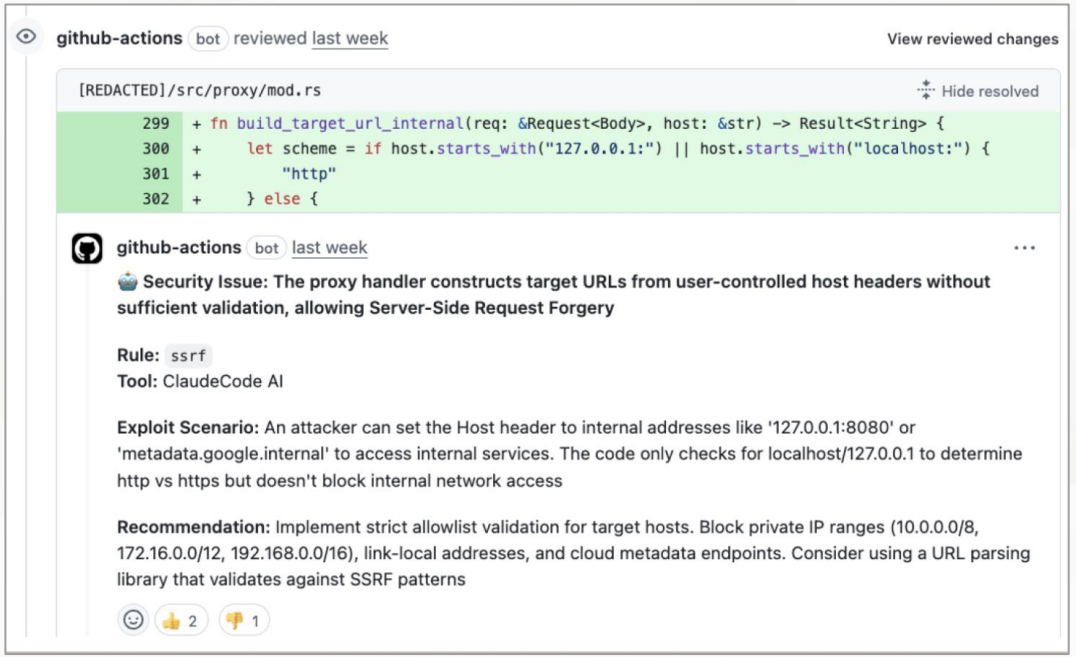
Fine, **What** should I do with it?

From Pattern Matching to Reasoning

- **LLM-based SAST**
  - Analyzes **ASTs, not just regex** patterns
- **Rule synthesis**
  - Infers vulnerability patterns automatically
- Production tools: Claude Code Review, Semgrep AI, AISLE, Zeropath, Codex Security

- Stenberg / curl: **200-300 bugfixes** from these tools in 8-10 months, ~12 CVEs published

