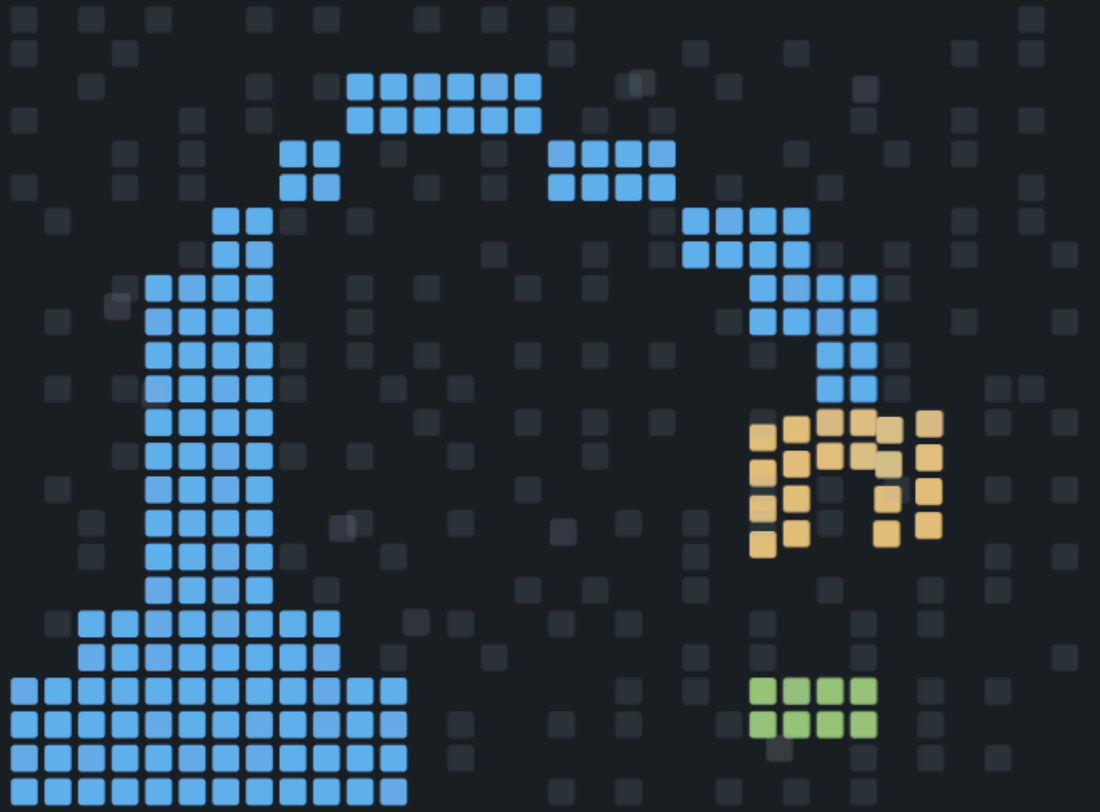
You can build this yourself.

Small teams, real tools, less waiting on vendors.

Khalid Omar · khalido.dev



build a robot arm

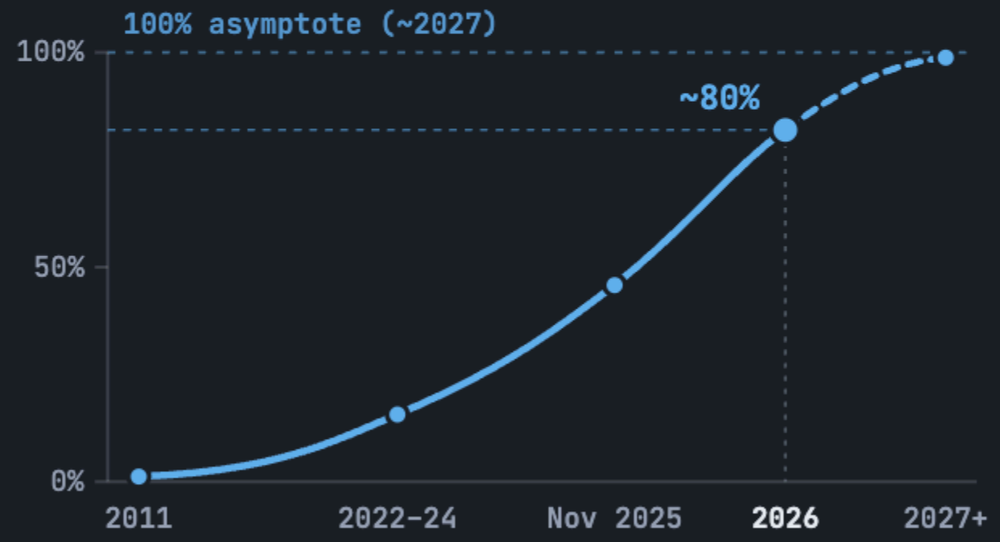
2011 → 2026

Software is eating the world — exponentially

Two separate shifts: AI share approaching 100%, while total software output explodes past the human baseline.

