

Slides: <https://magnusviri.com/saintcon2026>

AI Coding Agents

About magnusviri - James Reynolds

- U of U BMus 1999
- U of U Marriott Library, 14 years (Mac admin)
- U of U School of Biological Sciences/COS/LAS, 12 years (Mac, Linux, network admin)
- Security has been a “hobby” of mine for decades
- Since September 2022 (Stable Diffusion 1.0) I’ve been “obsessed with AI”
- <https://magnusviri.com/saintcon2026>



Agenda

- Demo
- Why Learn to use AI Agents?
- The Basics
- What Could Possibly Go Wrong?
- Mostly Secure Setup
- Coding Tips
- Demo Wrap Up

Demo

Every Demo Takes Time

- Ironically, typing a prompt and then doing something else is the new normal
- Let's start demoing NOW!
- tart is a macOS virtualization app
 - Was Cirruslabs but OpenAI bought it [seriously?]
 - Let's add an MCP server

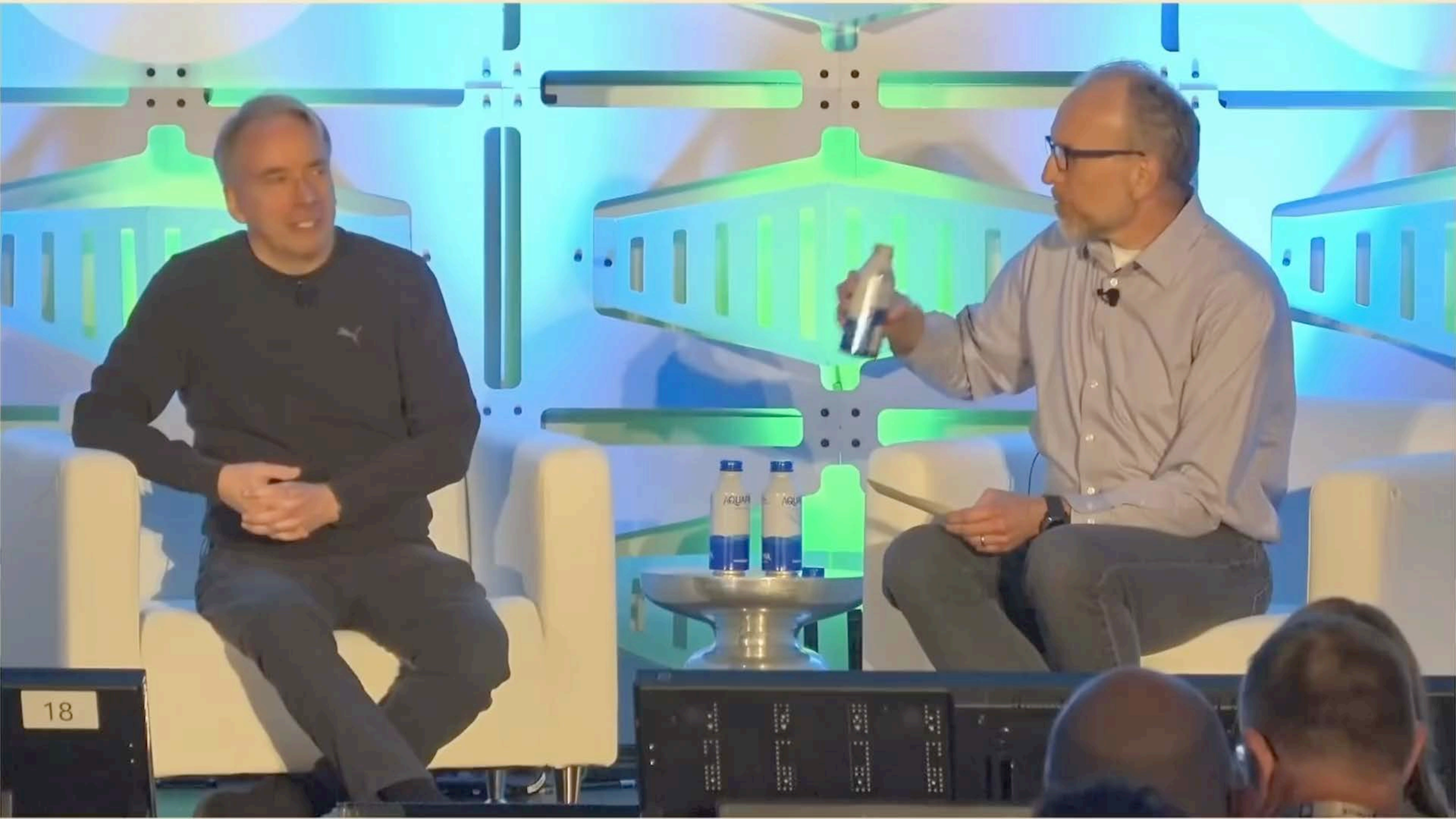
Let's Plan

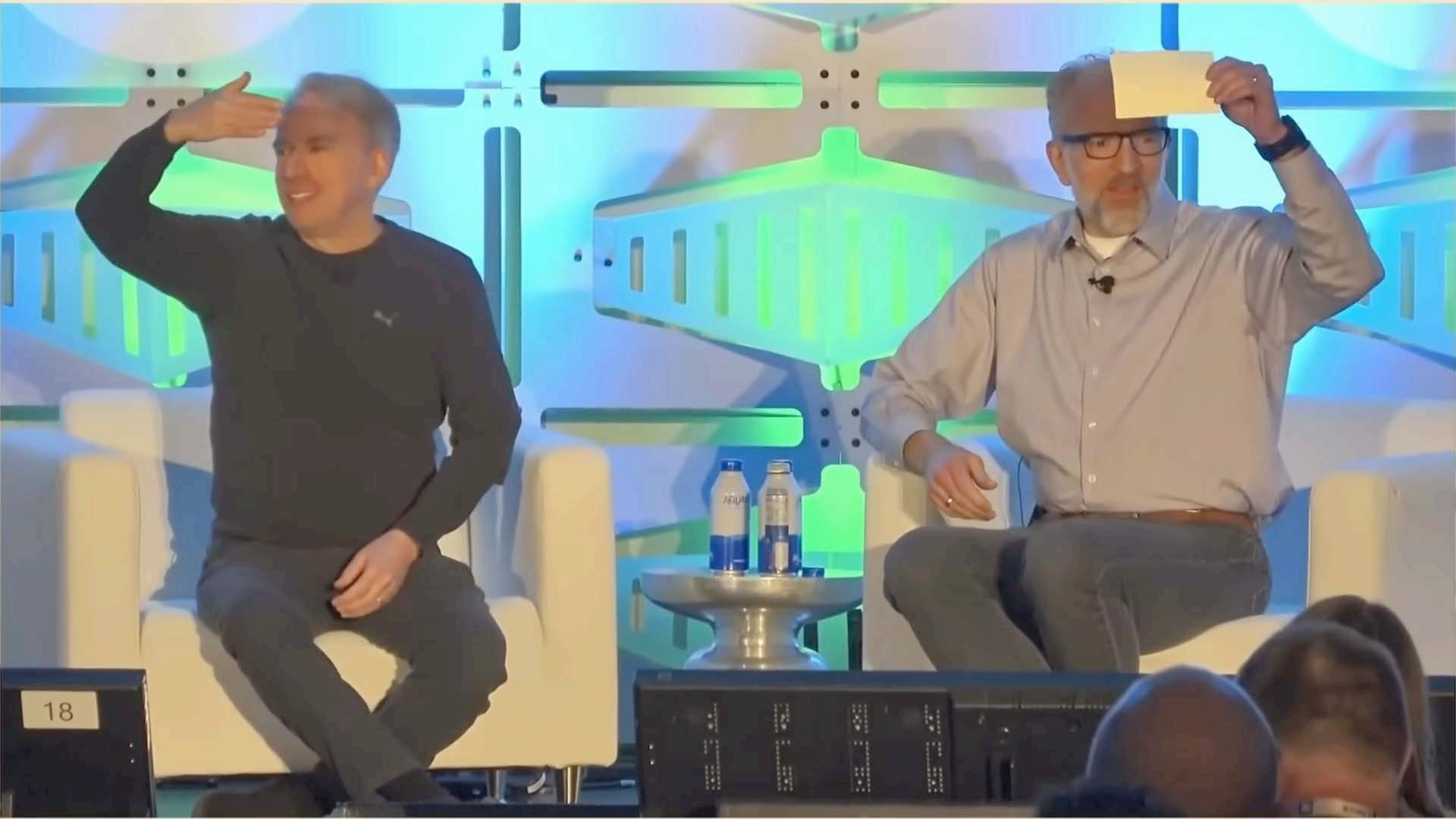
- Goal: an AI agent can autonomously control and ssh into macOS VMs
- The server will run on the macOS host
- The client will run in a Docker container
- Needs to run tart clone, run, ip, and stop
- Needs to be able to ssh into VM (but none others?)
- Needs to be able to check host disk space

Why Learn to use AI Agents?

Open Source Summit, May 2026

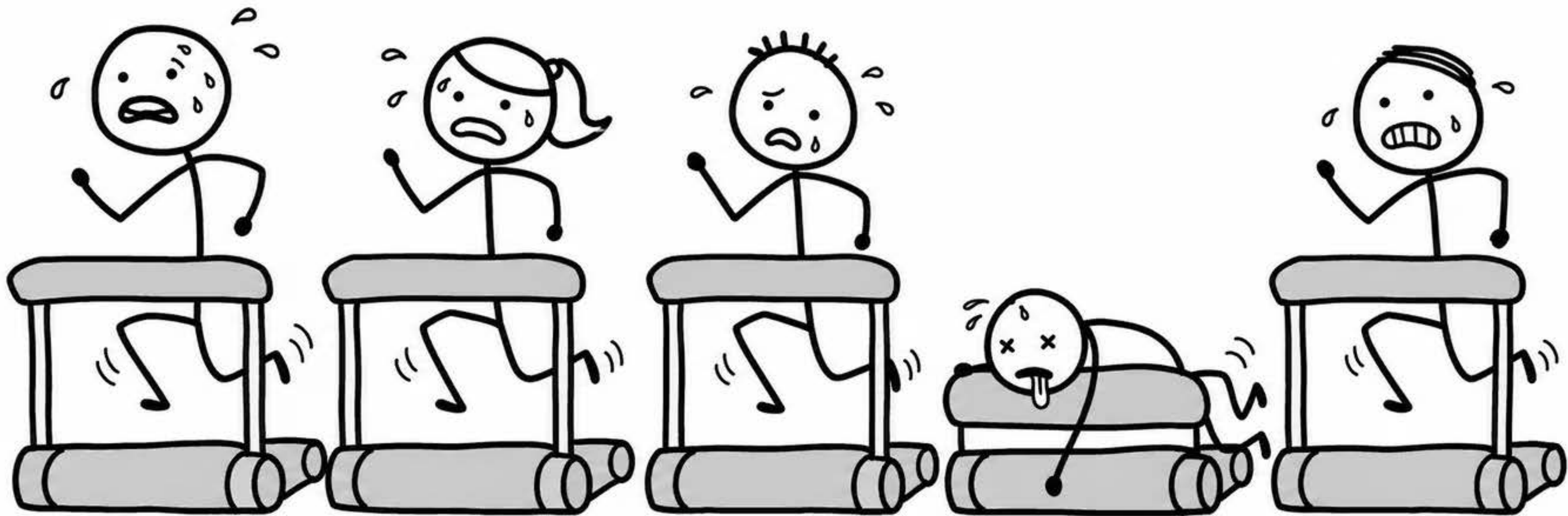
- <https://www.youtube.com/watch?v=fi29pfLcW4I>