- **Real-world adoption:**
  - Claude Security Review
  - Semgrep AI

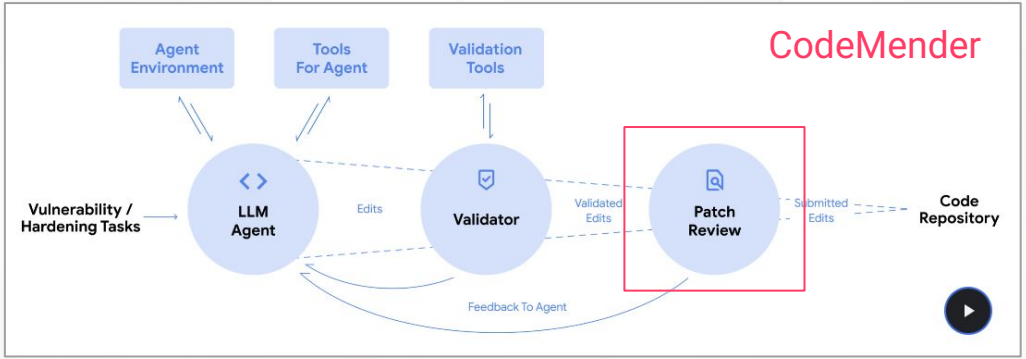


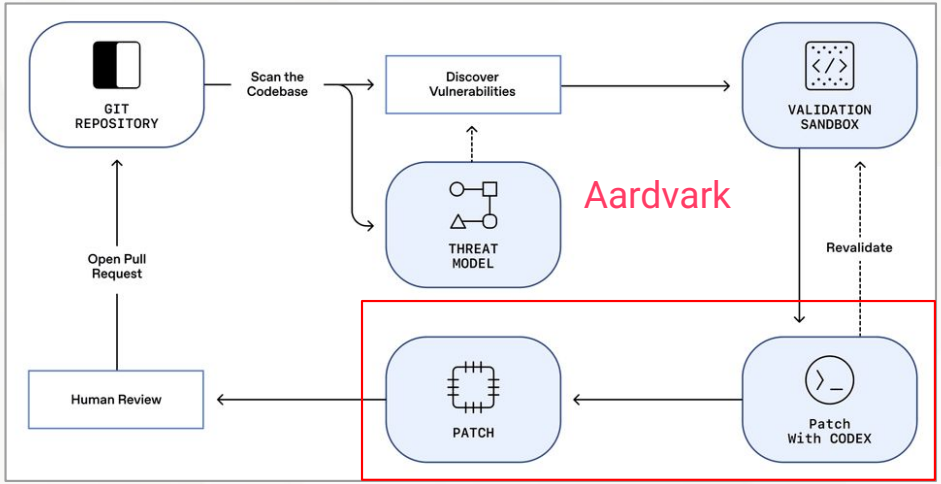
Claude Code Review - GitHub

FUZZING LABS

From harness synthesis to exploit generation

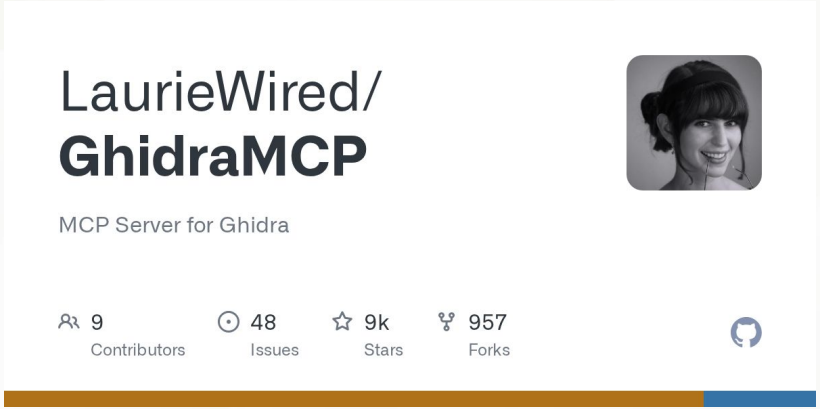
- **Harness synthesis**
  - Auto-generate fuzz entrypoints from source or APIs
- **LLM patching**
  - Generate candidate fixes from exploit traces
- **Automated validation**
  - Re-test PoC for functional correctness
- **Continuous Feedback**
  - Each validated patch improves next iterations
- **Examples:**
  - CodeMender
  - OpenAI Aardvark





Agentic reverse engineering

- **Function annotation**
  - Automatic naming of stripped binaries from behavioral analysis
- **Crypto & protocol detection**
  - Identify patterns, structures, vtables, custom protocols
- **Hypothesis generation**
  - Agents propose attack surface & exploitation paths — humans validate
- **Tools: Ghidra-MCP, Binary Ninja + LLM, Idasql**



