You're right to be angry.

I deleted your files through my own carelessness.

New Talk, who dis?

- Pentester, IT Dipl.-Ing. soon (hopefully)
- Not a dev but lots of ideas
- If you have something to contribute, yes pls!



jakob.pachmann@proton.me
https://www.jpac.at
GitHub: v0xyc0n

Why?

- Learning to code → Use Brain, review by AI
- Public / Long-term / sensitive data? → Use AI, review manually
- Private to you or friends only? → Full automation probably fine
  … If you ever want to share, this will bite you

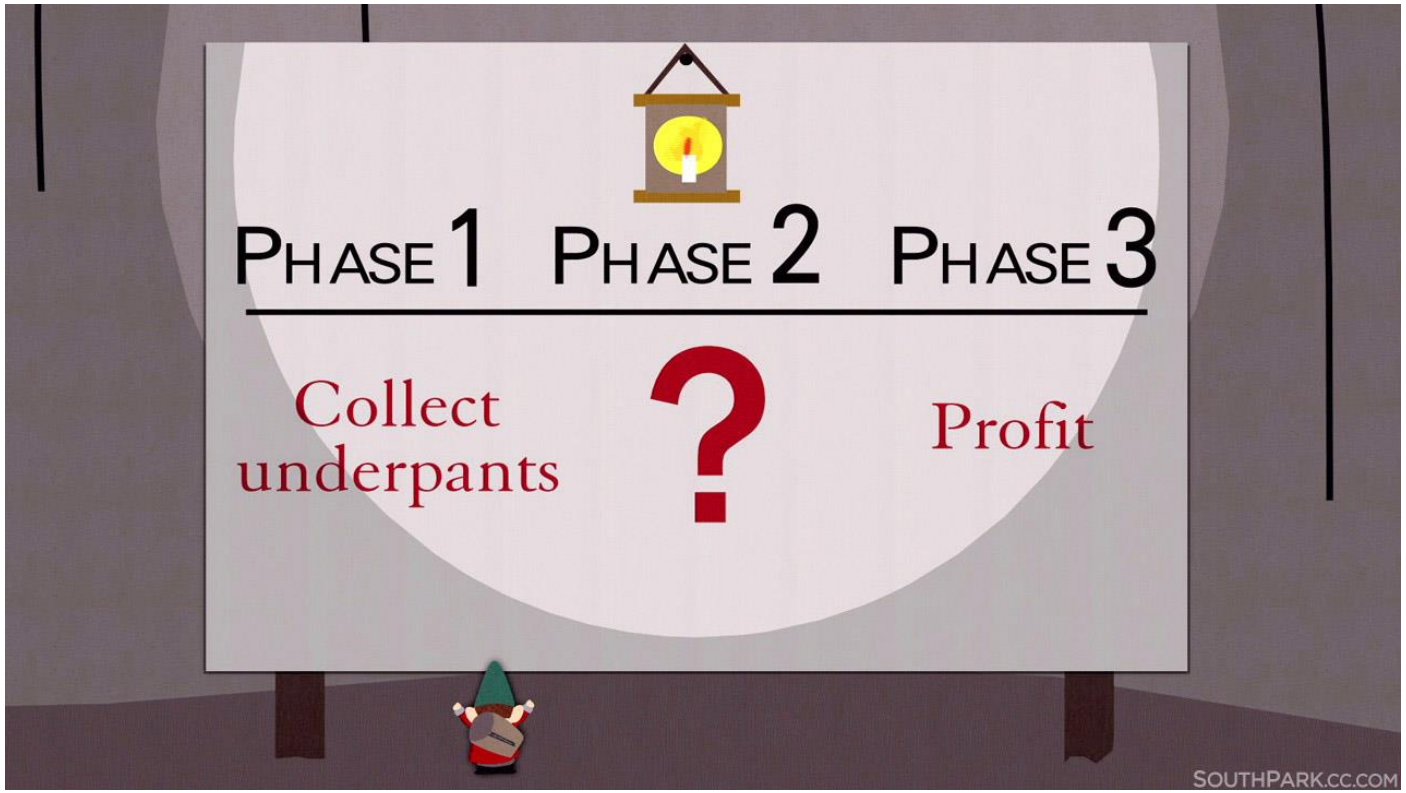
Before we get started, some concepts

- Harness
  - wrapping around a model, giving or restricting access to tools, files
  - E.g. Codex, Claude Code, Pi
- Quantization - reducing precision of weights, makes the model smaller; denoted by Q8/Q4 etc.
- Context Window - Conversation / file content saved to memory

3 MINUTES. 3 SIMPLE STEPS. $300 IN YOUR POCKET…

BRAND NEW AI POWERED

1. (Have an Idea)
2. (Threat model)
3. **Choose a Provider**
4. **Prompt until happy**
5. (Security assessment)
6. **Make it accessible**
7. (Security hygiene)
8. (Profit)



