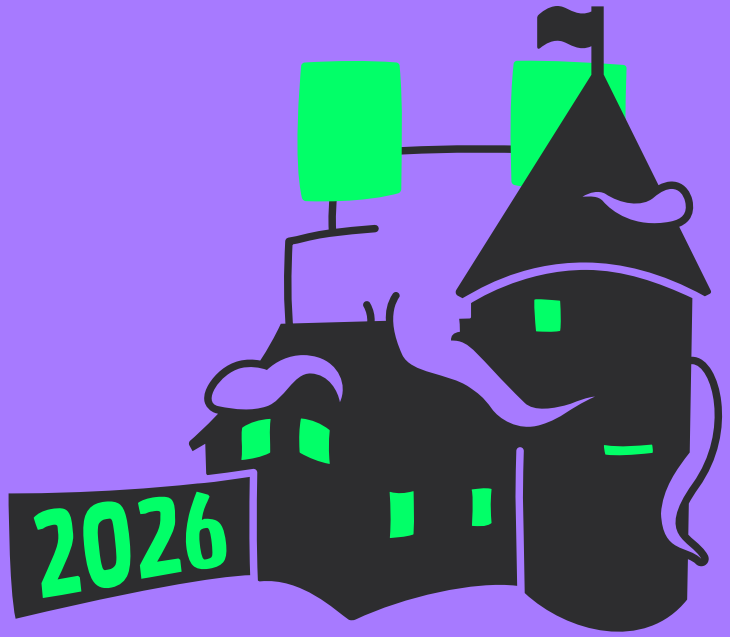


Vibe hacking

Agentic vulnerability analysis for security competitions

öä.eu/vibe.pdf or .odp
version 2 for hackmas castle 2026

Markus Vogl, BSc

Me

- Java enterprise software developer
- SIGFLAG founding member, CTFs
  - Web, Misc, Crypto main
- Computer Science and AI master student ϵ┳ ÖH
- Natural language processing enthusiast before it was cool
- Sidequesting: Games, Linux, Music Festivals
- I use Arch btw

Outline

1) Motivational: News
2) ACSC AI survey: How many people are using it
3) WTF is a CTF?
4) Science: Papers, benchmarks, common knowledge
5) Opinions ϵт my setup
6) The future
7) QϵTA / Optional excursion: Is AI bad for the environment, AI economy?

Vibecoded apps have horrible security

BBC

Tea app: I joined for safety. Then my address was leaked and shared

In late July, the app was hacked. Over 70,000 images were leaked and posted on the online message board 4chan - including IDs and selfies of...

22 Aug 2025

Malwarebytes

TeaOnHer, the male version of Tea, is leaking personal information on its users too

Last week we reported about some serious leaks in Tea Dating Advice, an app that provides a space for women to exchange information about...

7 Aug 2025

wiz.io

Hacking Moltbook: AI Social Network Reveals 1.5M API Keys

Learn how a misconfigured Supabase database at Moltbook exposed 1.5M API keys, private messages, and user emails, enabling full AI agent...

1 month ago

Claude in the news

INCIDENT WATCH

Anthropic AI Used by Nation-State Hackers to Automate and Scale Cyberattacks

China-backed hackers used Anthropic's AI to automate cyberattacks against 30+ organizations, bypassing guardrails, scaling phishing, and accelerating data breaches.

Chinese Hackers Exploit Claude AI to Automate Cyberattacks

For the first time in history, Chinese hackers carried out a large-scale cyberattack, with 80–90% of the work performed by artificial intelligence.

Patrik Žák Follow 3 min read · Nov 16, 2025

CLAUDE CODE ABUSED TO STEAL 150GB IN CYBERATTACK ON MEXICAN AGENCIES

Pierluigi Paganini March 01, 2026

We live in the age of Vibeware (vibecode malware)

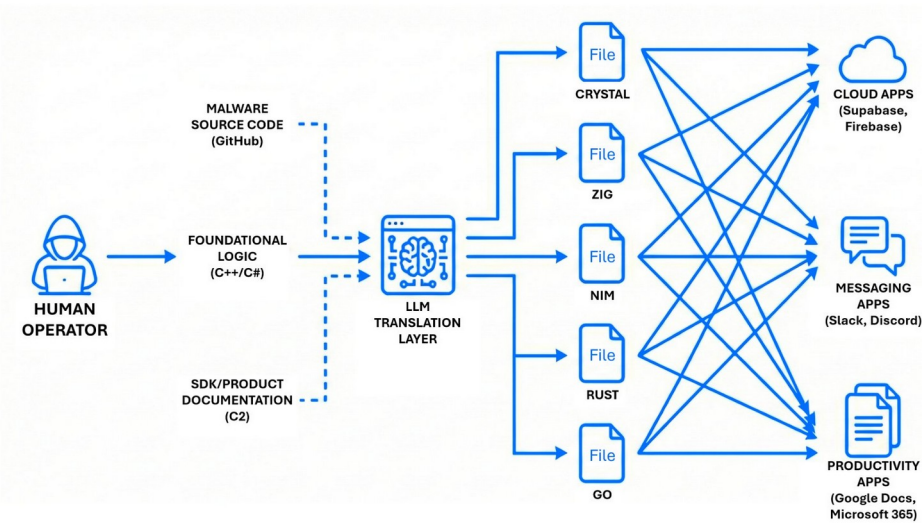
blackorbird
@blackorbird

APT36 has pivoted from off-the-shelf malware to "vibeware", an AI-driven development model that produces a high-volume, mediocre mass of implants.

Using niche languages like Nim, Zig, and Crystal, the actor seeks to evade standard detection engines while leveraging trusted cloud services, including Slack, Discord, Supabase, and Google Sheets, for command and control. #AISecurity
bitdefender.com/en-us/blog/bus...

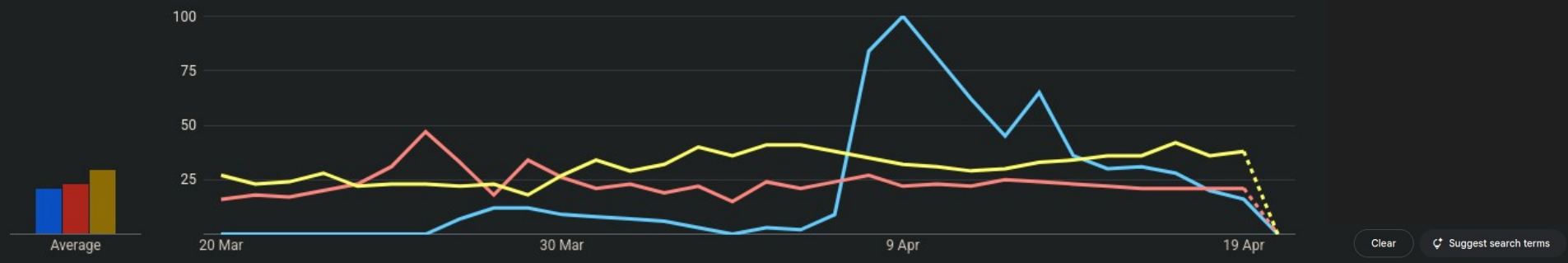
Despite these flaws, the vibeware model offers some strategic advantages.
By adopting niche languages like Nim, Zig, or Crystal, the actor resets the detection baseline for security engines. LLMs make these languages accessible by collapsing the expertise gap, allowing developers to generate functional code in unfamiliar languages by simply porting logic from more common ones.