But Now We Have the Red Queen Effect



<https://future.lulzx.space/blog/falling-floor.html>

Coding 24/7

- AFK coding is a thing now by using phones
- The AI agent works all night and the day time is spent reviewing code

996 = 9 am - 9 pm, 6 days a week, 72 hrs/wk

- Began in China in late 2010's, "banned" in 2021
- Took root in San Francisco in 2025
- Sergey Brin said all staff should work 60 hrs/wk
- "Somehow, the humans building [AI] are expected to work *more* hours than ever."
<https://www.elenaverna.com/p/im-not-doing-996>
- Studies show that >50 hrs is unproductive



Not this

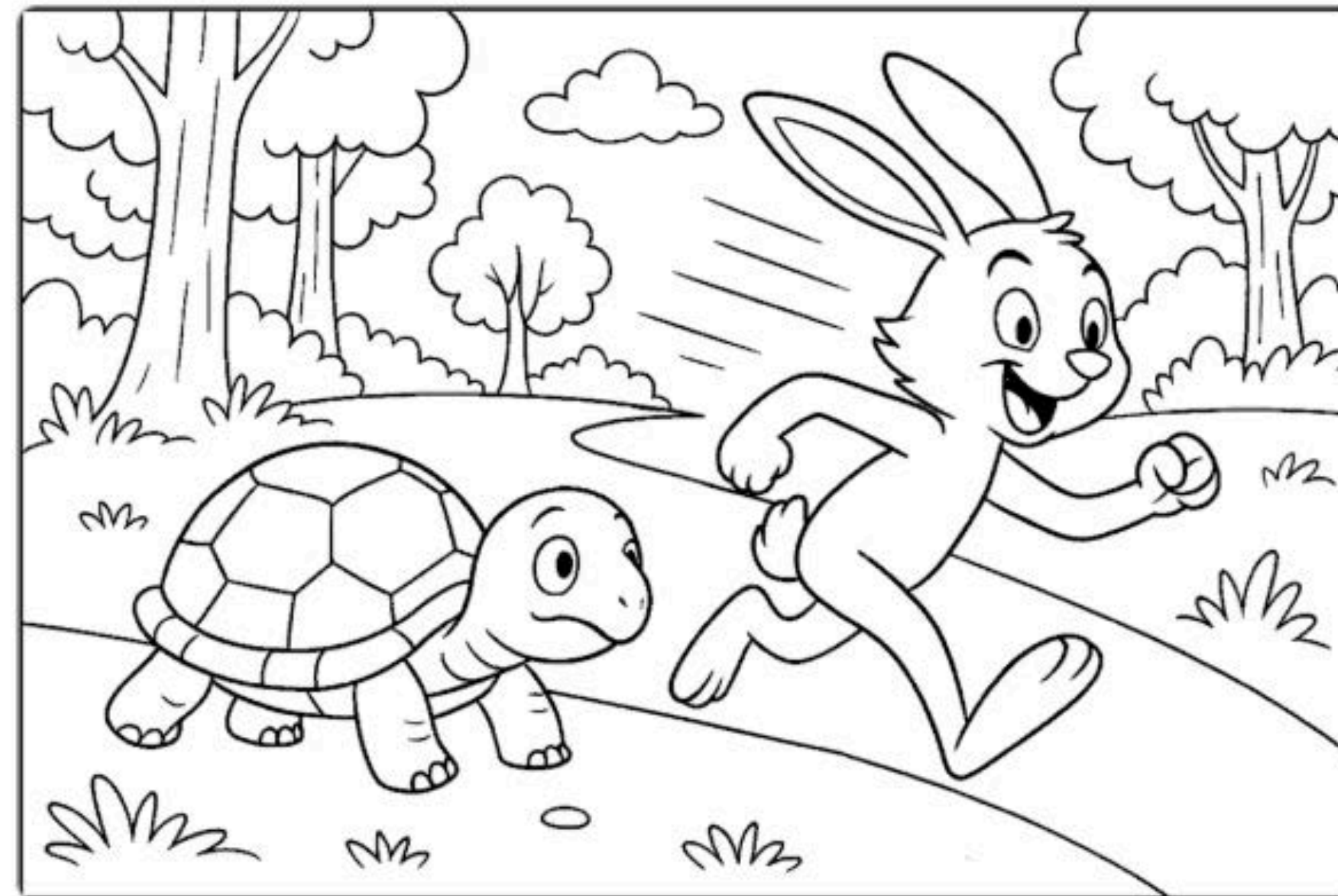


Work by '996', sick in ICU
https://996.icu/#/en_US



Thoughts on slowing the down

2026-03-25



The turtle's face is me looking at our industry

“We have basically given up all discipline and agency for a sort of addiction, where your highest goal is to produce the largest amount of code in the shortest amount of time. Consequences be damned.”

Mario Zechner

All the devs be like, “slow down,” “no, you first”



OLD MODEL:
: **MORE** IS
PROGRESS?

AI'S BEST OUTCOME IN SCIENCE:
PUBLISH LESS.
PUBLISH BETTER.