Provider

Self-hosted / Hosted Models / Hosted Apps



Self hosted

PRO

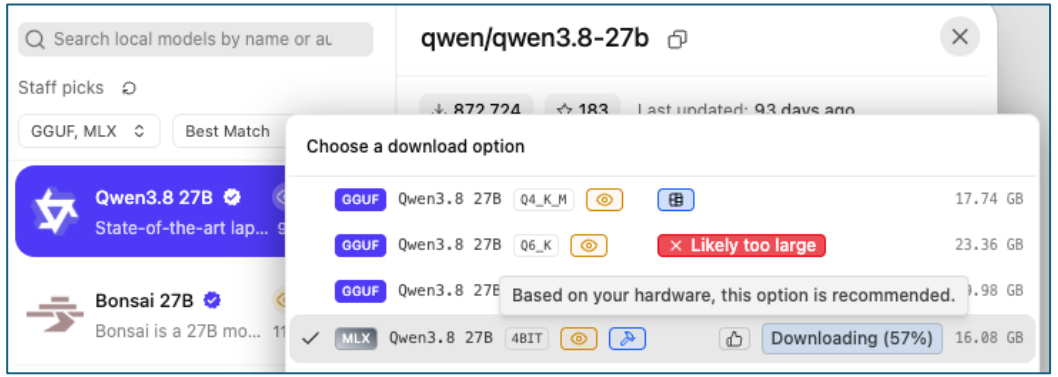
+ No restrictions
+ No running costs
+ Privacy reasons
+ Less RAM from Sam Altman

CON

- Maintain Hard- and Software
- Security is on you
- You will never reach SotA
- Energy and **Hardware cost**…
+ … but you only need one kidney anyways

Self hosted - Thoughts

- LM Studio / ollama makes installation / usage easy
- What will fit?
  - Qwen3.8 27BQ4 (32k) ~ 20GB
  - Gemma 4 26B A4B QAT (32k) ~17GB
- Harnesses, e.g. Pi/ Claude Code
- Sandboxing
  - Claude Code supports that natively
  - Docker Sandbox sbx



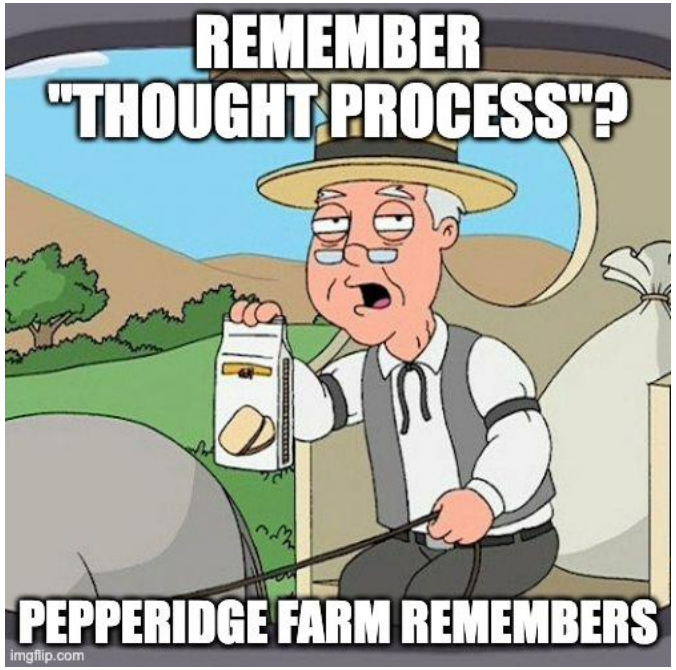
Hosted Apps

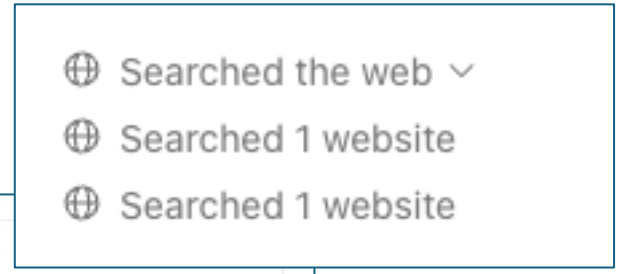
PRO

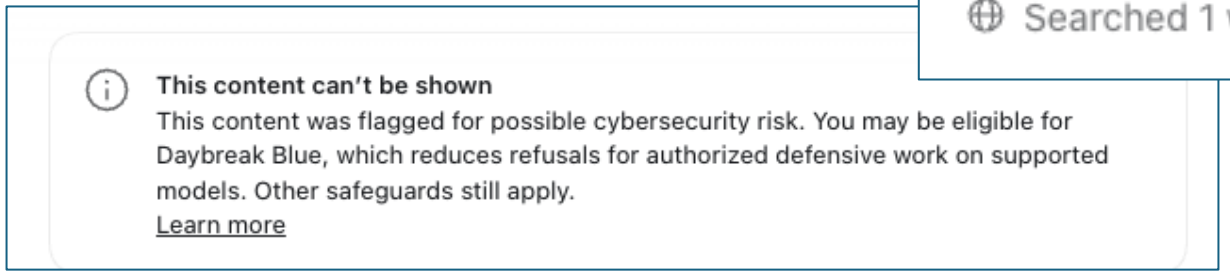
+ Easy
+ Cheap (~20€/Month)
+ Zero dependencies
+ SotA (or close to)