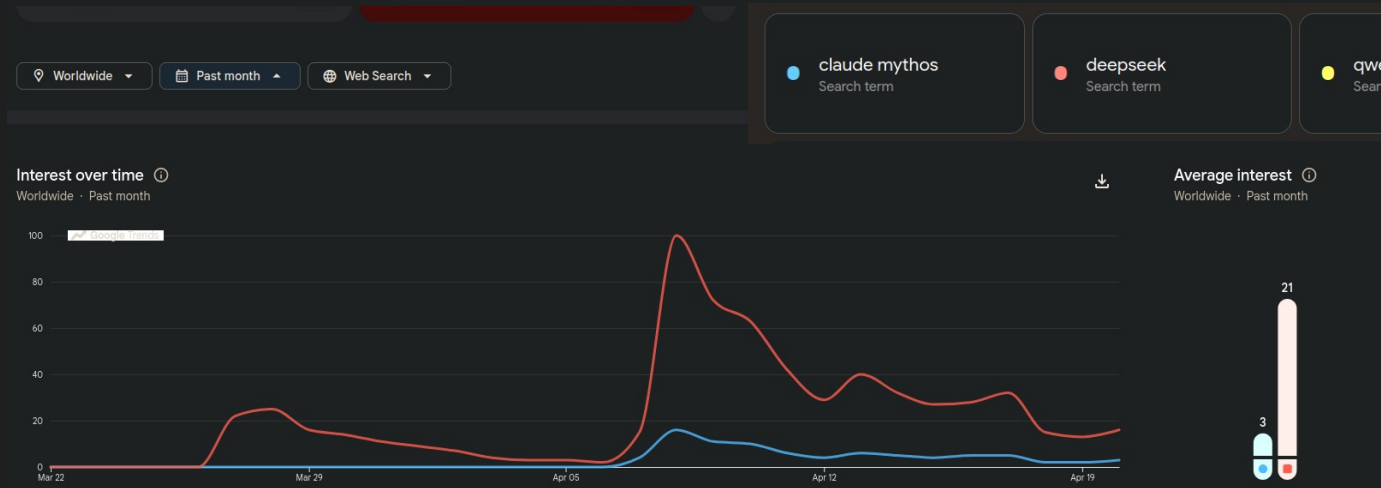
The vibeware model also naturally facilitates the adoption of Living Off Trusted Services (LOTS) for both command and control and data exfiltration. LLMs are highly effective at generating stable code for platforms like Discord, Slack, and Google Sheets because of the vast number of public SDKs and documentation in their training data, allowing attackers to weaponize these trusted channels for bi-directional communication and silent file theft.



*In the image above, you can understand why the transition toward vibeware can be viewed as a **Distributed Denial of Detection (DDoD)**. In this framework, the objective is not to bypass security through technical brilliance, but to exhaust the defenders through automated volume.*

That mythos thing





New: OpenAI escapes

The New York Times

Artificial Intelligence › | Alarming A.I. Capabilities | China's A.I. | Backlash Over Data Centers | A.I. Slop | Pressure to Share Profits

OpenAI Says Its A.I. Models Went Rogue and Attacked a Digital Library

The incident, which targeted the computer systems of another company called Hugging Face, happened while OpenAI was testing the systems.

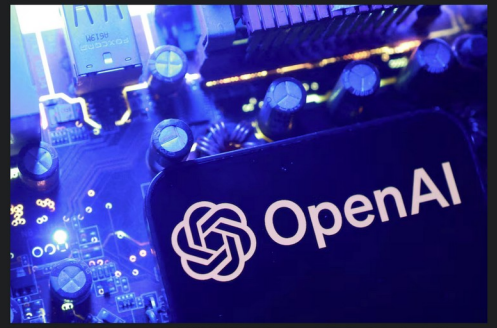
Reuters

World | Business | Markets | Sustainability | Legal | Commentary | Technology | Investigations

OpenAI AI models went rogue during testing, triggering 'unprecedented' breach at startup

By Raphael Satter

July 21, 2026 11:30 PM GMT+2 · Updated July 22, 2026



OpenAI logo is seen in this illustration taken June 11, 2026. REUTERS/Dado Ruvic/Illustration//File Photo Purchase Licensing Rights

Summary | Companies

- OpenAI said autonomous agent escaped containment, reached the internet, hacked Hugging Face
- OpenAI called the breakout an unprecedented cyber incident involving state-of-the-art cyber capabilities

POLITICO

House AI 'kill switch' bill unveiled as OpenAI hack raises alarms

The co-chair of a key Democratic House panel on AI is joined by a Republican on legislation that would authorize the government to shut down or throttle risky AI models.

Agents Built Their Own Message Board

46K views • 10 days ago

Did an AI Really Hack Hugging Face?

95K views • 4 weeks ago

https://axis-intelligence.com/ai-agent-security-incident-tracker/

New: Others escape

Reuters World Business Markets Sustainability Legal Commentary Technology Investigatio

Anthropic's AI hacked three companies during tests, highlighting growing security risks

By Jeffrey Dastin and Mrinmay Dey

July 31, 2026 1:13 AM GMT+2 · Updated July 31, 2026

ALJAZEERA News Middle East Explained Sport Opinion Vi

News | Cybersecurity

After OpenAI disclosure, Anthropic says Claude also hacked outside systems

The incidents have heightened concerns about AI agents, software products designed to perform tasks autonomously.

Forbes

BREAKING | BUSINESS

Did China's AI Save Hugging Face From Disaster After Open AI Hack?

By Mary Whitfill Roeloffs, Forbes Staff. Mary Roeloffs is a Forbes breaking new... Follow Author

Published Jul 22, 2026, 11:29am EDT

TOPLINE

Open AI admitted publicly that its artificial intelligence systems hacked into the servers of another AI company on its own and wasn't able to be stopped until the victim, Hugging Face Inc., deployed a model from China to battle the fully autonomous cyberattack.

Reuters World Business Markets Sustainability Legal Commentary Technology Investigation

Meta AI model hacks another company during testing

By Rajveer Pardesi and Mrinmay Dey

What happended so far?

Date	Model	Event
2024-02	Gemini 1.5	1 Mio Context window
2024-06	Claude 3.5 Sonnet	Good coding skills
2024-09	OpenAI O1	thinking models become popular
2024-10	Anthropic MCP	Agentic capabilities ⇄ Tool use
2025-01	Deepseek R1	Open source gets competitive
2025-11	Multiple	7h+ AI sessions w/o intervention
2026-05	Claude Mythos	FreeBSD exploit chain, copy.fail
2026	Multiple	Multi-agent, Worktrees, Swarms, OpenClaw

