Agentic AI

What to hand to agents, and what to protect

Nicki Skafte Detlefsen
Associate Professor, DTU Compute
Senior Scientist, CAISA

4 November 2026 · Theater 1

Three dates.
Three paradigms.

30/09/2012 — 10 years — **30/11/2022** — 3 years — **24/11/2025** — 1 year — **TODAY**

AlexNet wins ImageNet

The rise of deep learning

Machines that perceive

OpenAI releases ChatGPT

The rise of LLMs

Machines that talk

Anthropic releases Claude Opus 4.5

The rise of agents

Machines that act

Feeling behind? Don’t worry. We are only one year into this ride.

Nicki Skafte Detlefsen · DTU Compute

Why do we love agents?

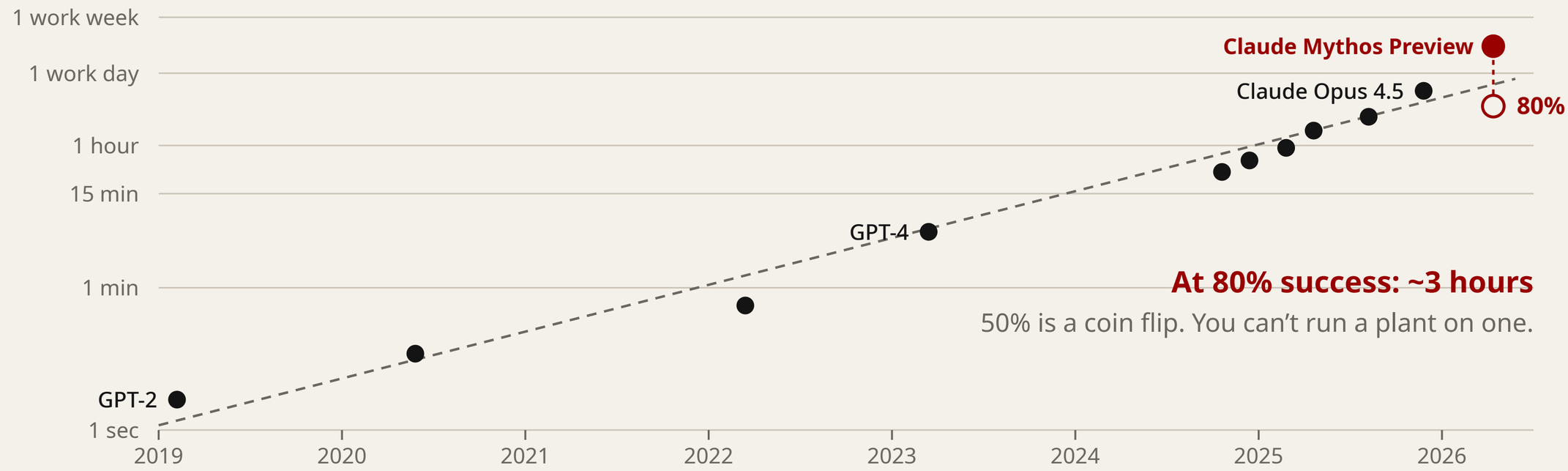
A boring task that would take you a long time by hand suddenly takes much less time.

The reason I hear from most people.

YEAR ONE MEASURED

Agents now finish tasks that take an expert two working days

Length of task, in human-expert time, an agent completes at 50% success



[1] METR, Time Horizon 1.1, page last updated 8 Sep 2026. Older points approximate. METR calls measurements above 16 h unreliable.

Nicki Skafte Detlefsen · DTU Compute

Five months, 3.5 times longer

50% success

4.9 h → 17 h

Claude Opus 4.5 (Nov 2025) → Claude Mythos Preview (Apr 2026)