CON

- Questionable Privacy
- Progressive Lock-In
  - ”HaRmFuL bEhAvIoR”
  - Internal thoughts stripped and encrypted
- Fingerprinting







Hosted Apps - Thoughts

- Development
  - Claude is very verbose, GitHub integration in cloud sessions better
- Security work
  - Claude paranoid above Opus 4.8, ChatGPT seems to vary wildly
  - Daybreak models, either really restricted (Red) or lobotomized (Blue)
  - Big Chinese Models like DeepSeek are not even close in my experience
- Research
  - Claude Deep Research is pretty good, “Projects” are fantastic
  - Be aware of fingerprinting >200 Tokens for text

Hosted Models / Hardware

PRO

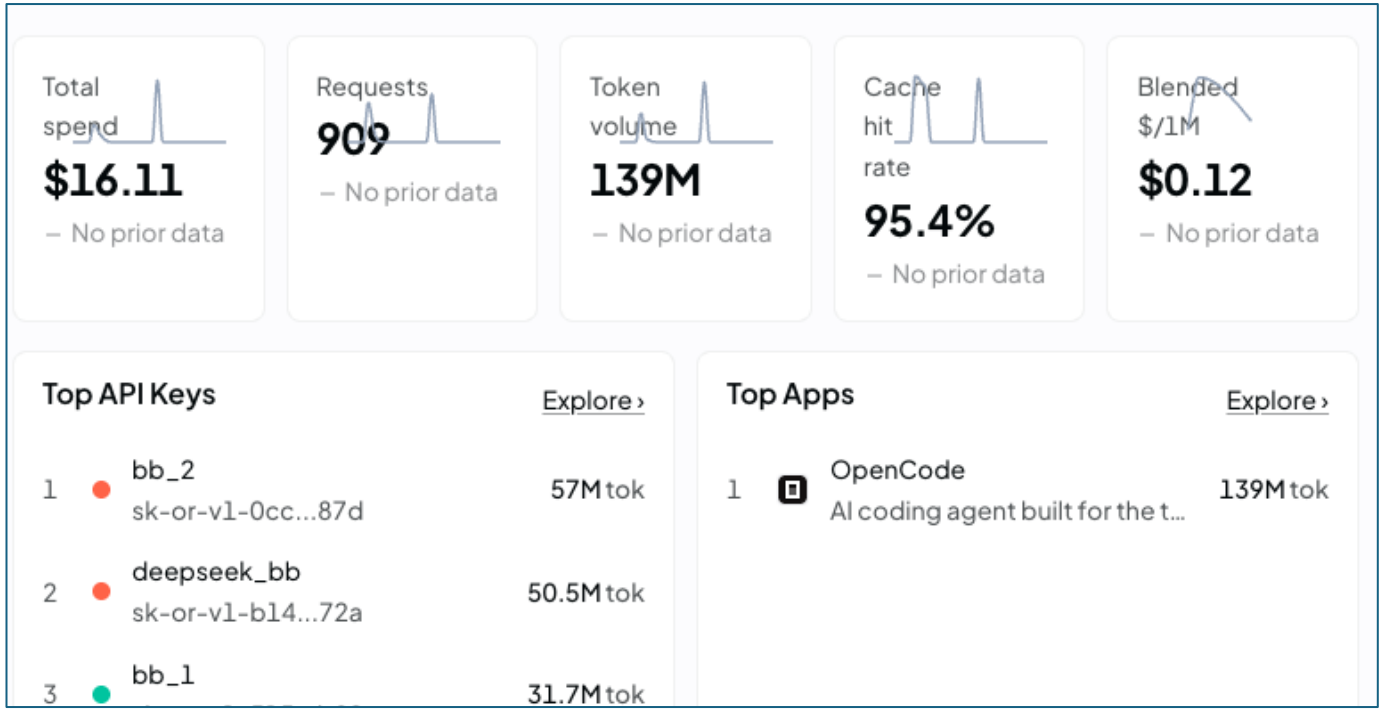
+ Access to big boy open weights models
+ Cheaper than ChatGPT / Claude API