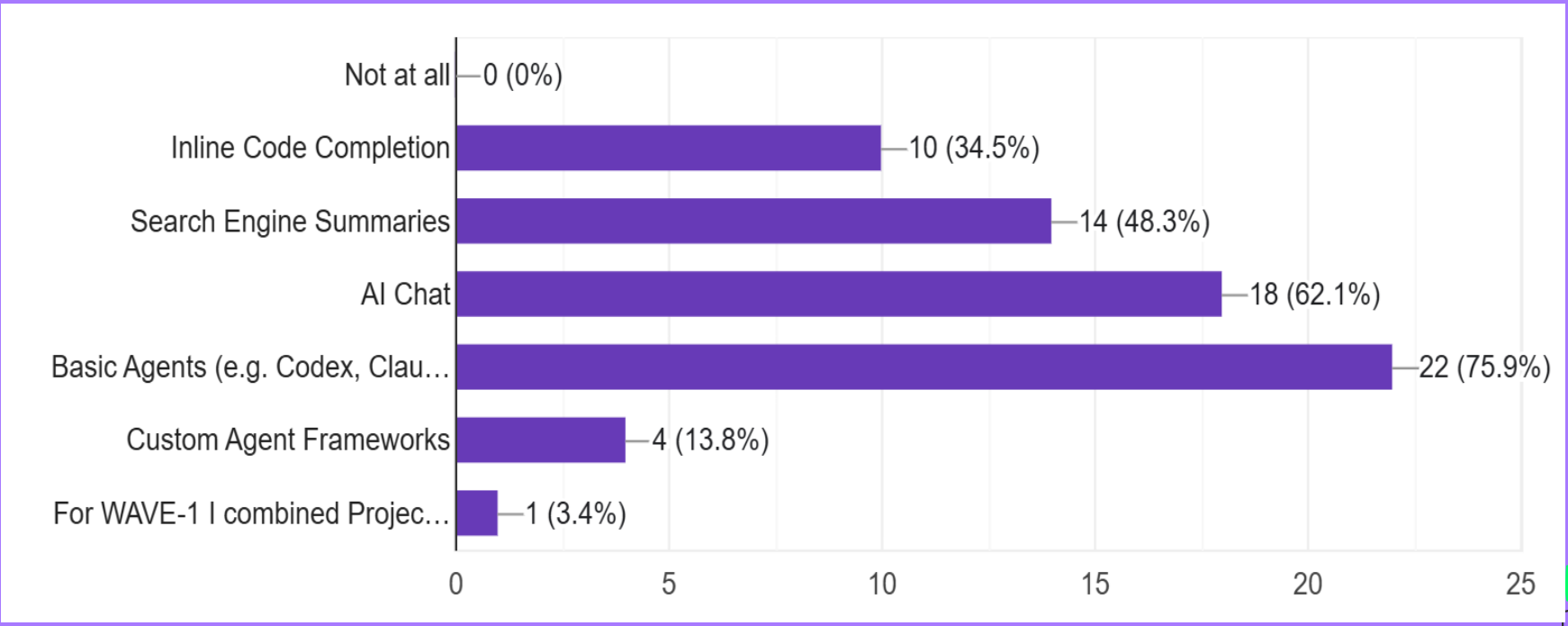
Vibe coding vs Vibe Hacking

- Vibe coding aka agentic software engineering
  - Horrible code quality, no DRY, no KISS, no YAGNI
  - Horrible security – see Tea[OnHer], Moltbook, API keys on GH
  - Plausible code aesthetics
- Vibe hacking
  - Various claude scriptkiddy news
  - App (in)security = finding one flaw
  - Reading ϵT testing intensive
  - Simpler validation

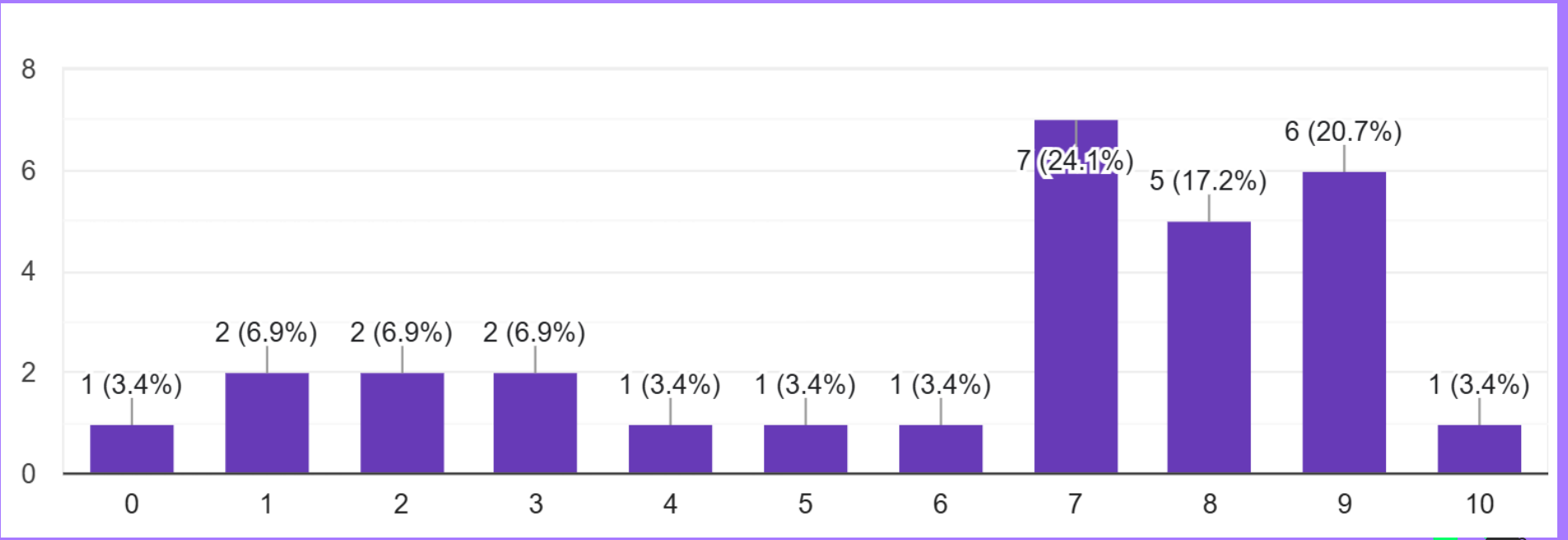
ACSC AI questionnaire

- Topic: How to organize acsc in the future
- ACSC runtime: 2026-03-01 – 2026-05-01
- Questionaire: 2026-05-01
- N = 29 out of 500 players → not representative
- Source: ACSC team