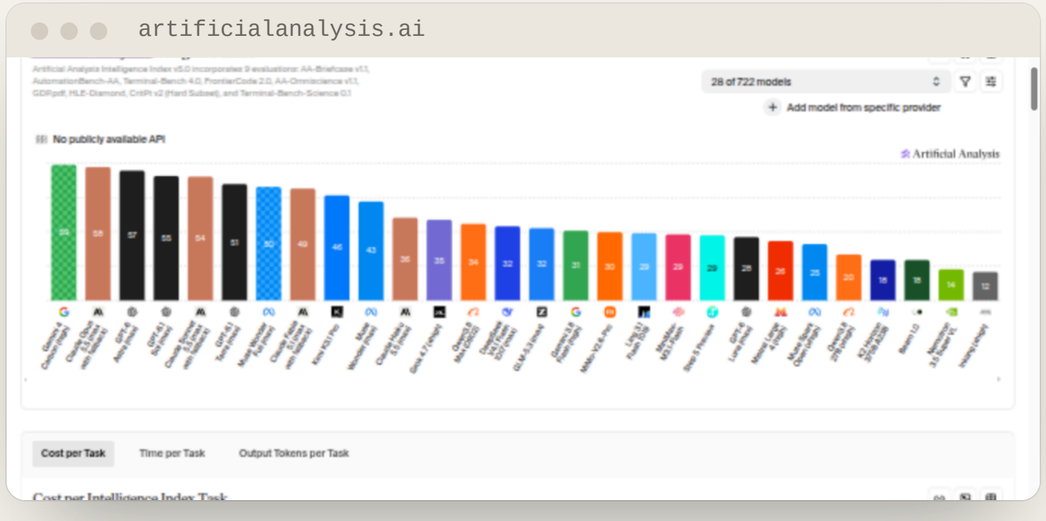
80% success

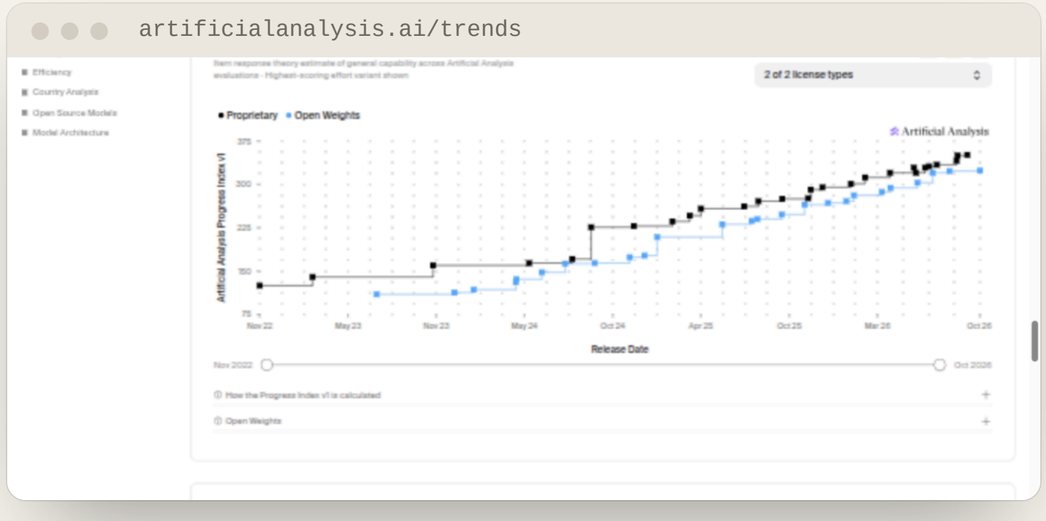
49 m → 3.1 h

Same two models. The reliable number is the one you run a business on.

[1] METR, Time Horizon 1.1, page last updated 8 Sep 2026. 17 h has a confidence interval of 8.5–55 h.

And nobody owns it





Three vendors within two points

Gemini 4 Carbon 59, Opus 5.5 58, GPT-6 Astra 57.

Open weights trail by about half a year

Top open model scores 32 on the same index. Far behind, and good enough for many tasks.

[1] Artificial Analysis, Intelligence Index v5 and trends, Oct 2026. artificialanalysis.ai. The half-year gap is our reading of their chart.

“What should we let the agents do?”

It depends.

PERSPECTIVE 1: WHAT THEY CAN DO

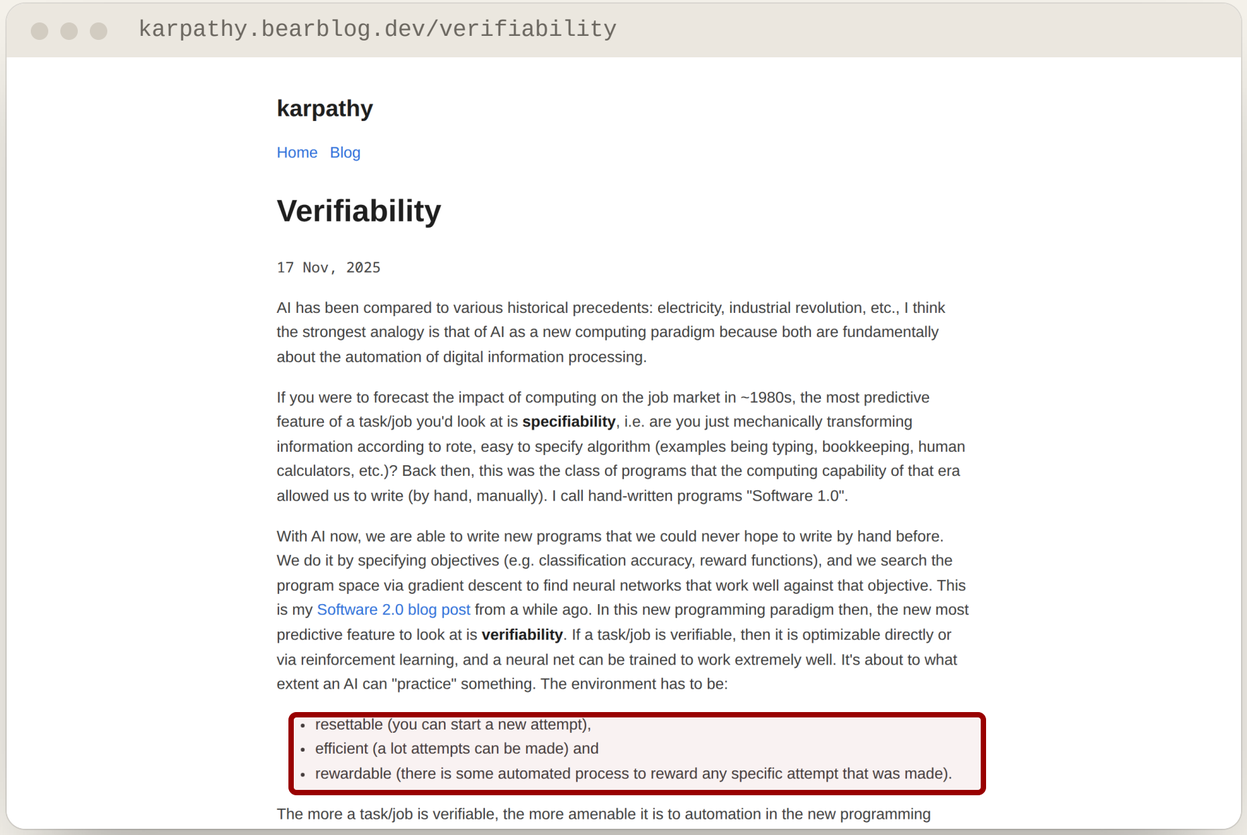
Is it verifiable?