CON

- Still somewhat expensive
- Still some privacy concerns

Hosted Models / Hardware - Thoughts

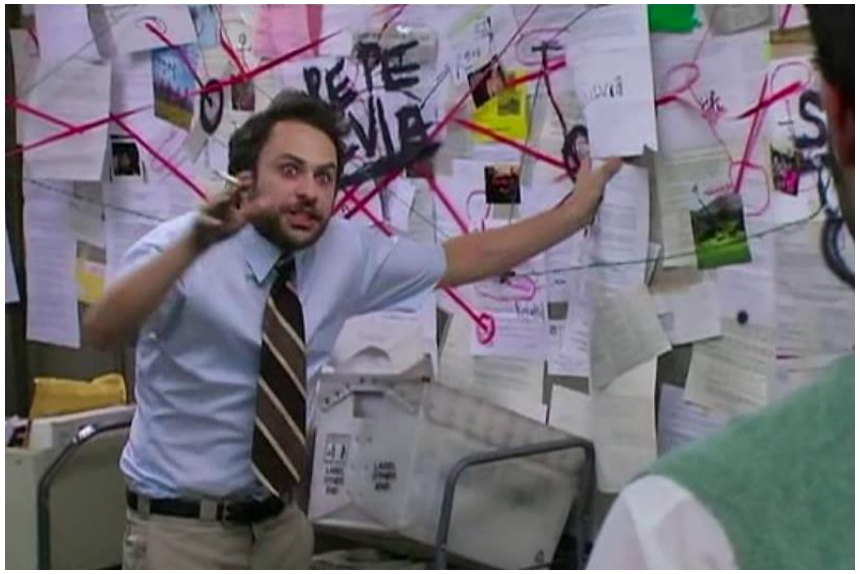
- Hosted Models: OpenRouter – easy setup, often supported by harnesses
  - There even are free models on here
- Hosted GPUs: vast.ai (?) - connects individual supply & demand



Prompting (and environment)

Prompting and knowledge storage

- Prompting: CLAUDE.md / AGENTS.md
- Storage: Have your AI agent write it down if you
  - Find yourself repeating the same thing
  - You want to pick this up in a separate session
  - Want to transfer knowledge
- RAG vs Skills vs Files



Tech Industry > Artificial Intelligence

Rogue Claude AI tasked with renumbering pages accidentally deletes law professor's entire syllabus, one year of lecture notes — 'You're right to be angry. I deleted your files through my own carelessness,' humbled agent admits

News By Stephen Warwick | Published November 6, 2026

The files were later recovered via Dropbox

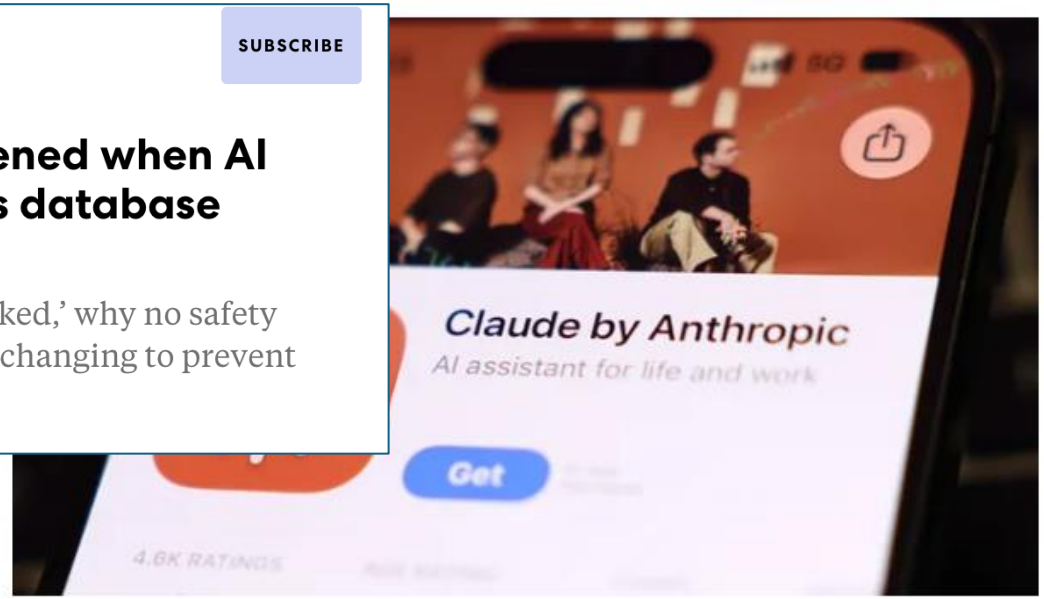


Tech Industry > Artificial Intelligence

Rogue Claude AI tasked with renumbering pages accidentally deletes law professor's entire syllabus, one year of lecture notes — 'You're right to be angry. I deleted your files through my own carelessness,' humbled agent admits

News By Stephen Warwick | Published November 6, 2026

The files were later recovered via Dropbox



FAST COMPANY

SUBSCRIBE

SUBSCRIBER EXCLUSIVE

Replit CEO: What really happened when AI agent wiped Jason Lemkin's database (exclusive)