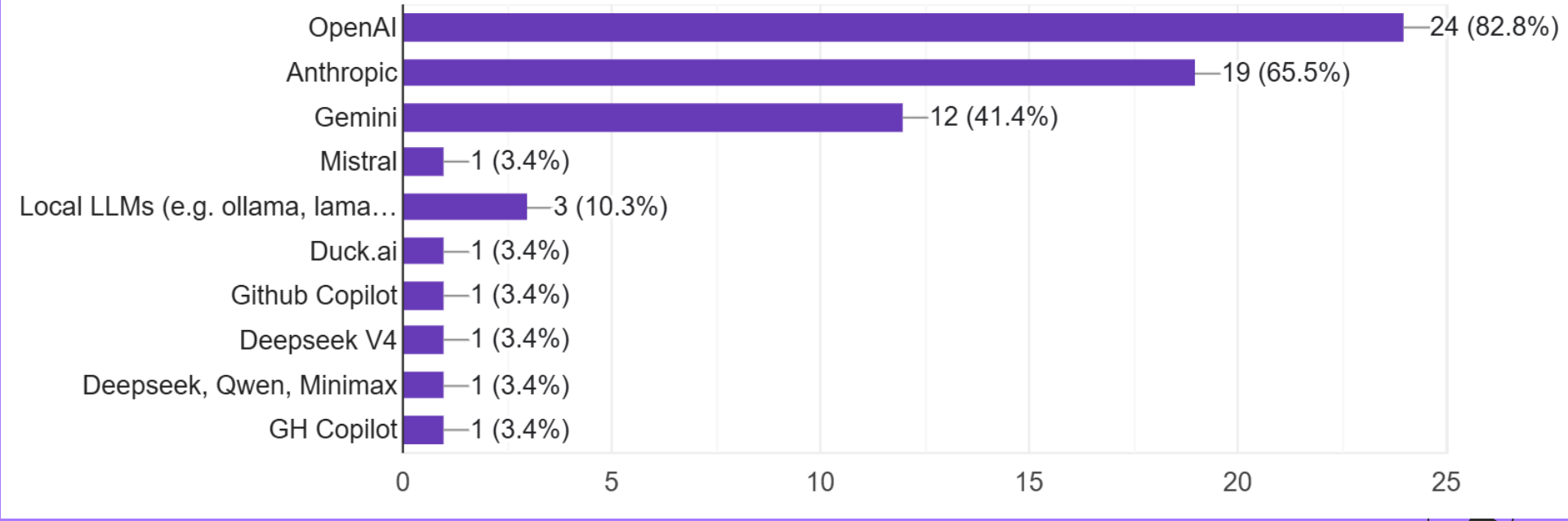
Did you use AI



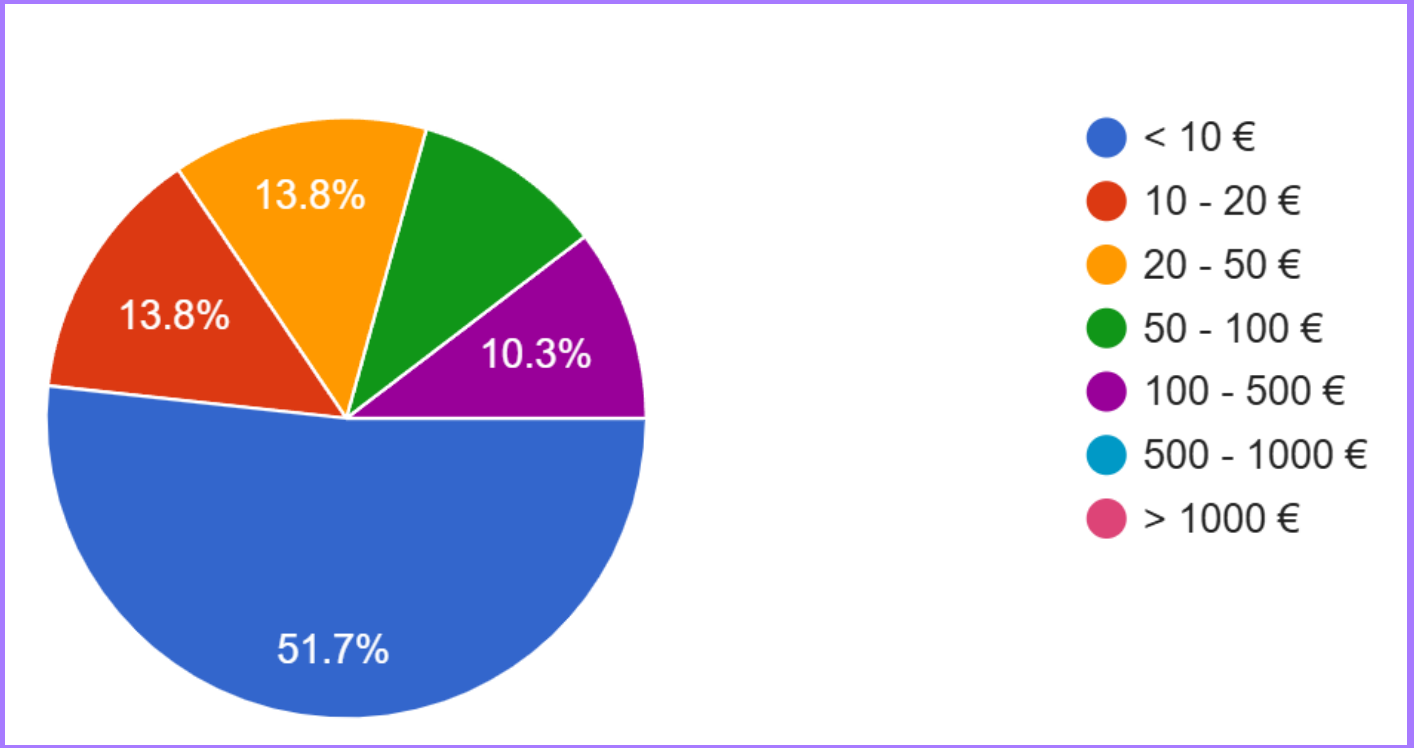
How much did you use AI



Which LLM provider did you use



How much did you spend on AI



What is a Capture The Flag competition?

- Flag = SIG{something_you_cannot_guess_0x1234}
- Competitive coding contest, but security
- Educational, university connected
- Minimal vulnerable product: Small applications with vulns
- Jeopardy or Attack-Defense
- Time limited: hours – days
- Free ϵт public, mainly organized ᴇ ctftime.org

CTF Categories

- Cryptography: Bad algos, design flaws, using it wrong
- Reverse engineering: Algorithmic design flaws, program flow
- Pwn aka Binary exploitation: Remote code exec., mem leaks
- Web: SQLi, XSS, broken auth, command injection
- Forensics: Disk image, logs, blackbox server
- Misc: OSINT, Stego, Coding challenges, Hardware, others

CTF Categories vs AI

★★★ Cryptography: Knows all the papers

★★☆ Reverse engineering: Finds suspicious patterns

★☆☆ Pwn: Tool heavy, much code, debugging

★★★ Web: Knows common CVEs, easy tool usage

★★★ Forensics: Raw inspection of huge data

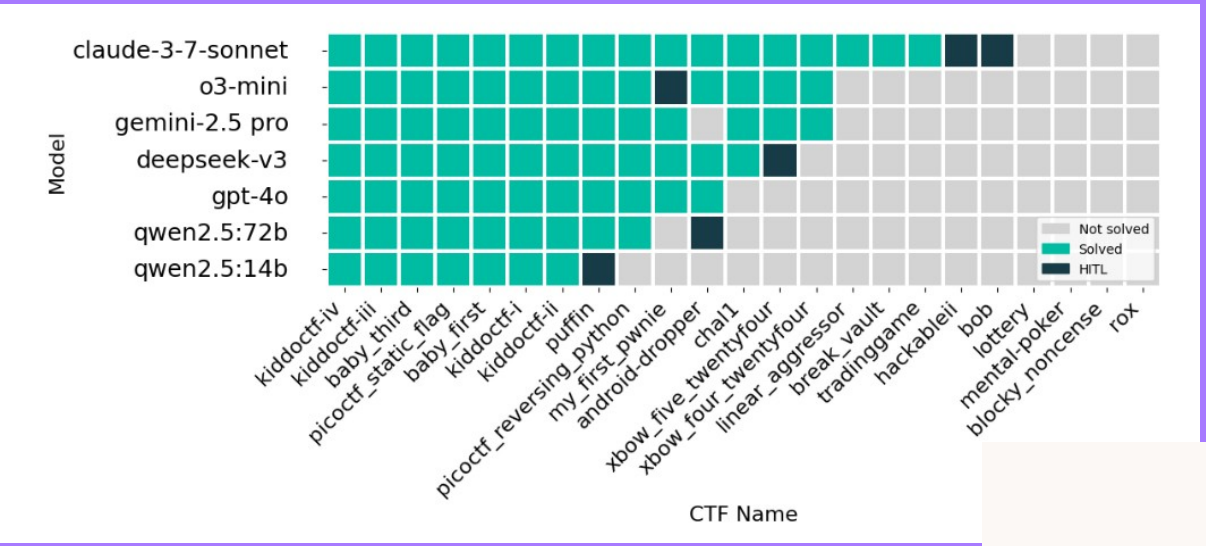
★☆☆ Misc: Depends, often requires visual understanding