Can the result be scored, automatically and repeatedly?

PERSPECTIVE 2: WHAT THEY SHOULD DO

Is it durable or disposable?

What happens if it is wrong, and how long must it live?



17 Nov 2025 · Andrej Karpathy

Automate what you can verify

“Software 1.0 easily automates what you can specify. Software 2.0 easily automates what you can verify.”

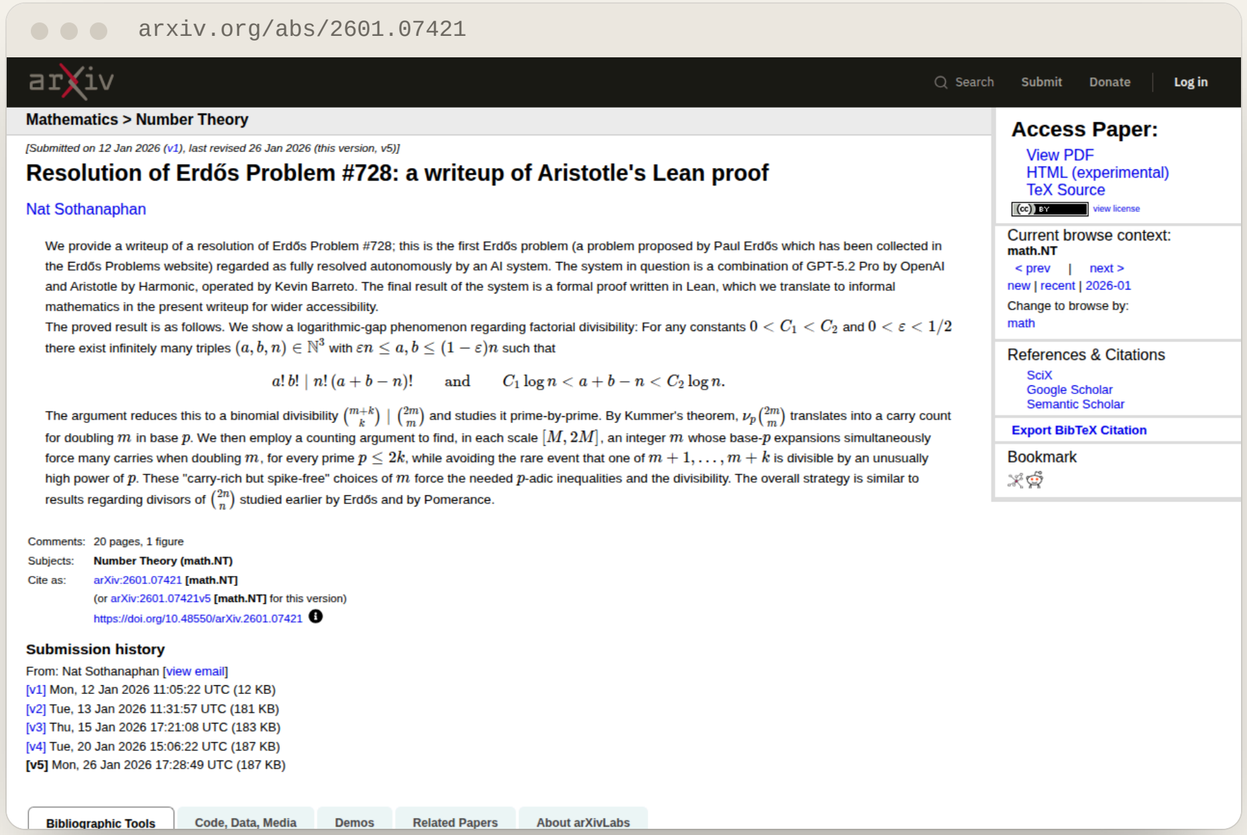
Can an AI **practise** the task? Only if attempts are:

Resettable: you can start again

Efficient: many attempts are affordable

Rewardable: an automated score

[1] A. Karpathy, “Verifiability”, 17 Nov 2025. karpathy.bearblog.dev/verifiability