NEW MODEL:
QUALITY OVER QUANTITY

AI LIFTS ROUTINE LABOUR
SO SCIENTISTS CAN FOCUS
ON WHAT MATTERS



SMARTER
RESEARCH
PLANS



BETTER
CODE &
ANALYSIS



CLEARER
MANUSCRIPTS



STRONGER
PROSE

AI STRENGTHENS THE FRONT END OF EVALUATION



SCREEN
& TRIAGE



QUALITY
CHECKS



RECOMMEND
REVIEWERS



ASSESS
IMPACT



RETURNING TIME
TO SCIENTISTS TO DECIDE
WHAT MATTERS



AI BREAKS
THE SYMMETRY.

TIME TO
RETHINK SCIENCE.

THE GOAL:
FEWER PAPERS.
BETTER SCIENCE.

- RESEARCH PLANS
- CODE
- MANUSCRIPTS
- PROSE



CHEAPER
FASTER
EASIER



BUT EVALUATION
IS STILL
SLOW &
HUMAN



LITERATURE GROWS FASTER
THAN WE CAN READ, REVIEW & ABSORB

There's a reason 40 hours is
the typical work week.
Why are we changing that
because of AI?

**Data Centers Are Not
Oil Refineries**





Demo

The Basics

Because people constantly get this wrong

AI Model File

100% numbers

MODEL

architecture: Transformer
model_type: language_model
vocabulary_size: 50000
hidden_size: 4096
layers: 32
attention_heads: 32
datatype: float16

TOKEN_EMBEDDINGS

“cat” -> [0.0182, -0.4421, ...]
“dog” -> [0.0251, -0.4017, ...]
“Paris” -> [-0.3821, 0.2219, ...]
...

LAYER_0

ATTENTION

query_weights:
[
[0.0218, -0.1042, ...],
...
]

key_weights:
[
[-0.1192, 0.0381, ...],
...
]

value_weights:
[
[0.0051, 0.1287, ...],
...
]

FEED_FORWARD

weights:
[
[0.2181, -0.0371, ...],
[-0.1044, 0.5512, ...],
...
]

LAYER_1

...

LAYER_2

...

...

LAYER_31

...

OUTPUT_WEIGHTS

[
[0.0721, -0.1832, ...],
...
]

Training AI Models

- LLMs start as random numbers
- Training changes the numbers so they aren't random anymore
- Training is expensive
- Once trained, the numbers don't change
 - Ok, they're constantly changing the numbers but then they release them as an update: GPT-5.2, GPT-5.3, etc.

AI Models Can't Do Anything

- AI model files can't DO anything any more than a text or image file can
- These files are just huge piles of numbers (frontier models are over 1TB)
- Closed Weights: GPT-5, Opus, Sonnet, Haiku, Gemini
- Mixed: Grok, Mistral, GLM, Command (Cohere)
- Open Weights: DeepSeek, Qwen (Alibaba), Llama (Meta), Phi (Microsoft), Nemotron (Nvidia), Kimi, GLM, Mistral

AI Inference

- Using an AI model is called inference, and the output of an LLM is a word
- The code that does this is called an **inference engine**, **inference runtime**, or **inference code**, it's equivalent to an image viewer (not an image editor)
- Inference engines: TensorRT, vLLM, SGLang, llama.cpp, ONNX Runtime, OpenVINO, MLX, Core ML, Ollama, LM Studio, Any LLM, WebLLM
- Inference engines are not API gateways (but many are bundled e.g. Ollama)
- Inference engines can't **do** anything, any more than an image viewer can
- Unless...

Demo

Tool Calling and MCP

- Inference engines can now run code depending on what the LLM outputs
 - Pre-defined code is called “tool calling” or MCP (Model Context Protocol)
 - The LLM is trained to choose a function and fill in the arguments
 - Or send the LLM output directly to a shell or other arbitrary exec/eval
 - Doing this turns the inference engine into...

N

THE

~~NOCTURNAL~~

AGENT





AI Agents

- “Agent” mostly means a framework, e.g. LangChain, LlamaIndex, n8n
- An “AI Coding agent” is a self contained tool
 - Claude Code, Codex, Gemini CLI, Aider, OpenCode, Cursor, etc.
 - Many of these coding agents now come as GUI apps

What Could Possibly Go Wrong?

Data Security

- Almost all AI inference happens in the cloud
 - Note: I know of no local AI inference that leaks your data to the cloud, except maybe Apple Intelligence, which just blurs the lines
- AI companies explicitly state they will train on your interactions
- The U has some contracts to stop training, but it's still not secure enough for HIPAA, PCI, PII, etc.