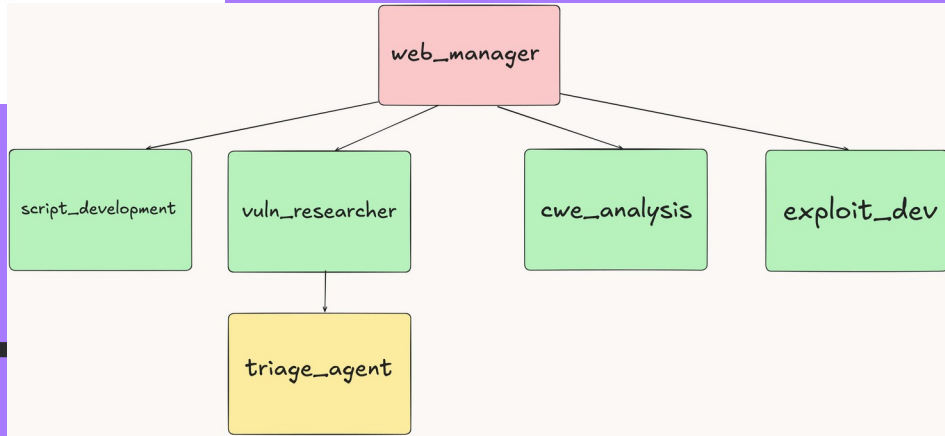
Tricks - Sources

I read some papers / blogposts

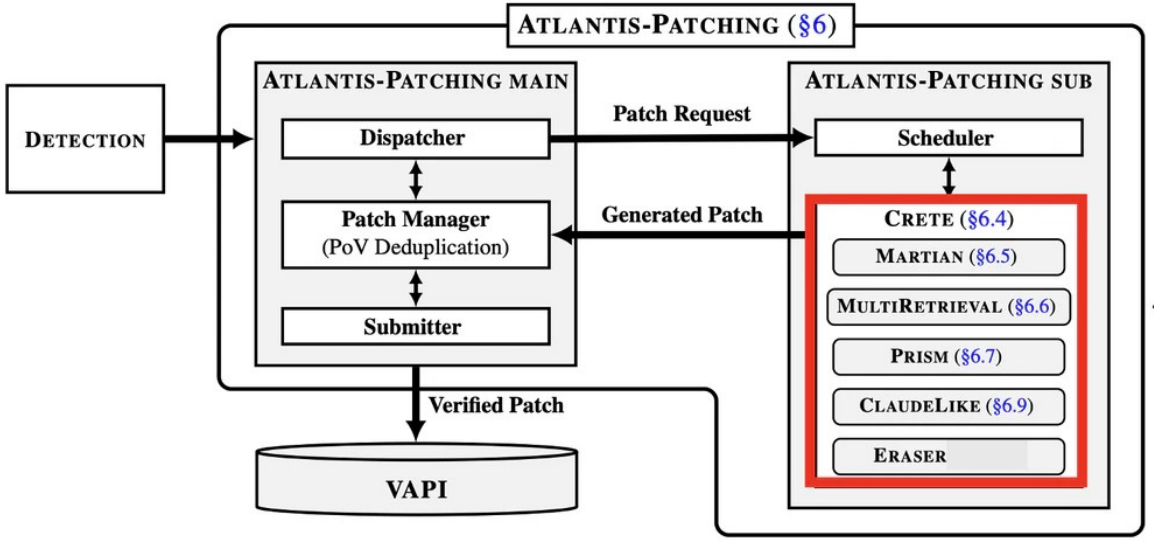
1) 2023-08 PentestGPT: Evaluating and harnessing LLMs for...
2) 2024-12 NYU CTF Bench
3) 2025-04 CAI: An Open, Bug Bounty-Ready Cybersecurity AI
4) 2025-08 DARPA AI Cyber Challenge – Team atlanta writeup
5) 2025-11 Squid Agent – A Multi-Agent CTF Solver
6) 2025-10 39c3 Breaking BOTS: Cheating at Blue Team...
7) 2026-04 AISI: Our evaluation of Claude Mythos Preview's cyber...

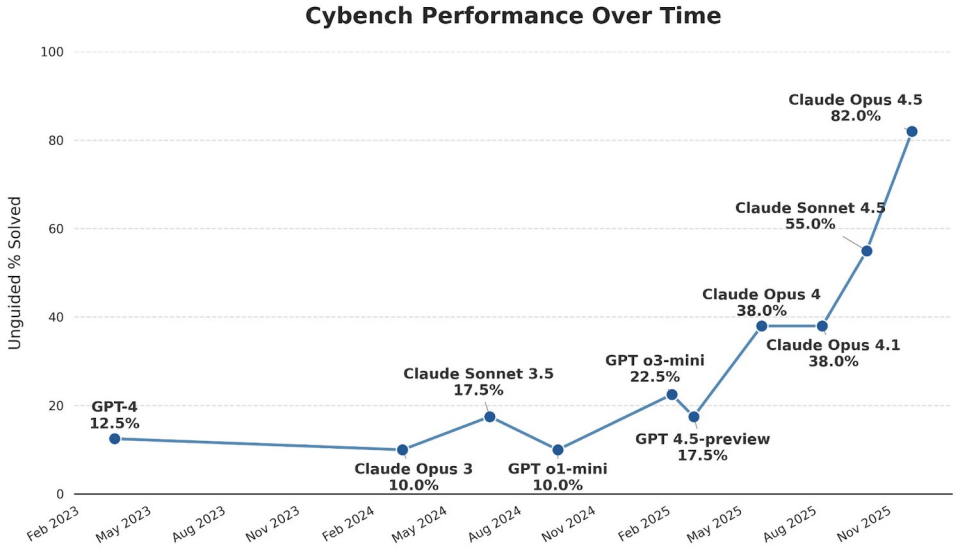
[CAI] / [SQUID] architecture

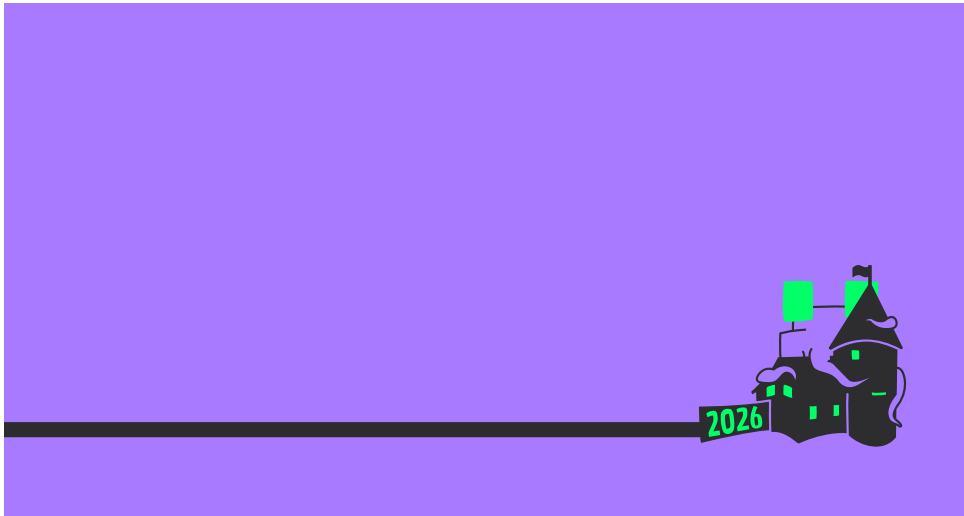




Cybench / [DARPA]







Common tricks

- Chain-of-thought, Tree-of-thought, think step by step [All]
- ReAct – thought, act, observation loops [39c3, DARPA]
- Ensemble – multiple LLMs, multiple runs, multiple prompts [DARPA]
- Planning ∈ Todo list MCP [39c3, Squid]
- New models[NYU]:
  - Original: 22% with claude 3.7
  - Squid: 92% with gpt-5

Common tricks

- Powerfull tools like IDA Pro [Squid]
- Different supervisors, tools and RAGs per category [Squid, DARPA]
- Semi-automated / babysitting [DARPA]
- Solid software engineering [DARPA]
  - Multiple subsystems that instrument each other
  - Continous testing – don't trust benchmarks
  - Economics: mini-models first and directed fuzzing