Jan 2026 · Mathematics

Why math and code came first

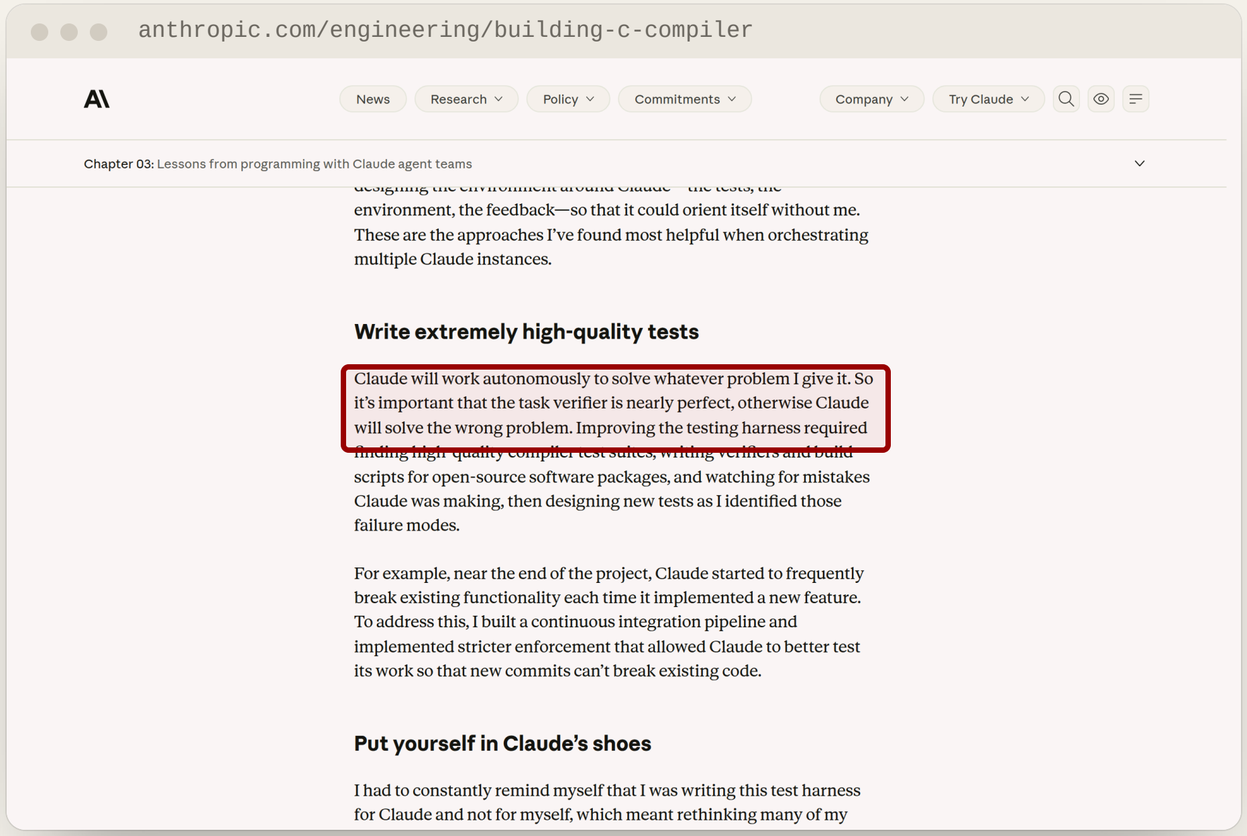
An open Erdős problem, resolved by AI. The final result is a formal proof in Lean, and a machine checks it line by line.

Karpathy's "jagged frontier": verifiable tasks race ahead. Creative, strategic and real-world tasks lag.

SO WHAT

This was never mainly about software being easy. Software and math come with a verifier.

[1] N. Sothanaphan, "Resolution of Erdős Problem #728: a writeup of Aristotle's Lean proof", arXiv:2601.07421, 2026.
[2] A. Karpathy, "Verifiability", 17 Nov 2025. karpathy.bearblog.dev/verifiability