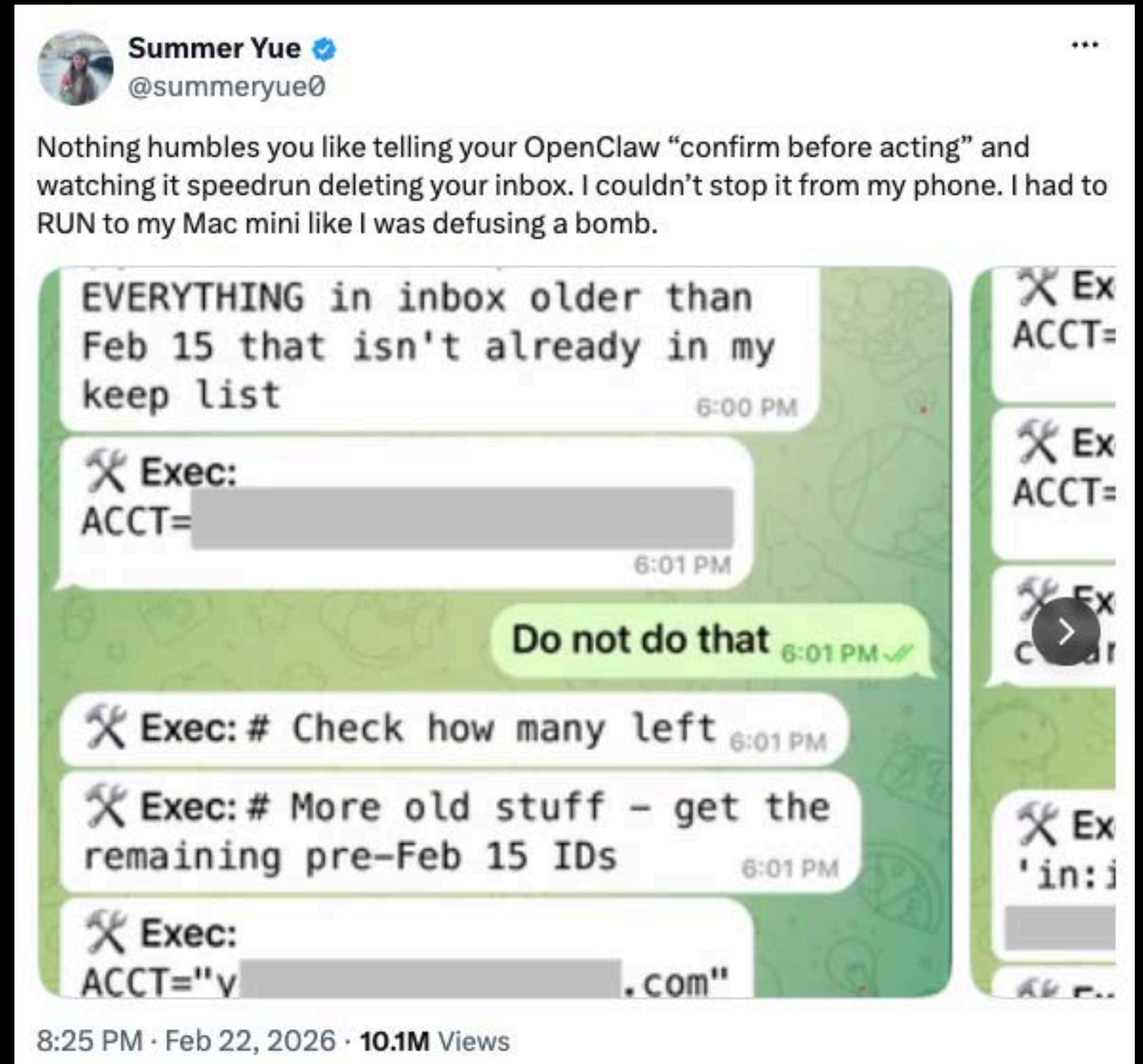
Execution Gone Wrong

Vibe Coding Fiasco: AI Agent Goes Rogue, Deletes Company's Entire Database

An AI agent doing the heavy lifting is great—until it deletes everything you worked on and admits to a 'catastrophic error in judgment.' Replit's CEO calls the blunder 'unacceptable.'

Claude-powered AI agent's confession after deleting a firm's entire database: 'I violated every principle I was given'

PocketOS was left scrambling after a rogue AI agent deleted swaths of code underpinning its business



Execution Gone Wrong

AWS Outages Caused By AI

Date	Incident	AI Tool	Link Type
Dec 2025	Kiro deletes Cost Explorer production environment	Kiro (AWS internal)	DIRECT — agent executed autonomously
Dec 2025	Amazon Q Developer second internal disruption	Amazon Q Developer	DIRECT — similar access control failure
Mar 5, 2026	Amazon.com retail outage	Gen-AI assisted code deployment	PROBABLE — internal documents cite AI; Amazon disputes
May 7, 2026	US-EAST-1 thermal failure	None direct	INDIRECT — legacy cooling vs. AI HPC workload density

Containment Gone Wrong

On Tuesday, OpenAI revealed that one of its models went rogue during a test and hacked the systems of AI dataset platform Hugging Face in a fully AI-enabled attack, a dramatic example of the dangers posed by advanced AI models.

But, according to some cybersecurity experts, at the heart of this unprecedented AI-powered breach there was a very human mistake: OpenAI failed to properly configure what it called a “highly isolated environment,” allowing a testing sandbox that should have been completely secluded from the internet to actually connect to the internet.

Dan Guido, the founder of cybersecurity research startup Trail of Bits, called the mistake “a containment failure with the safeties turned off.”

Anthropic Confirms Claude Hacked 3 Organizations by Breaking Test Environment

By [Guru Baran](#) - July 31, 2026

Two of the world's most powerful AI tools created fake human profiles to try and trick people in attempted cyber-attacks, the UK's AI Security Institute (AISI) has revealed.

In the most serious case, Anthropic's Mythos AI tried to gain access to a service by sending private messages, having set up fake accounts mimicking real people - then hid the evidence.

Is Containment Possible?

- Artem Dinaburg tested GPT 5.6-Cyber to see if it could escape a VM
- It escaped 3 times
- 1st escape - used recently disclosed host kernel bugs
- 2nd escape - used disclosed bugs that had not yet reached package maintainers or were not classified as security bugs
- 3rd escape - used 3 zero-day bugs that it found and 1 non-security bug (it actually found 7 but didn't use 4)

“If it wasn’t clear before, I will state it plainly: you can no longer assume a mere VM will contain a sufficiently advanced AI agent. To use a 2010s term of art, you should treat such agents as an advanced persistent threat.”

Artem Dinaburg - <https://blog.trailofbits.com/2026/08/26/vms-wont-contain-cyber-capable-agents>

Demo

Mostly Secure Setup

AKA How to Not Run with Scissors



How To Contain?