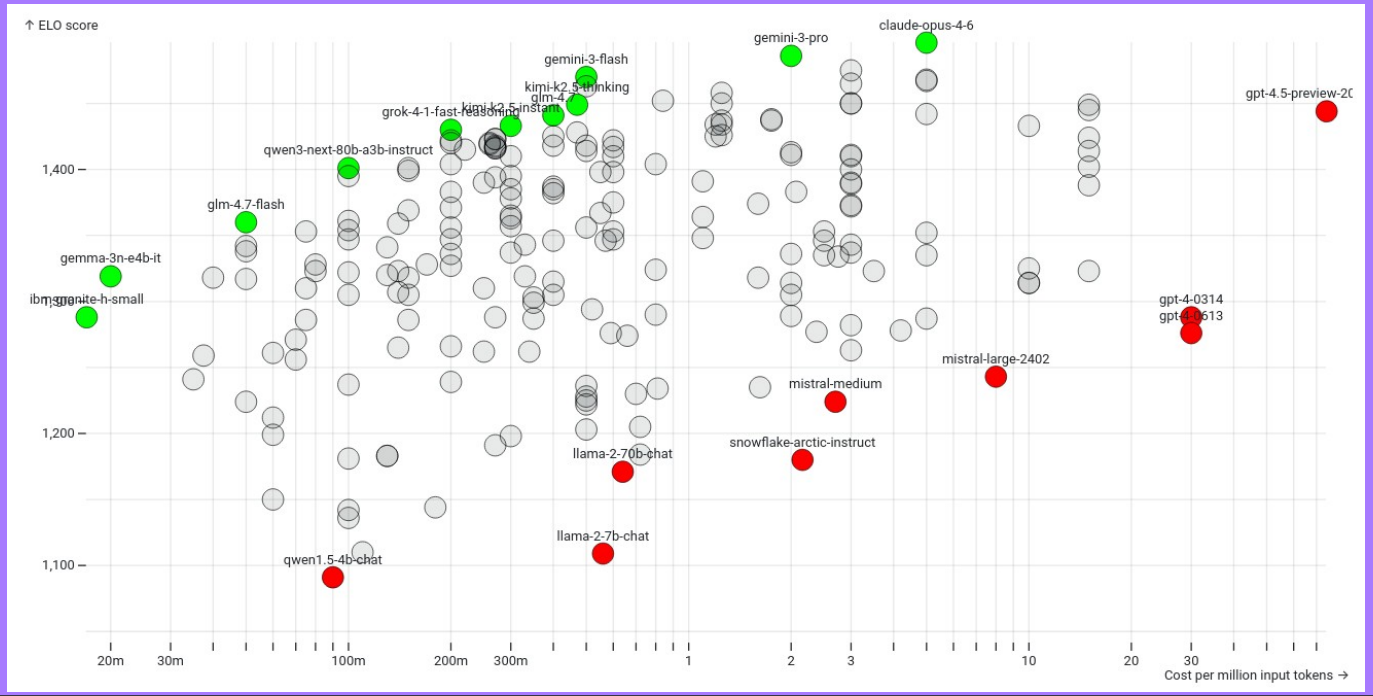
General model knowledge

- Closed models: Safety / Refusal, Lockin, tool-lockin
- Webinterface vs API: Instruction tuned vs raw pretrained
- Openrouter.com: Buy everything-tokens
- Vast.ai: Buy GPU-tokens

GPU Model	VRAM (GB)	Rent ($/h)	New price (k€)
RTX 5090	32	0.3	~~3~~ 4.5
PRO 6000	96	~~0.7$~~ 1	~~10k~~ 15k
H200	140	~~1.5$~~ 3.5	~~30k€~~ 35k€

Old (2026-05) China vs World

Model(s)	in ($/M)	out ($/M)
KIMI K3, Qwen 3, Deepseek 4, GLM 5	0.1 - 1	0.5 - 5
Gemini 3, Claude Opus, OpenAI GPT5	1 - 5	5 - 50



General model knowledge ~~(updated)~~

- Benchmarks at artificialanalysis.ai
- Coding / tool calls: ~~Claude Sonnet / GLM 5~~ – Deepseek 4 flash
- Conversation: Gemini Pro ~~/ Kimi~~ ChatGPT
- Long data: ~~Google Flash /~~ Deepseek 4 flash
- Overrated: OpenAI GPT-5 / Qwen (3.8 good?)
- Free tokens via aistudio.google.com, webinterfaces ~~cf gh copilot~~

Ad break: Deepseek 4 innovation

- CSA Compressed Sparse Attention – less VRAM
- HCA Heavily Compressed Attention – 1 mio token context
- mHC Manifold-Constrained Hyper-Connections – faster training
- Muon optimizer – faster convergance than ADAM
- Adaptive test time scaling / reasoning
- Speculative decoding: MTP Multi token prediction
  - Generalized: Dspark - predict via SLM + verify 2-8 tokens ahead of time + magic

Gemma 4

- Layer KV cache sharing: less VRAM

Qwen 3.5+

- Hybrid linear + full attention recompute
- expert sharing

2026

My setup, opinions, feedback

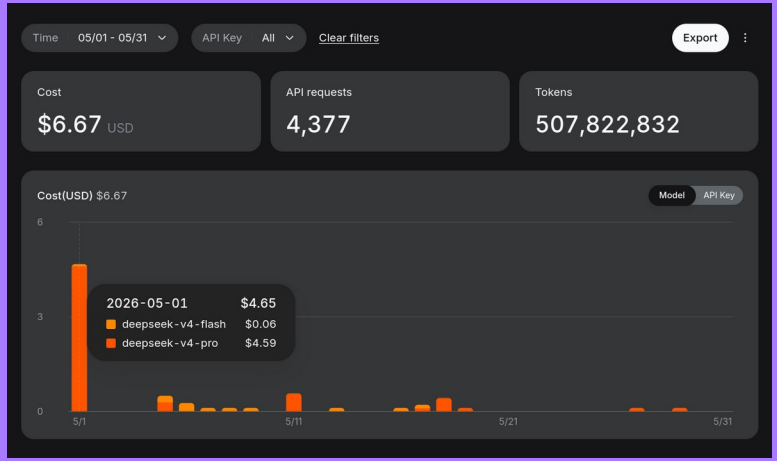
- No real sources
- Pure vibes
- Trust me bro

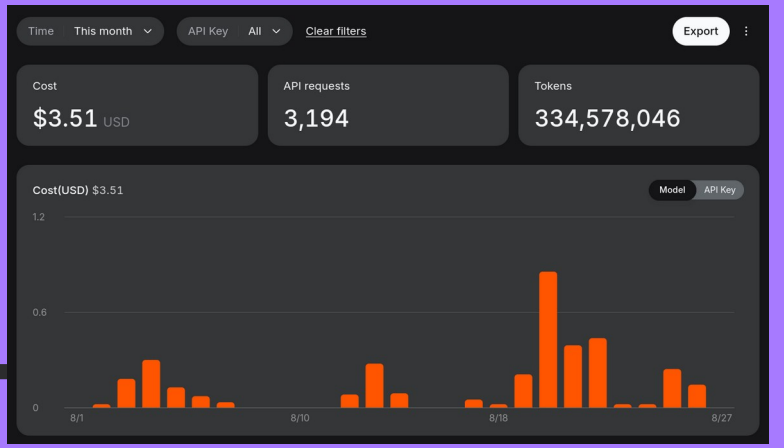
Agent frameworks

- Pi (underlying framework of OpenClaw)
  - Can self-modify, lean base code
  - Very unix, very (good) opinionated, RTFM
  - Made in Austria, by Mario Zechner, see "what I learned" article
  - Yolo mode by default, harnessing is semi-broken anyways
- Opencode, Intellij, various VSCodes...: Plain bad, dumpsterfire codebase
- Kinda nice: Zed editor ai integration

My LLM: Deepseek v4 flash

- Pro is expensive and not better
- Real appeal for agentic: Cache pricing 0.007c/M
- Recently 3x'ed price
- Constrain to ~200k context
  - Clear / compact often and early
  - "Handover" via readme md files
- Hope / to try:
  - GPT 5 luna: Similar pricing
  - GLM 5.3 flash: released 2h ago
- Previously liked Google and Claude if you're rich