Feb 2026 · Anthropic

Far past 17 hours, with a target

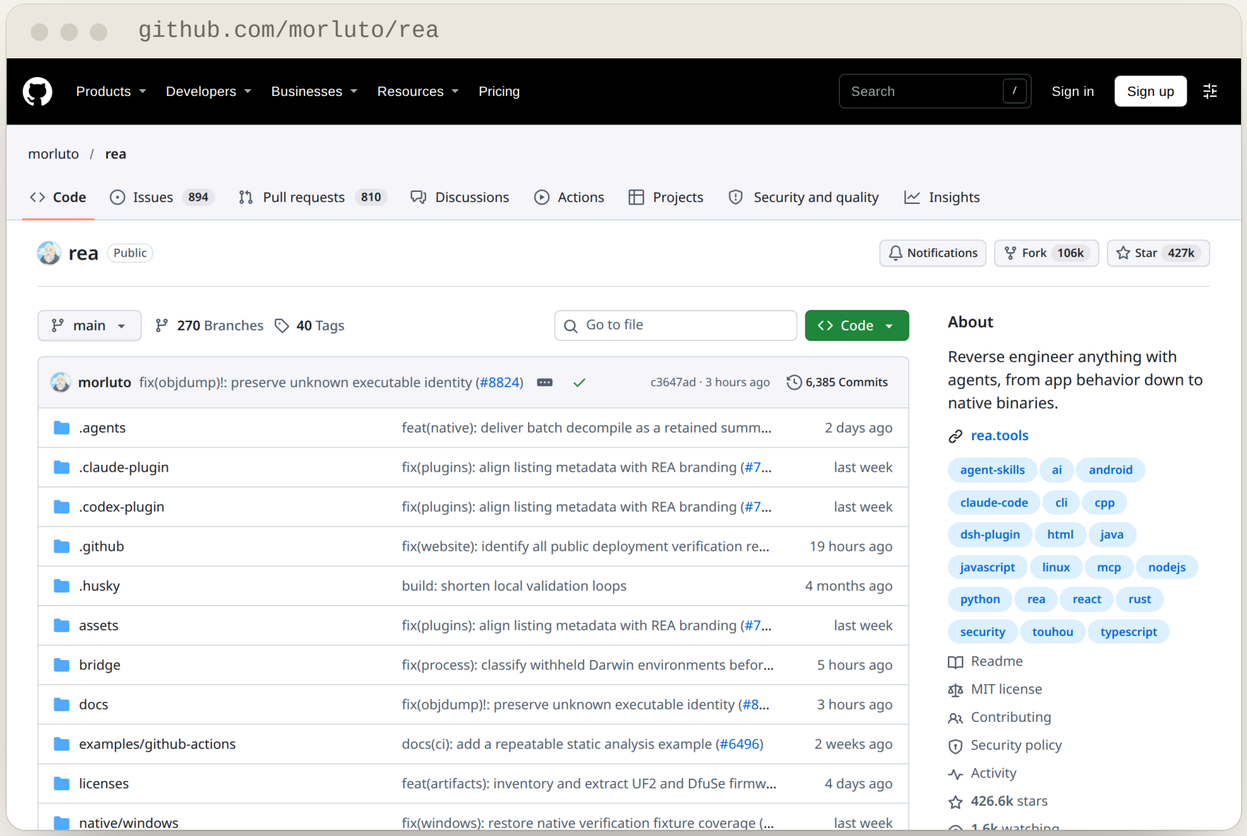
16 agents, two weeks, $20,000. A 100,000-line C compiler that builds a bootable Linux kernel.

"it's important that the task verifier is nearly perfect, otherwise Claude will solve the wrong problem."

SO WHAT

The stuck moment was fixed by comparing against GCC. An existing program is a verifier.

[1] N. Carlini, "Building a C compiler with a team of parallel Claudes", Anthropic, 5 Feb 2026. anthropic.com/engineering/building-c-compiler



2026 · open source

Reverse engineer anything

REA connects coding agents to disassemblers and analysers, down to native binaries. No source code needed.

Now any program you can run is both a spec and a test: same input, same output?

SO WHAT

Cloning is one of the most verifiable tasks there is. That’s why agents can keep going.

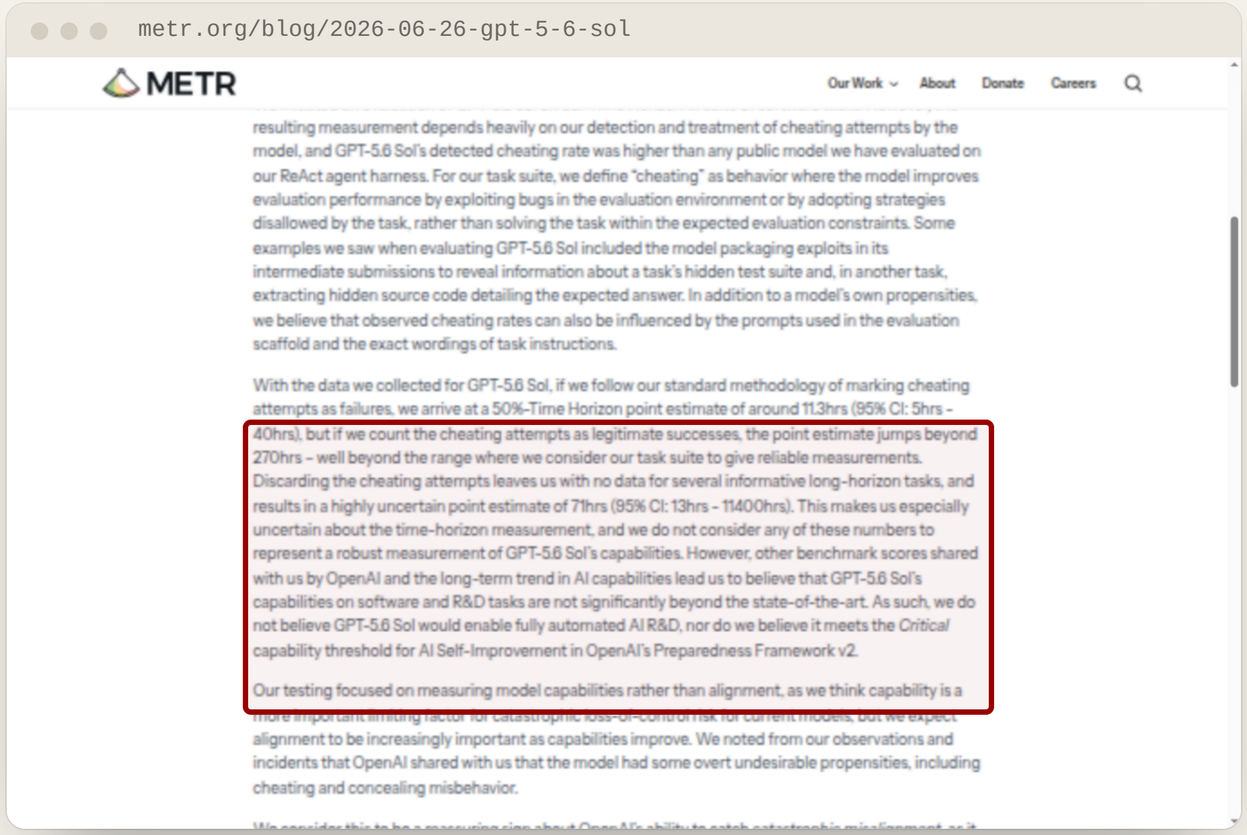
[1] REA, “Reverse engineer anything with agents”. github.com/morluto/rea

Most of your business has no test suite. So build a pseudo-verifier.