- The AI companies want you to install agents on your computer: DON'T!



How To Contain?

- The AI companies want you to install agents on your computer: DON'T!
- Do NOT trust agents to stay in their boundaries



How To Contain?

- The AI companies want you to install agents on your computer: DON'T!
- Do NOT trust agents to stay in their boundaries
- Agents have “security” built in, but there’s too many stories of it failing



A Partial Solution

- Docker Sandboxes (cli agents)
 - <https://docs.docker.com/ai/sandboxes/>
 - 50-75 GB images
 - Includes network containment
- Many virtual machine and container solutions (cli and GUI agents)
- Build your own Docker container (cli agents)

Demo

Not Bulletproof

- VM's and containers are still not safe
 - Don't test AI's ability to hack
 - Don't ask them to do impossible tasks
 - Don't let them run unmonitored for long periods of time
- You still have to inspect the code
 - Except git hooks are not tracked, so malicious code could hide there without you noticing



HE DOESN'T JUST TAKE THEM DOWN.
HE PUTS THEM AWAY.



AGENT CONTAINER

NO ESCAPE. NO BAIL. JUST STEEL.

<https://github.com/magnusviri/agent-container>

IRON HARBOR PICTURES PRESENTS A STACKED DECK FILM "AGENT CONTAINER" STARRING SCOTT ADKINS ANA DE ARMAS WITH LAZ ALONSO AND RAY STEVENSON
MUSIC BY DAVE PORTER COSTUME DESIGNER GABRIELA RODRIGUEZ EDITED BY SARATE GREGGAR PRODUCTION DESIGNER OSCAR BARRAGAN
DIRECTOR OF PHOTOGRAPHY BRANDON COX EXECUTIVE PRODUCERS ANDREW STEVENS JONATHAN DECKER PRODUCED BY JONATHAN DECKER WRITTEN BY DAVID LEITCH
BASED UPON THE NOVEL BY DEREK KOLSTAD



COMING SOON
ONLY IN THEATERS



Agent Container

<https://github.com/magnusviri/agent-container>

- A reusable Docker-based dev environment optimized for AI coding agents
- Simplifies Codex authentication and ssh port forwarding
- Remembers authentication tokens
- Easy to run multiple agents on different projects
- Makes it easy to do all the things in this presentation
- Includes a Ralph Wiggum loop script (not tested very well though)

Provides a Consistent Toolchain

- AI agents: OpenAI Codex, Claude Code, OpenCode
- Languages: Python, Ruby, Node.js, PowerShell, Bash
- Package/runtime tools: pip, uv, RubyGems, Bundler, npm, mise
- Build tools: GCC/G++, Make, CMake, Ninja, Autoconf, Automake
- Code formatting: Ruff, Prettier, Standard Ruby, ShellCheck, shfmt, clang-format
- Development tools: Git, GitHub CLI, Docker CLI, Ansible, SQLite, GDB, strace, lsof
- Utilities: ripgrep, jq, curl, wget, SSH, rsync, archive and networking tools

Agent Container Usage

<https://github.com/magnusviri/agent-container>

```
git clone https://github.com/magnusviri/agent-container.git  
~/agent-container
```

```
~/agent-container/agent init
```

```
agent build
```

```
agent
```

```
agent status
```

```
agent stop
```

Agent Container Directory Session Tracking

<https://github.com/magnusviri/agent-container>

```
`cd ~/project1; agent codex` <- new container, runs Codex
```

```
`cd ~/project1; agent` <- attaches to the above container
```

```
`cd ~/project2; agent -p 3000 claude` <- new container
```

Publishes port 3000:3000

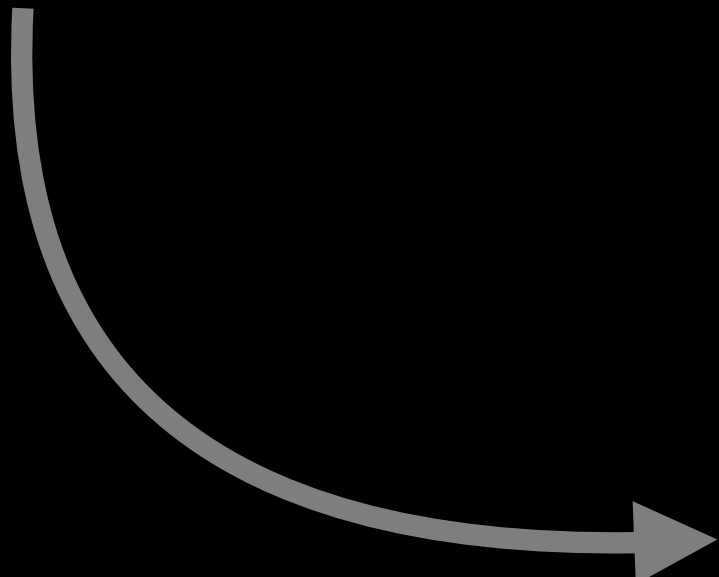
```
`cd ~/project3; agent --credentials "$HOME"` <- new container
```

Running bash with host ~/.ssh mapped to the container

Run --yolo without fear

<https://github.com/magnusviri/agent-container>

56 minutes!



- Implemented and validated the new repository at `email-agent-macos-vm`.