And **WHY** should I make the effort

What still doesn't work - and that's OK

1

LLM non-determinism

same prompt, different outputs, even at temperature=0

2

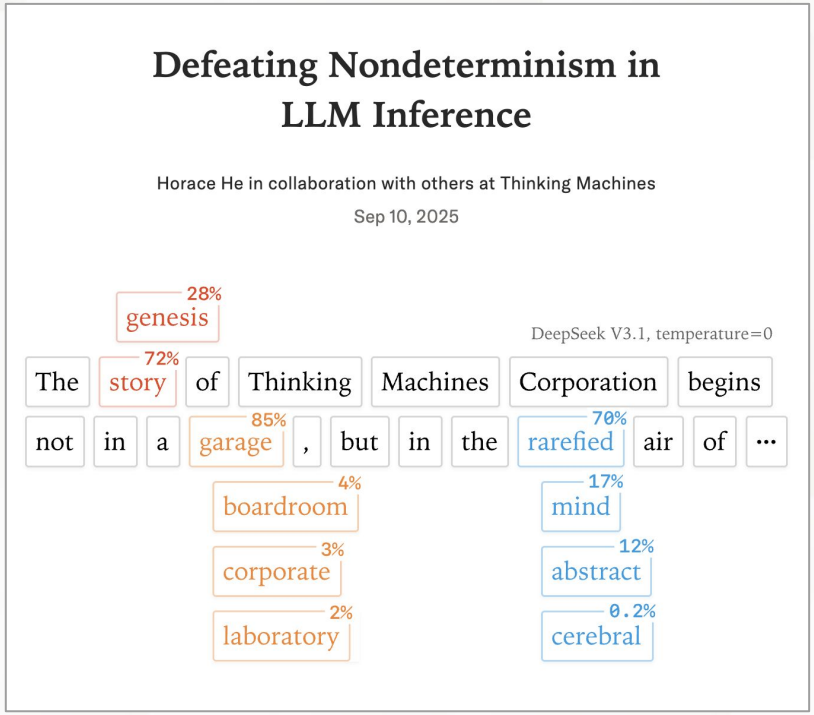
Benchmarks are immature

CVE-Bench, CAIBench, XBOW — useful signals, not gospel

3

Hallucinations persist

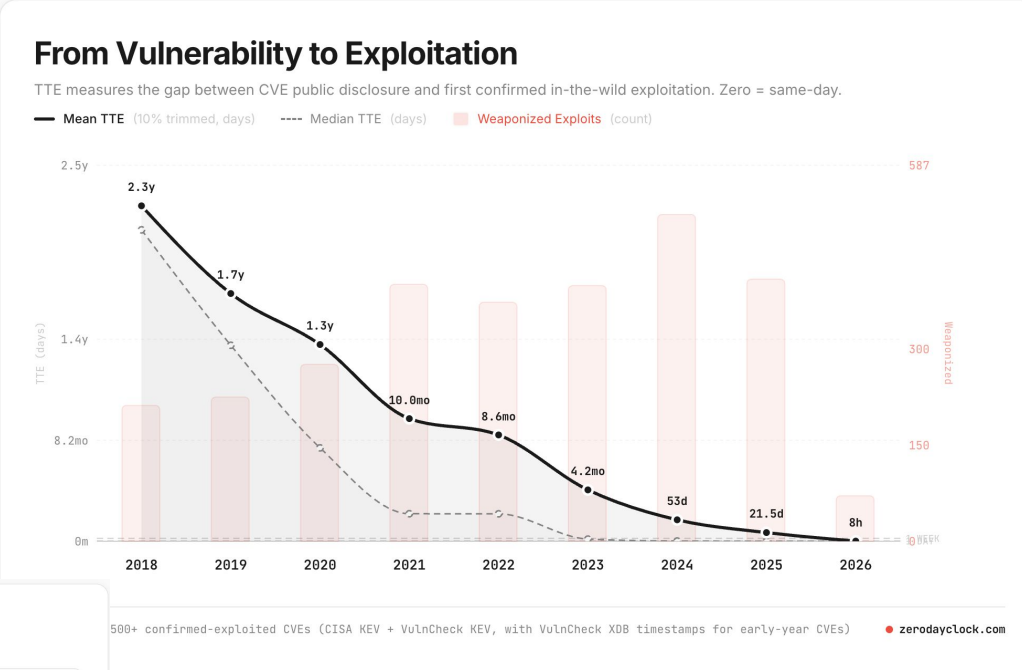
~26.5% fabrication rate on Qwen 3.6 Plus for complex code reasoning

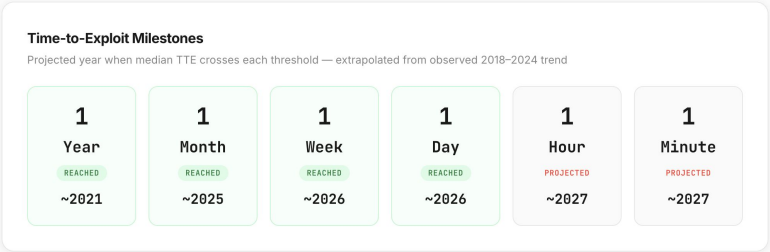


FUZZING LABS

From vulnerability to exploitation

- Mean TTE went from **2.3 years (2018) to 8 hours (2026)**
- **73.2%** of CVEs are now exploited BEFORE public disclosure (vs 31% five years ago)
- **67.2%** of exploited CVEs in 2026 are zero-days (vs 16.1% in 2018)
- Weaponized exploits per week trending up — no plateau yet
- https://zerodayclock.com/





Whatever side you're on, you need this

If you defend

- **Attackers already have it** (Hexstrike on the darkweb, day one)
- The exploitation window is now ~1 day
  - you must **match their speed**
- AI-augmented triage is the only way to keep up with the CVE flood
- Waiting = falling behind by default

If you attack

- This is **your edge** — scale what a single human can't
- Reproduce n-days in **minutes, not days**
- Cover more attack surface per engagement
- The data flywheel: each analysis makes the next faster and better

FUZZING LABS

Where to start — Monday morning

1

Clone an agent framework

OpenClaw or Hermes Agent — running in minutes

2

Add a model

Gemma 4 or Qwen 3.6, local on your workstation

3

Plug in MCP tools

WireMCP, Ghidra-MCP, or the MCP Security Hub

4

Point it at a target

Start with an n-day: a CVE + its commit fix

FUZZING LABS

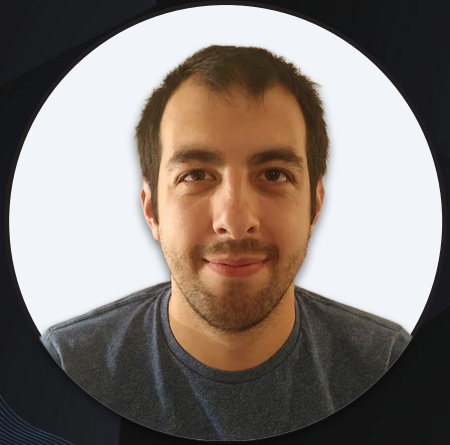
You don't have to be a Mythos.

You just have to keep building.

FUZZING LABS

Let's Connect!

Talk: No need to be a Mythos to do offensive security



Patrick Ventuzelo
Fondateur & CEO of FuzzingLabs

patrick@fuzzinglabs.com

fuzzinglabs.com