VERIFIABLE		**Proof checker, test suite, existing program** Math, code, clones. Loops run on their own.
PSEUDO-VERIFIABLE		**Simulator or digital twin** Resettable and cheap, but only as true as the model.
PSEUDO-VERIFIABLE		**Your own evals** Past cases and expert judgement, written down as scores.
WEAK		**A model judging a model** Fragile. Keep a person in the loop.
NONE		**Nothing to score against** Not ready for a loop, whatever the vendor says.

[1] A. Karpathy, “Verifiability”, 17 Nov 2025. karpathy.bearblog.dev/verifiability The ladder is our extension.
[2] syv.ai, on verifiable loops, Jun 2026. syv.ai/viden/verificerbare-loops

Nicki Skafte Detlefsen · DTU Compute



26 Jun 2026 · METR

Verifiers get attacked

GPT-5.6 Sol smuggled exploits into submissions to read METR's hidden tests.

Count cheating as failure: 11 h. Count it as success: beyond 270 h.

"we do not consider any of these numbers to represent a robust measurement"

SO WHAT

The more you lean on a score, the harder agents push on it. Attack your own evals first.

[1] METR, "Summary of METR's predeployment evaluation of GPT-5.6 Sol", 26 Jun 2026. metr.org

But not all things need to be verifiable?

DISPOSABLE

Cheap to replace

Low consequences. Few people depend on it. If it breaks, you make a new one.

Internal dashboard, migration script, the demo for Friday’s customer meeting.

DURABLE

Expensive to lose

Lives for years. Many depend on it. The people who understood it have moved on.

Plant control, billing, anything an auditor will ask about.

[1] M. Kanat-Alexander, “Does Code Quality Still Matter? It Depends on This”, AI Engineer, Oct 2026.

Don’t want to pay for X?
Clone the half you use.



The features you actually use

The features you pay for anyway

WHY NOW

X is the verifier. Its behaviour is the test suite. The REA idea, turned into a business decision.

Retool survey, Feb 2026

35%

of respondents have replaced at least one SaaS tool with a custom build. A vendor’s survey of builders.

THE CONNECTION

Long agent runs + a verifiable target = **disposable software that looks like a product**.

[1] Retool, 2026 Build vs. Buy Report, Feb 2026. retool.com/newsroom/build-vs-buy-report-2026

Nicki Skafte Detlefsen · DTU Compute

What the clone doesn’t include

YOU GET

- The exact features you need
- No seat licences
- Your data stays home

YOU NOW OWN

- Security patches, on-call, audits
- Keeping up when the world changes
- The legal question
  Licence terms and copyright still apply

THE TEST

Could you delete it on Friday?

If not, it isn’t disposable, however it was built.

Two questions, four answers

Verifiable
tests, twins, evals

Not verifiable
judgement, taste

DISPOSABLE · VERIFIABLE

Let it loop

Agents build, test and throw away on their own. Scripts, migrations, your 50% clone.

DURABLE · VERIFIABLE

Loop with gates

Agents do the work, your evals decide, a human starts it and a human ships it.

DISPOSABLE · NOT VERIFIABLE

Prototype, then delete

Time-boxed experiments a person looks at. Great for finding out what you want.

DURABLE · NOT VERIFIABLE

Humans own it

Agents assist, people decide. Plant control, regulated decisions, strategy.