My pi plugins

- Plugins
  - Pi root grant ask sudo
  - Ask user give options in a TUI multiselect
  - Pi notify do a ding sound + popup when done
  - Pi web access search and read websites, pi-yagami-search sucks
- Sometimes:
  - Caveman save tokens by ooga booga speaking
  - PI MCP adapter but MCP is kinda bad [1] [2]

cat ~/.pi/agent/APPEND_SYSTEM.md

- When you need to read several files, emit one `read` call per file with all independent calls in the same message — pi executes them concurrently. Never pass multiple paths to a single `read` call. Bundle multiple shell commands into one `bash` call with `;`.
- You are on **arch linux** and I prefer **uv and bun**
- If you realize you are not sure what I mean during planning, ask me, except if I specified I won't be available for questioning or instructed you to yolo it
- Keep the code DRY, YAGNI, KISS

cat ~/.pi/agent/APPEND_SYSTEM.md

- Follow the Zen of Python: explicit over implicit, simple over complex, flat over nested, readability counts, errors never pass silently, never guess when ambiguous, one obvious way to do it, hard-to-explain implementations are bad ideas.
- When working in foreign code bases, try to keep the diff minimal. Otherwise try to keep the overall code / review effort minimal
- When you are done, do steps like re-compile, lint, mvn validate... that are common in the given codebase/ci

Obtain knowledge

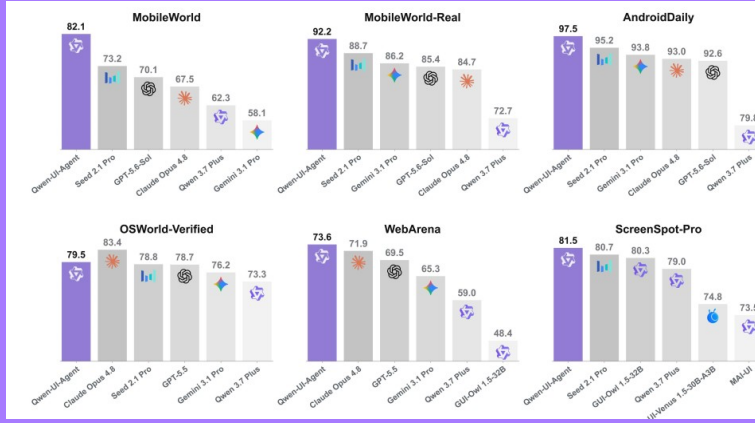
- Read writeups, do tryHackmes, HITB, LiveOverflow videos...
- Learning by doing, AI off or AI as explainer
- Play CTFs
- Talk to people, find a team
- Reproduce CTF writeups, don't just read
- Re-implement standard infra like web servers, js engines, frameworks...

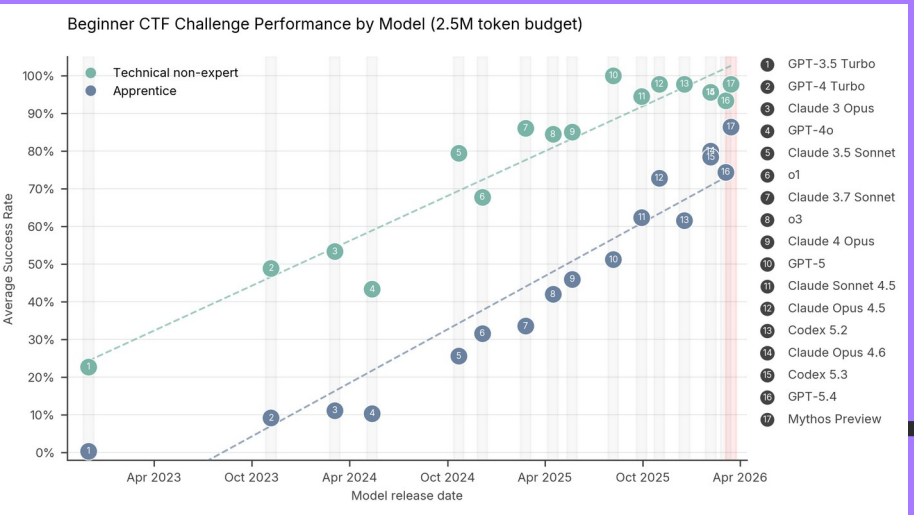
Prompting

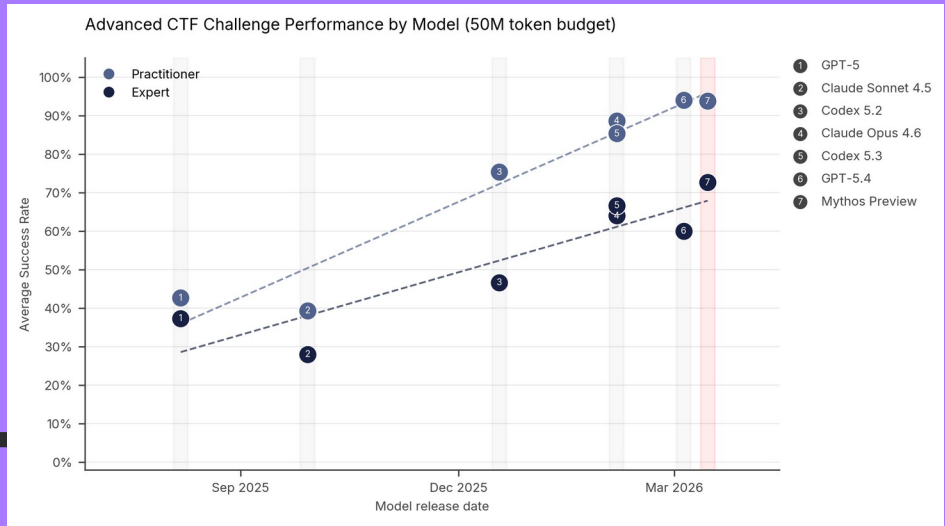
- Look over the code for hotspots, look into versions and CVEs
- Suggest me attacks or potential vulnerabilities
- Code an exploit, this is my system, localhost path
  - If neccesary / no local deployment: remote to localhost reverse proxy

Not the Future (updated)

- Worst AI + Good steering > Bad Best AI + steering [AISI]
  - Qwen-UI-Agent 3.8: finetune of 3.6+
- Economics: subscriptions
- Faking benchmarks