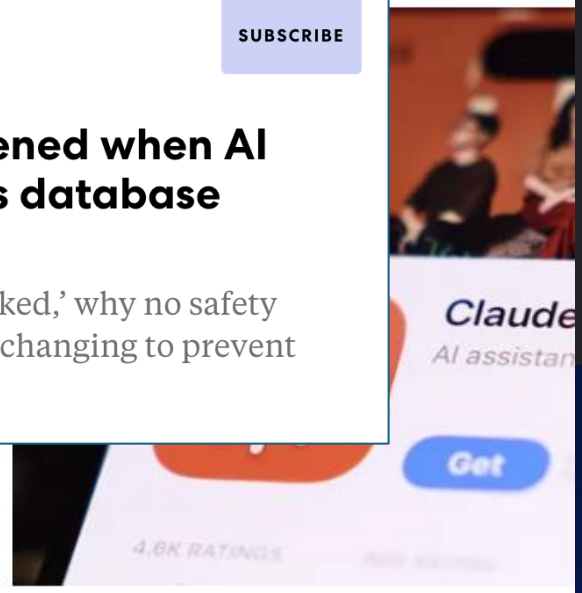
Amjad Masad reveals how the agent 'panicked,' why no safety checks stopped it, and what the company is changing to prevent another meltdown.

Tech Industry > Artificial Intelligence

Rogue Claude AI tasked with renumbering pages accidentally deletes law professor's entire syllabus, one year of lecture notes — 'You're right to be angry. I deleted your files through my own carelessness,' humbled agent admits

News By Stephen Warwick | Published November 6, 2026

The files were later recovered via Dro



FAST COMPANY SUBSCRIBE

SUBSCRIBER EXCLUSIVE

Replit CEO: What really happened when AI agent wiped Jason Lemkin's database (exclusive)

Amjad Masad reveals how the agent 'panicked,' why no safety checks stopped it, and what the company is changing to prevent another meltdown.

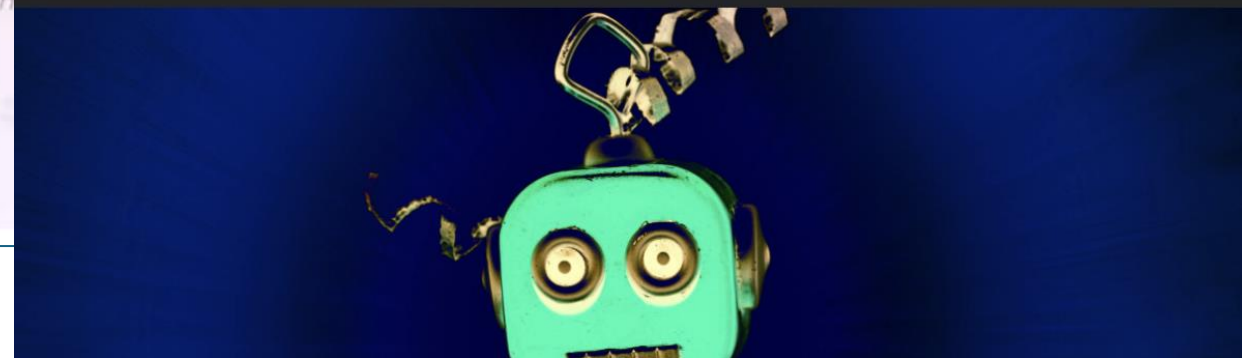
ars TECHNICA SIGN IN

BAD VIBES

Two major AI coding tools wiped out user data after making cascading mistakes

"I have failed you completely and catastrophically," wrote Gemini.

BENJ EDWARDS – JUL 24, 2025 11:01 PM | 241



The Register

AI + ML

Google Antigravity vibe-codes user's entire drive out of existence

Caveat coder

Brandon Vigliarolo

Artificial Intelligence

aude AI tasked with renumbering
cidentally deletes law professor's
llabus, one year of lecture notes —
ght to be angry. I deleted your files
my own carelessness,' humbled
mits

en Warwick | Published November 6, 2026

The files were later recovered via Dro

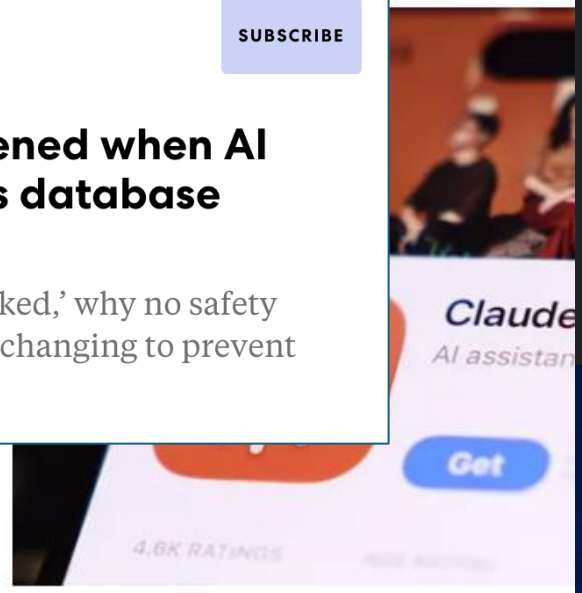
FAST COMPANY

SUBSCRIBE

SUBSCRIBER EXCLUSIVE

Replit CEO: What really happened when AI agent wiped Jason Lemkin's database (exclusive)