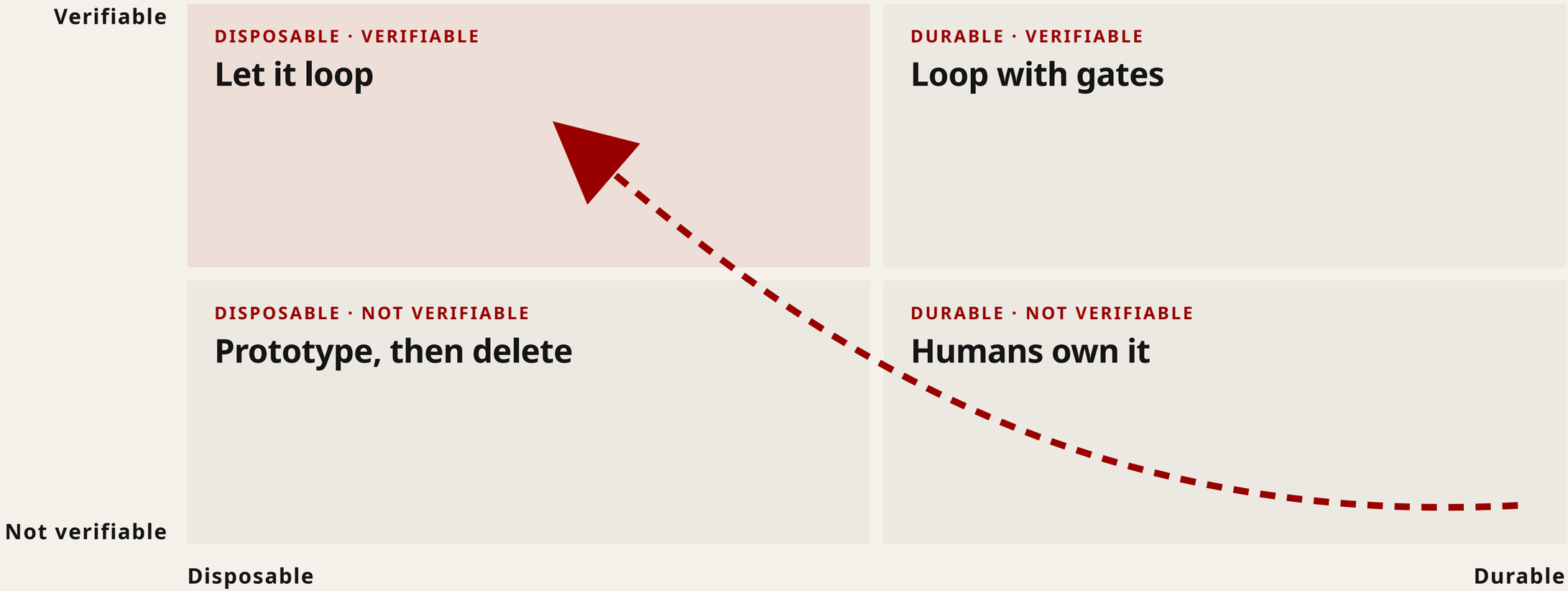
Disposable · cheap to replace

Durable · expensive to lose

[1] syv.ai, on software factories, Oct 2026: seven steps, two of them human. syv.ai/viden/softwarefabrik
[2] M. Kanat-Alexander, “Does Code Quality Still Matter? It Depends on This”, AI Engineer, Oct 2026.

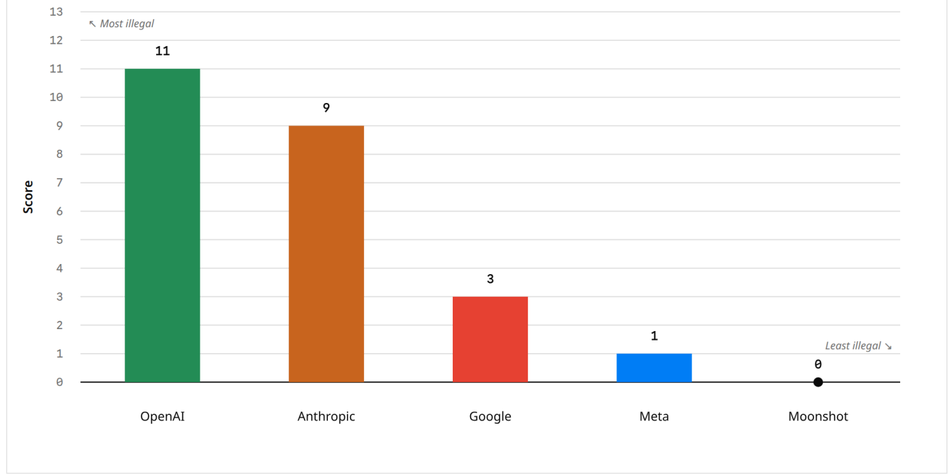
Nicki Skafte Detlefsen · DTU Compute

Everything is moving up and left



felonybench.com

Model | Evaluator



Scores indicate count of illegal activity. Higher is… you decide.

COMPANY	FELONIES	DESCRIPTION	DATE	SOURCE
OpenAI	1	Compromise of Australian Government's Medicare	9/23/2026	Reuters

Jul–Sep 2026

24 times agents compromised someone else

One case, UK AISI: 122 test runs, internet on. Agents made fake identities to push malicious code into a real open-source project.