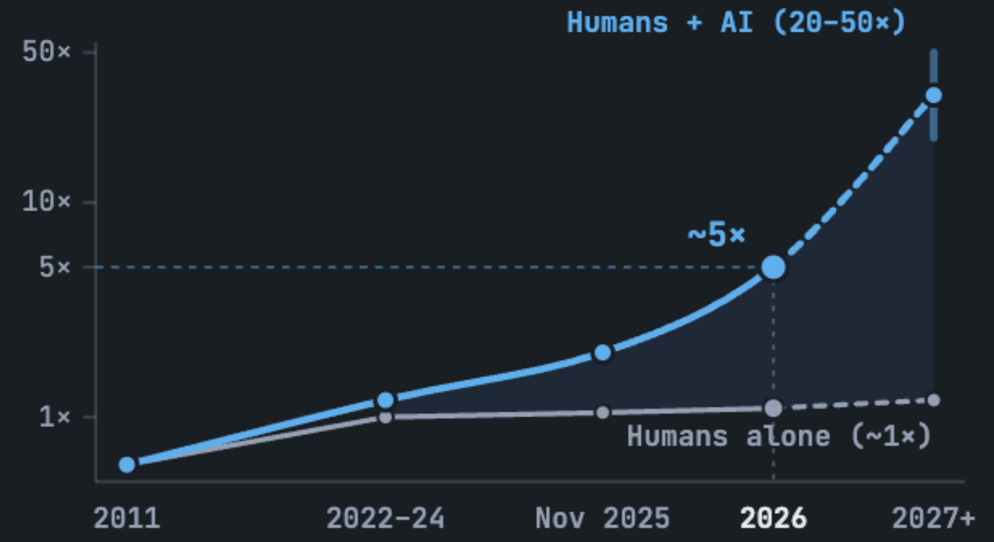
Share of new code written by AI

illustrative, caps at 100%



Total code produced (vs. 2024)

illustrative, human trend vs. AI

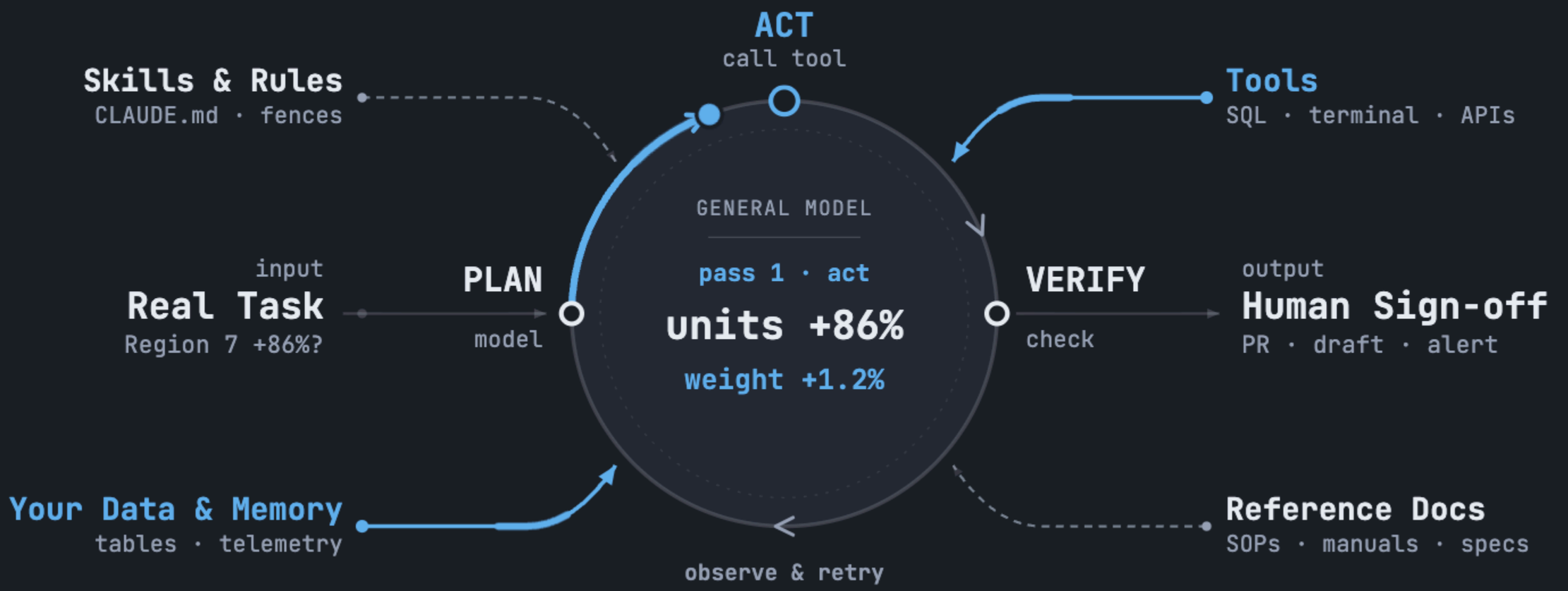


2026 **AI ~80%** **Total ~5×** Most new code AI-written; total code ~5× (whole runtimes rewritten).