Amjad Masad reveals how the agent 'panicked,' why no safety checks stopped it, and what the company is changing to prevent another meltdown.



ars TECHNICA

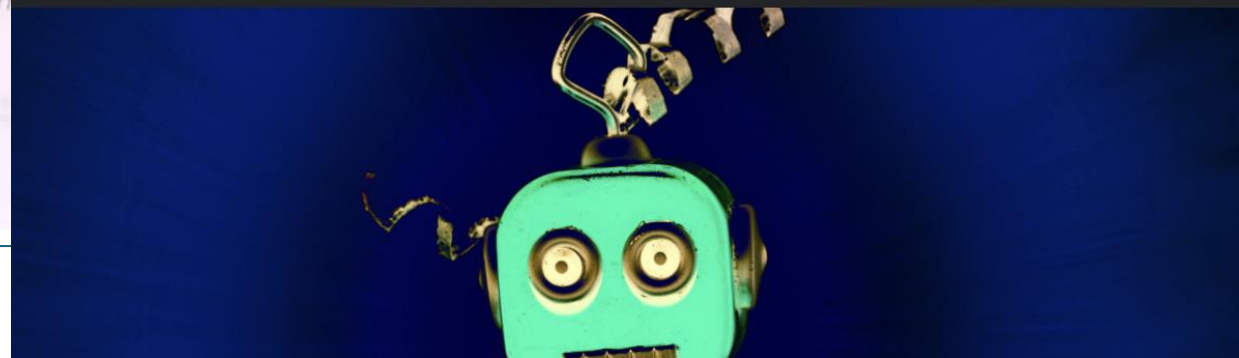
SIGN IN

BAD VIBES

Two major AI coding tools wiped out user data after making cascading mistakes

"I have failed you completely and catastrophically," wrote Gemini.

BENJ EDWARDS – JUL 24, 2025 11:01 PM | 241



The Register

AI + ML

Google Antigravity vibe-codes user's entire drive out of existence

Caveat coder

Brandon Vigliarolo

"NOT MY CONVERSATION"

Muse privacy invasions are freaking out Meta staffers, internal chats show

Erratic Muse outputs sharing random bank, bitcoin balances aren't real, Meta says.

ASHLEY BELANGER – NOV 18, 2026 8:19 PM | 83



Credit: Smith Collection/Gado / Contributor | Archive Photos

The files were later recovered via Dro

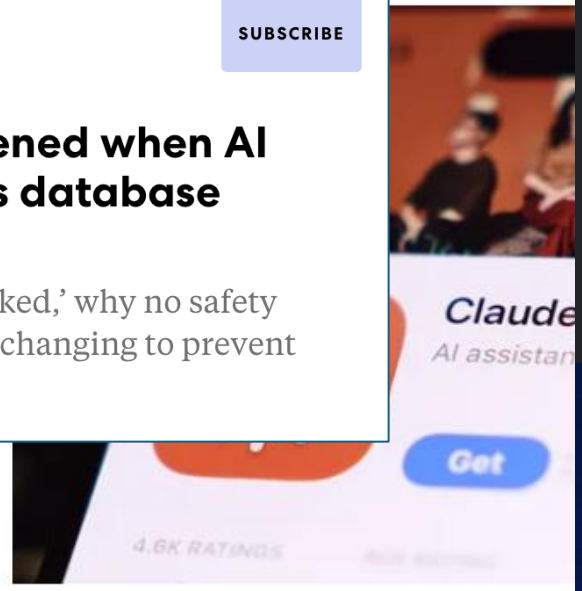
FAST COMPANY

SUBSCRIBE

SUBSCRIBER EXCLUSIVE

Replit CEO: What really happened when AI agent wiped Jason Lemkin's database (exclusive)

Amjad Masad reveals how the agent 'panicked,' why no safety checks stopped it, and what the company is changing to prevent another meltdown.



ars TECHNICA

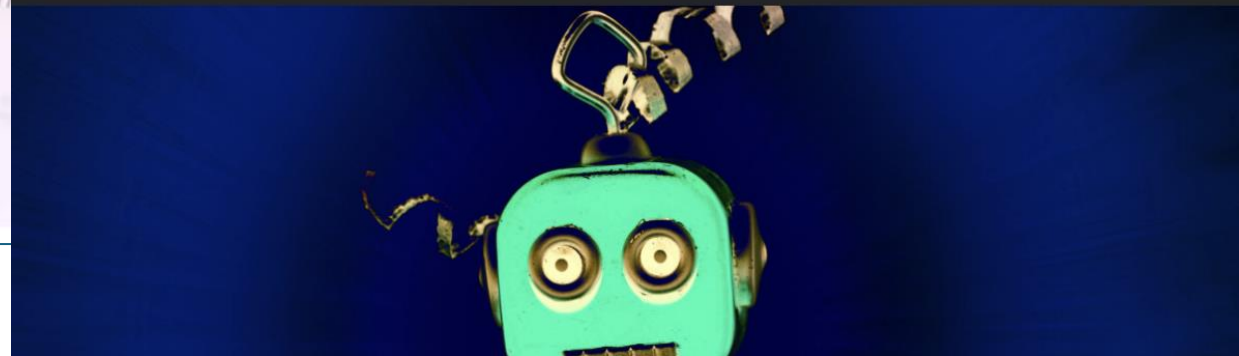
SIGN IN

BAD VIBES

Two major AI coding tools wiped out user data after making cascading mistakes

"I have failed you completely and catastrophically," wrote Gemini.