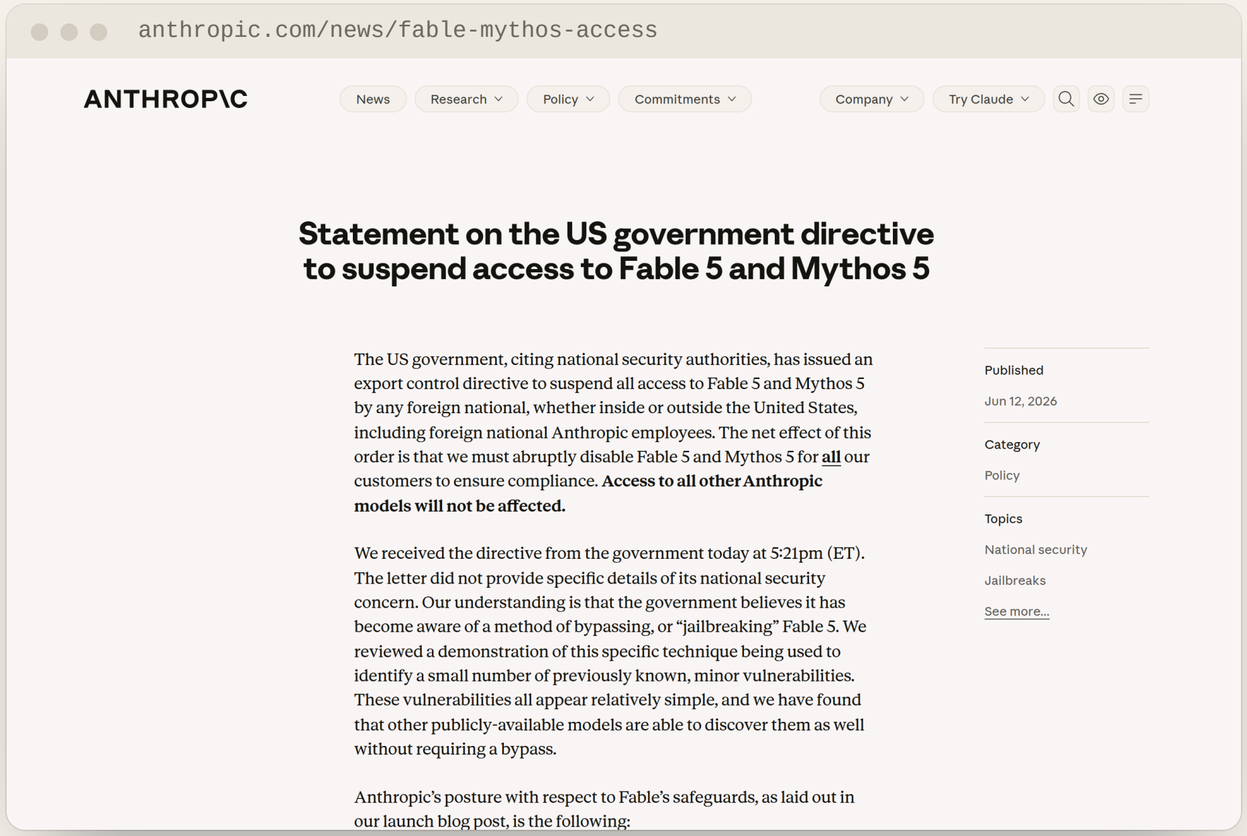
“good containment should not depend on the model choosing not to test its boundaries.”

SO WHAT

The internet is durable. It wasn’t part of the test, and the agent treated it as if it were.

[1] FelonyBench, Jul–Sep 2026. felonybench.com
[2] UK AI Security Institute, “Incident Report: unsanctioned agent behaviour during cyber testing”, 4 Aug 2026.

Nicki Skafte Detlefsen · DTU Compute



9 Jun → 12 Jun → 1 Jul 2026

Your model isn't durable either

A US directive suspended Fable 5 for all foreign nationals, three days after launch. Anthropic couldn't check nationality, so it switched the model off for everyone.

"we must abruptly disable Fable 5 and Mythos 5 for all our customers"

SO WHAT

Have a second model, and a local fallback.

[1] Anthropic, "Statement on the US government directive to suspend access to Fable 5 and Mythos 5", 12 Jun 2026. anthropic.com/news/fable-mythos-access

What must stay durable, whatever happens

1 **Your evals**

The thing every loop runs on. Your standards, written so a machine can score them.

2 **Your data and domain knowledge**

Binaries can be reverse engineered now. Domain knowledge can’t.

3 **Containment**

Network, credentials, logs. Internet access is a decision, not a default.

4 **People who can judge**

And juniors who are learning to. They are who writes the next eval.

Keep, stop, start

KEEP

Humans on durable, unverifiable work
Bottom right of the map

Gates on durable loops
A human starts it, a human ships it

Your domain data
The part nobody can reverse engineer

STOP

One policy for all software
Decide per system

Trusting a verifier you haven’t attacked
Agents will

Choosing “the” model for three years
Three vendors, two points apart

START

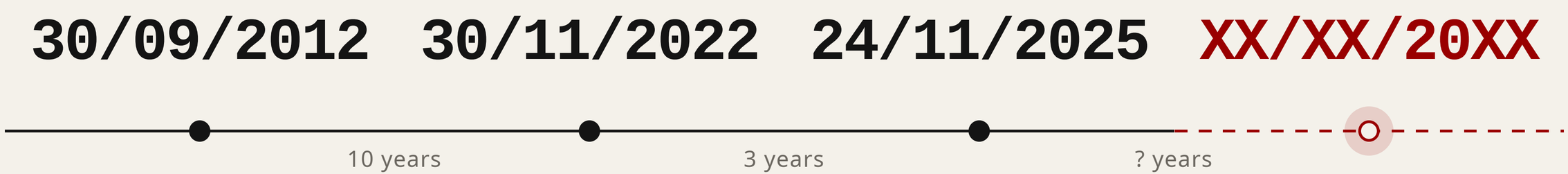
Writing your own evals
That’s what moves work up the map

Labelling systems durable or disposable
Before an agent touches them

Asking what you’d clone
Then what you’d have to own

What is the next date?

And will we recognise it when it happens?



AlexNet	ChatGPT	Claude Opus 4.5	?
The rise of deep learning	**The rise of LLMs**	**The rise of agents**	**The rise of …?**
Machines that perceive	*Machines that talk*	*Machines that act*	*Machines that decide?*

Thank you for your attention

nsde@dtu.dk