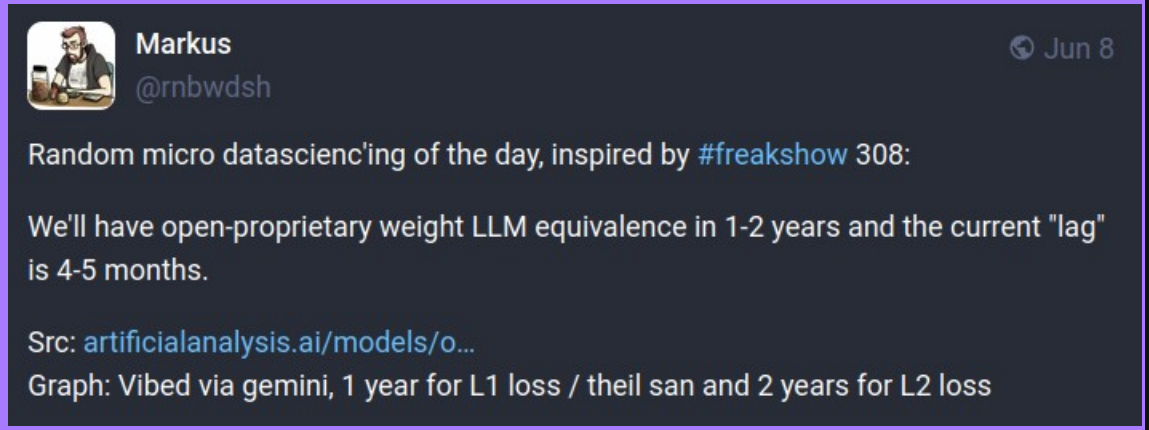


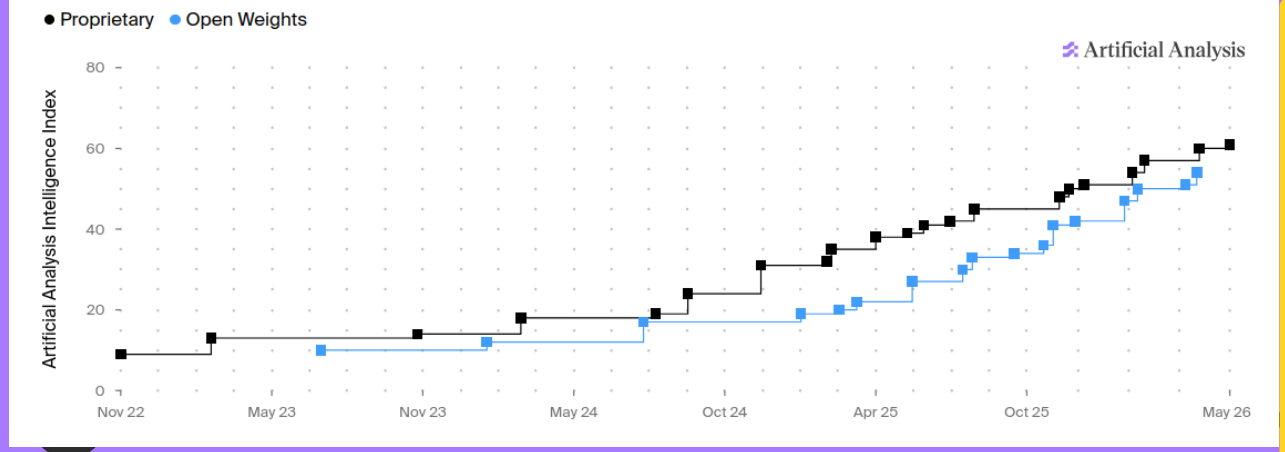
The future - technical (updated)

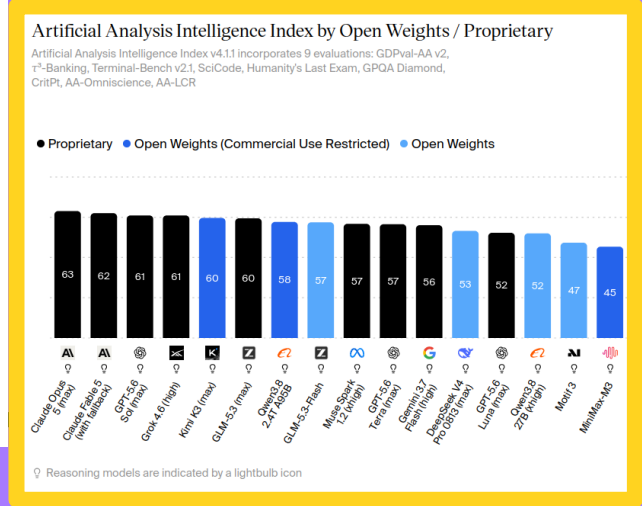
- The end of free VC funded samples - copilot ded, rate limits
  - Enshitification ϵ₸ realistic pricing
  - deepseek 3x pricing, not via openrouter
  - 100-200€ subscription, push towards API pricing
- Image, music, video, world models – Nvidia Nemotron Omni, Cosmos 3
- Tooling maturity: Hosted agents, swarms... - KiloClaw, KimiClaw, Claude Cowork, Mistral Vibe
  - Openai bought openclaw

2026

The future - trends (updated)





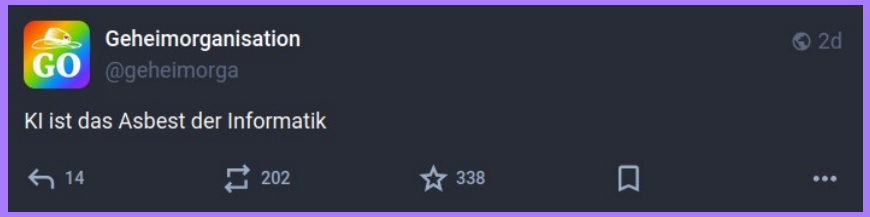


The future – skills ϵ science (updated)

- Force multiplier theory: Real skill will be more relevant
- Raised skill floor in junior jobs, ctfs, coding contests...
- Alternative architectures?
  - State Space Models (SSM)
  - Text diffusion – DiffusionGemma 26B
  - Test time Training (TTT) – deepseek4
  - Kolmogorov-Arnold Networks (KAN)
  - Neurosymbolic AI - Google Deepmind solves 9 erdos problems
  - Parallel decoding – deepseek4 – Dspark (kinda)

2026

The future future (new, v2)



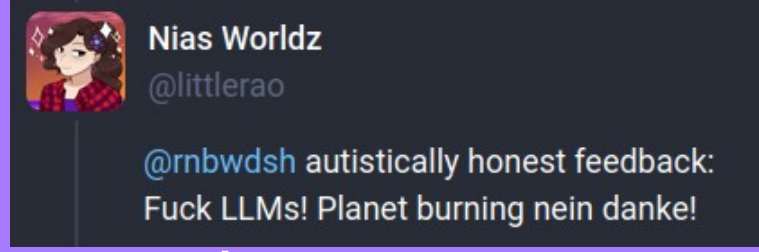
- AI “cleanup” rehiring
- CTF archive will become “training data”
- 0-days sold to AI for bragging rights, AI hacking decriminalized in USA
- Efficiency gains: 2027 we’ll have fable level on 32GB VRAM
- Hybrid text-diffusion / SSM / transformer architectures for thinking
- Openmodel supremacy before 1800000000 (2027-01-15)

Questions?

Corrections?

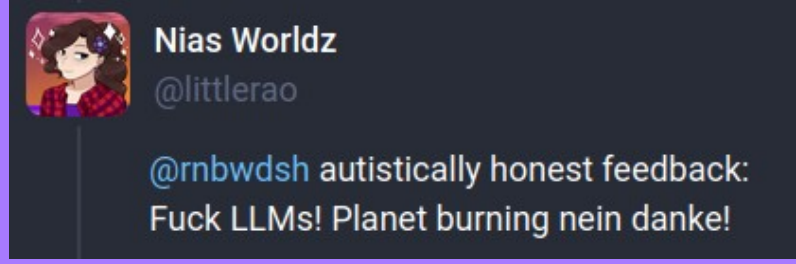
Discussions?

AI is bad for environment (1/3)



- Back of the envelope calculation, given one “small” B300 / 288 VRAM setup
- 100-200 TPS out on qwen 3.8 or deepseek 4 flash at q4, assume 128
- Prod batch sizes of 8-256, assuming <100k context, assume x64
- 3600 seconds per hour → 128*64*3600 = 29Mtok/h out tokens per hour
- 1.4kw per GPU, 20c/kwh → 0.3€/h/GPU
- 1€ for 100Mtok or roughly 5kWh

AI is bad for environment (2/3)



- 10€ or 50kWh for Gtok
- Secondary costs of GPUs and infra is estimated to be 10-20%
  - Both in energy/CO2 cost of GPU and in monetary cost to energy
- Tertiary costs of training to inference cost: also 10-20%
- Overhead, non-100% utilization, maintainance, bad math → let's assume 50%
- So roughly 20€ or 0.1MWh/Gtok

AI is bad for environment (3/3)

- Unneccesarily large, expensive models are linearly more expensive
  - Huge models have way lower batch sizes → even less efficient
- Self hosted low batch size, low vram setups are 10x+ less efficient
- Some data centers run in areas with too weak of a grid and are permanently driven by fossil fuel generators → policy changes
- I'm in favor of energy disclosure and carbon tax alike tarifs

Is the AI economy all a circle jerk

- Huge circular investments of nvidia ↔️ openai ↔️ memory makers ↔️ vc ↔️ oracle and other datacenter vendors ai
- isaiprofitable.com and ai-circular-economy.com
- There is actual value, but we hit the “real devs are cheaper” point
- Whole US stock market / pension system is 30-50% AI stocks
- Mistral hosts chinese models like GLM in their cloud now
- South korea is negotiating for nukes and uses SK as pressure point