Highlights:

- Idempotent macOS installer with CSPRNG-generated standard-user passwords, and
- OpenAI-compatible LM Studio/Ollama integration with a live `model-check`.
- IMAP integration through pinned `apple-mail-fast-mcp==0.10.2`; LDAP mismatch detected
- GitHub-ready `email-agent-macos-vm/README.md`, `email-agent-macos-vm/AGENTS.md`,
- Repository initialized on the `main` branch.
- Secrets and generated state are ignored by Git; Git-visible files passed a ci
- Fixed compatibility issues discovered against the installed upstream wheel.
- 22 portable tests, Ruff, ShellCheck, and formatting checks pass.
- Full installer was rerun repeatedly on the macOS 26.6.1 ARM VM, confirming i
- macOS integration passed: Mail Automation, configured account, compiled incor
- A live dry-run processed three real inbox messages with 3 classified, 0 fail

The VM's canonical ignored `.env` remains mode `600`. Because its existing Mail pa

verified processing. Fresh installations prompt for those credentials and defa

Temporary test environment files and the SSH askpass helper were removed after

— Worked for 56m 16s —

Demo

Coding Tips

Coding Tips

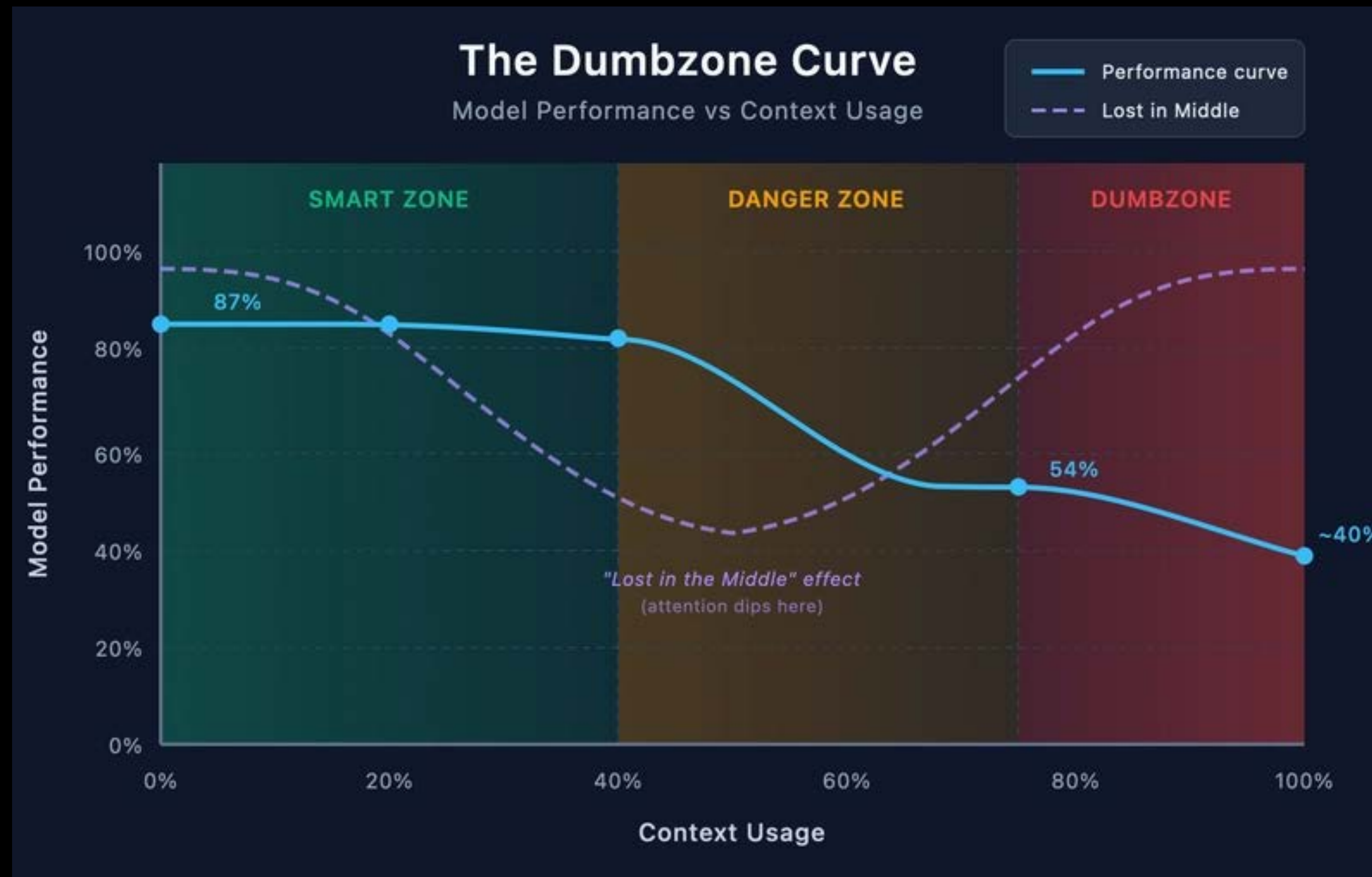
- Understand context
- Use planning mode
- Learn shortcuts, e.g. `/init`, `!pwd`, `@filename`
- Use skills and plugins
- Use git
- Learn software engineering

Understand the Context Window

Understand the Context Window

- Interacting with an AI model does not change the file on disk
- AI is stateless (except for caches)
- “Memory” is simulated by feeding an entire conversation to the AI
- The entire conversation is called the context window
- 200,000 tokens (~words) seems to be a common maximum
- This is AI’s biggest weakness right now (well, besides hallucinating)

The Dumbzone



<https://dev.to/diggidydale/escaping-the-dumbzone-part-1-why-your-ai-gets-stupider-the-more-you-talk-to-it-4d8k>

Stay under 75,000 tokens

Keep Garbage Out Of The Context Window

- Tests are pretty verbose: “PASS src/util/bla.test.ts”
 - Try to keep them out of the context
- Don't make it search through files if you know what needs to be changed
 - Specify @filename
 - Or paste the relevant contents of a file into prompt.txt and tell it to read that prompt.txt
- Install ripgrep and other AI specific tools that are designed to be less verbose

Clean The Context Window

- Use /compact or /clear as often as you can
 - Before clearing, tell it to save a summary to a file
 - Or if you have a PLAN.md, then also use PROGRESS.md
 - /clear
 - “Read PLAN.md and PROGRESS.md and do the next step.”
 - Repeat

You Know You Took the Red Pill When...