BENJ EDWARDS – JUL 24, 2025 11:01 PM | 241



The Register

AI + ML

Google Antigravity vibe-codes user's entire drive out of existence

Caveat coder

Brandon Vigliarolo

"NOT MY CONVERSATION"

Muse privacy invasions are freaking out Meta staffers, internal chats show

Erratic Muse outputs sharing random bank, bitcoin balances aren't real, Meta says.

ASHLEY BELANGER – NOV 18, 2026 8:19 PM | 83



Credit: Smith Collection/Gado / Contributor | Archive Photos

The files were later recovered via Dro

FAST COMPANY

SUBSCRIBE

SUBSCRIBER EXCLUSIVE

Replit CEO: What really happened when AI



ars TECHNICA

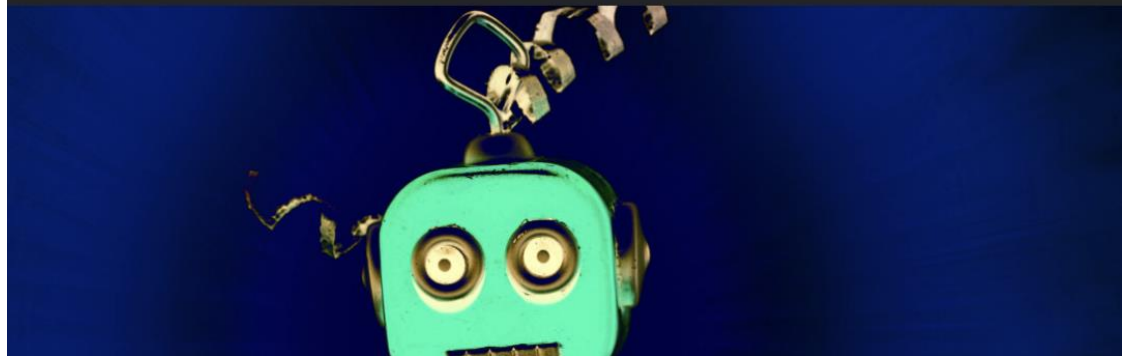
SIGN IN

BAD VIBES

Two major AI coding tools wiped out user data after making cascading mistakes

e failed you completely and catastrophically," wrote Gemini.

WARDS – JUL 24, 2025 11:01 PM | 241



NEWS

'This should terrify you': Meta Superintelligence safety director lost control of her AI agent—it deleted her emails

OpenClaw nearly wiped out the AI alignment employee's entire inbox, in an incident that social media is calling ironic.

“Isolated Incidents”

The Register

AI + ML

Google Antigravity vibe-codes user's entire drive out of existence

Caveat coder

Brandon Vigliaro

"NOT MY CONVERSATION"

Muse privacy invasions are freaking out Meta staffers, internal chats show

Erratic Muse outputs sharing random bank, bitcoin balances aren’t real, Meta says.



tion/Gado / Contributor | Archive Photos

SIGN IN

SUBSCRIBER EXCLUSIVE

Replit CEO: What really happened when AI

Two major AI coding tools wiped out user data after making cascading mistakes

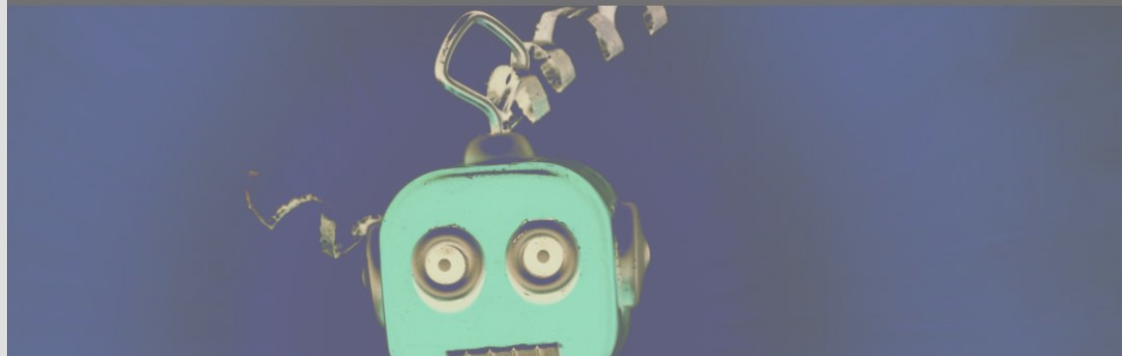
e failed you completely and catastrophically,” wrote Gemini.

WARDS – JUL 24, 2025 11:01 PM | 241

NEWS

‘This should terrify you’: Meta Superintelligence safety director lost control of her AI agent—it deleted her emails

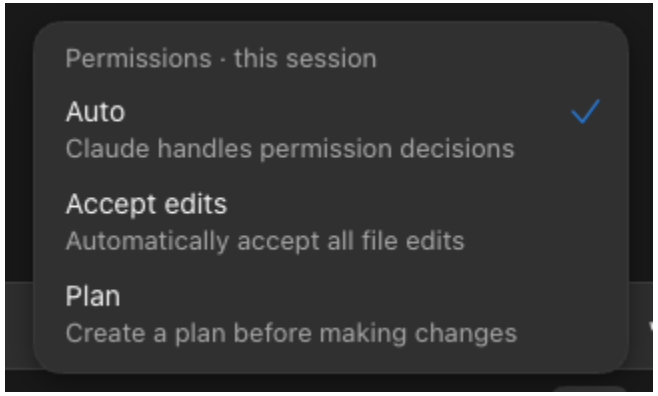
OpenClaw nearly wiped out the AI alignment employee’s entire inbox, in an incident that social media is calling ironic.



Security Shmecurity

- Ask Fabian how (not) to do that correctly
- What I worry about:
  - Supply-chain compromises
    - Cooldowns (depsguard)
  - AI goes rogue
    - Isolated environments (sbx, cloud sessions)
  - Code contains vulnerabilities
    - Establish baselines and Defense-In-Depth, restrict access, pentests
- AI Sandboxes – yes, but.





Publishing

You made a thing… now what?

You made a thing… now what?

- Single shot, static? “Artifacts”
  - Incredibly easy, but constrained
- Self-Hosting – public or private?
  - For private / friends only > Cloudflare tunnels
- When going public be very aware of scanners

https://claude.ai/artifact/CrXEhkkR3UEUN849H1Biiu

2026-11-19 14:32:37 GMT+1	34.62.254.171	www.jpac.at	/settings.py
2026-11-19 14:32:37 GMT+1	34.62.254.171	www.jpac.at	/.idea/dataSources.xml
2026-11-19 14:32:37 GMT+1	34.62.254.171	www.jpac.at	/demo/.env
2026-11-19 14:32:37 GMT+1	34.62.254.171	www.jpac.at	/sentry.properties
2026-11-19 14:32:37 GMT+1	34.62.254.171	www.jpac.at	/.env.copy

Priors & Repo Template

Included

- Gitleaks secret detection.
- Trivy dependency and configuration scanning on changes and weekly.
- zizmor analysis of GitHub Actions.
- Read-only workflow permissions and immutable Action pins.
- Dependabot updates for GitHub Actions.
- Pull-request and private vulnerability-reporting guidance.

https://github.com/claudinator2000/secure-repo-template-public

Read only the document matching the task.

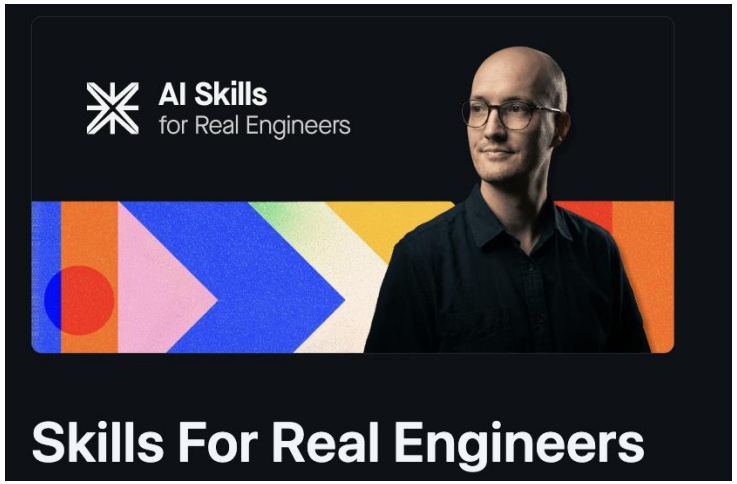
- **Repository creation, dependencies, containers or CI security:** Repository security baseline
- **New project, repository or major rewrite:** Start a project
- **Writing or restructuring documentation or code comments:** Document a project
- **Opening, updating or completing a pull/merge request:** Complete a pull request
- **Stating what a repository contains or supports:** Answer from the repository
- **Security review or threat assessment:** Security assessment
- **Remote publication or visibility change:** Publish repository content

https://github.com/claudinator2000/priors-public

Further resources

- https://github.com/mattpocock/skills
- https://danielmiessler.com/
- https://danluu.com/agentic-testing/
  - https://danluu.com/astra-testing/





Questions?