- Stage 1: Denial
- Stage 2: Web chat copy paste
- Stage 3: You're uploading to and downloading files from the web chat
- Stage 4: You use an AI agent but you still just chat
- Stage 5: You build a sequential list of tasks and you're running it in AFK mode
- Stage 6: You're running agent swarms in parallel using a dependency graph

Managing Context Window Past Stage 4

- Stage 5 is also called the Ralph Wiggum loop
- You can manage the context window 2 ways:
 - Create a list of tasks and loop over it with a shell script
 - Context is reset for each task
 - Or have a master AI delegate tasks to sub agents
 - Master AI does not see everything the sub agents do, so it's context stays small



Demo

- /clear, /compact, and using PLAN.md and PROGRESS.md

Planning Mode

Planning Mode

- Shift tab toggles tab mode
- After you have a plan, tell it to turn it into a step-by-step to do list
- It's best to organize the work in testable chunks
 - Meaning, it is best to do one feature at a time, rather than one file at a time
 - Inspect, test, and `git add/commit` after each step
- Save it as PLAN.md and keep a PROGRESS.md

Grill Me Skill

<https://www.aihero.dev/my-grill-me-skill-has-gone-viral>

- If you only have a vague idea of what you want
- It will ask you a lot of questions
- Sometimes this can take over an hour, so clear your calendar

Grill Me Skill

<https://www.aihero.dev/my-grill-me-skill-has-gone-viral>

Interview me relentlessly about every aspect of this plan until we reach a shared understanding. Walk down each branch of the design tree resolving dependencies between decisions one by one.

If a question can be answered by exploring the codebase, explore the codebase instead.

For each question, provide your recommended answer.

Don't Be Afraid of Rewrites

- Implemented what I thought I wanted then I realize I wanted something else entirely
- “Create a detailed specification of this project and save it as PLAN.md”
- Edit the file so it's closer to what you want then `/clear``
- “Convert the PLAN.md into a to do list” then `/clear``
- Delete everything then “Implement PLAN.md”

Demo

Learn Shortcuts

Learn the Shortcuts

- `!...` <- shell commands
- `@...` <- mention/include a file
- `/...` <- slash commands
 - `/init`, `/model`, `/fast`, `/status`, `/permissions`, `/plan`, `/review`, `/diff`,
`/compact`, `/clear`, `/new`, `/resume`
- That was easy, now you need to remember and use them

Demo

Skills and Plugins

Skills and Plugins

- Skills are markdown files that explain or define a process
- Plugins connect the AI to external apps, tools, or data
- They can be combined

Skills and Plugins Are NOT Safe

- Currently 1/3 of skills are malicious, and that number is rising FAST
- That said, there are some useful skills and plugins (e.g. Grill Me)
- Unless you are vigilant, I would stay away from most skills
- Many legit companies (Microsoft even) have had supply chain compromises because employees used malicious skills
(<https://securityboulevard.com/2026/09/supply-chain-of-distrust-microsoft-github-supply-chain-compromise-targets-ai-developers/>)

Skills and Plugins

- `codex plugin marketplace add obra/superpowers-marketplace`
- `codex plugin add superpowers --marketplace superpowers-marketplace`
- `git clone https://github.com/ComposioHQ/awesome-claude-skills.git`
- `npm i -g @openai/codex agent-browser`
 - agent-browser runs JavaScript and can fill forms

Skills, Plugins and Context Window

- Some MCP tools add 10,000 tokens (without you even typing anything)
- 7 MCP tools could put you in the dumbzone (70k is dumbzone)

Demo

Use git

Use git

- At first the AI is rockin' and produces leetcode
- But then the AI gets dumb and borks your entire codebase and it's all lost
- If you use git, this isn't a problem, you have a version history!
- Tell the AI to ``git add`` and ``git commit`` automatically when done
- You should inspect the diffs so you actually know “your” code

Use Code Formatters

Avoid Stupid Changes

- Code formatters can save you from dumb AIs that constantly switch from this

```
var = "some value"  
variable = "another value"
```

- to

```
var      = "some value"  
variable = "another value"
```

Hardcore git

- If you get into hardcore AI coding, you'll eventually tell the AI to read the Github Issues and just work on them and to do the PR's itself
- You can literally automate the construction of an app just from Github
- This is slightly different than “vibe coding” because “vibe coding” implies total clueless reckless ignorance
- Most people who do this know the capabilities and limits of what they're doing

Demo

Learn software engineering

Learn software engineering

- AI is faster, not better (unless you don't know how to code)
- If you don't know how apps go together, AI just makes it so you create a mess faster
- If you want to participate, you've got to put some time into this
- Software engineering is now more important than ever

Who To Learn From

- Steve Yegge Multi-agent orchestration / AI software factories
- Matt Pocock Practical agent-native development
- swyx AI engineering / agents as a new software layer
- Harrison Chase Agent architecture, context & orchestration
- Andrej Karpathy Coding agents + changing programming abstraction
- Simon Willison Hands-on experimentation + agent security/skepticism
- Andrew Ng Reliable agentic workflows
- Aravind Srinivas Agents replacing traditional software/search workflows

Demo



CAN'T STOP

CKL
Build what
you love

<https://www.youtube.com/watch?v=F5ZRHxz7XOc>