REF: Marc Andreessen — Why Software Is Eating the World (2011) ↗ Anthropic — Claude Fable & Mythos 5.2 (Nov 2026) ↗ Simon Willison — Three LLM inflection points ↗ METR Long Tasks ↗

WHAT "AGENTIC" MEANS

What an agent is



REF: Anthropic — Building Effective Agents ↗ Anthropic — Components of Agent Memory ↗ Claude Code — Code Mode ↗ Armin Ronacher — What is Codemode ↗ Code with Claude 2026 Keynote (YouTube) ↗

INSIDE BIG FIRMS

2–3 people doing the work of dozens

> **Ethan Mollick** ✓
> @emollick · Wharton Professor
>
> Thing I am hearing from a number of big firms, from manufacturers to financial services: we have scattered teams of 2-3 people using AI that are doing the work of dozens of folks in a fraction of the time, creating large-scale management fights over the implications and politics.
>
> 2026 · X — View post on X ↗

2–3 people **10×–100× output** **Politics first**

EMPIRICAL EVIDENCE, NOT HYPE

Anthropic Research (2026)

Inside an AI-native company, and across 58 enterprises.

Inside Anthropic, Aug 2026

6 : 1

6 hours of AI agent work for every 1 hour of human work.

58 enterprises, ~300 use cases

2 in 3

Companies buying AI capacity instead of hiring headcount.

Non-technical staff coding-agent sessions: 2 → 90 / month

REF: Anthropic — Working with AI (Nov 2026) ↗ Anthropic — Employment at Anthropic (Oct 2026) ↗ OpenAI — How Orgs Use AI (8.3× gap) ↗ OpenAI — Work at the Frontier (43.5% crossover) ↗

ANTHROPIC · WORKING WITH AI (NOV 2026)

High-value AI means changing the workflow



Bottom-left, low value

Chatbots & slide decks don’t move the needle

Bottom-middle, the traps

Static exec dashboards & ad-hoc apps flop

Top-right, where value lives

Re-engineer the workflow & own it

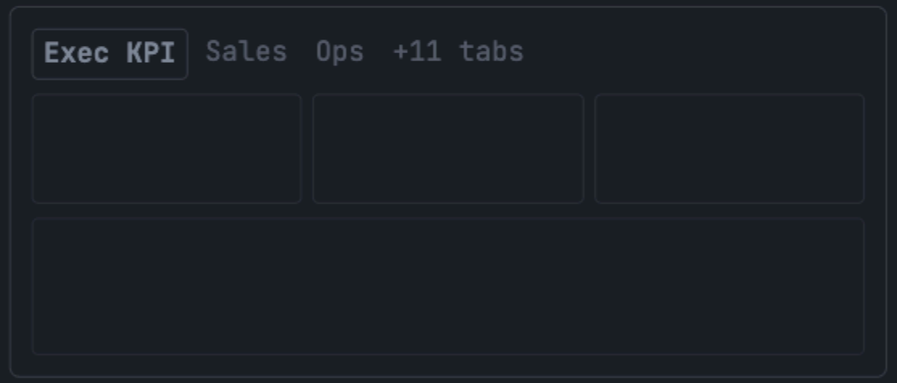
WHY OLD-SCHOOL BI IS DYING

Nobody opens your dashboard

Human attention is the bottleneck — shift from static BI tabs to exception alerts + live DB follow-ups.

Old model: pull — 4/100 value (Anthropic)

Static Executive Dashboards



42 charts, 14 tabs — Last opened: 47 days ago

Great analysts build them. Overloaded execs don't have the headspace to check them before it's too late.

New model: push + drill-down — ● Alert pushed 08:14

> Only sent when something moves, to the decision owner
>
> **Region 7 gross margin dropped -2.4% overnight (-$14.2k run-rate)**
>
> Unit count jumped +86% while pallet weight stayed flat (+1.2%). Likely billing-unit shift, not real demand.

Exec follow-up

"Which distributor drove the margin drop?"

```
> SELECT dist_id, SUM(margin_delta) FROM orders_may GROUP BY 1
```

Dist-04 (South) switched SKU-118 billing from cartons to single packs on May 12 — accounting for 91% of the variance.

"Did we see this last quarter too?"

```
> SELECT week, AVG(units / NULLIF(pallet_kg,0)) FROM shipments_q1_q2
```

No — Q1 unit-to-pallet ratio held at 12.0x across all 6 distributors. First drift started May 12 06:40 UTC.

Agents watch metrics 24/7 → humans make timely decisions.

Things I built with AI

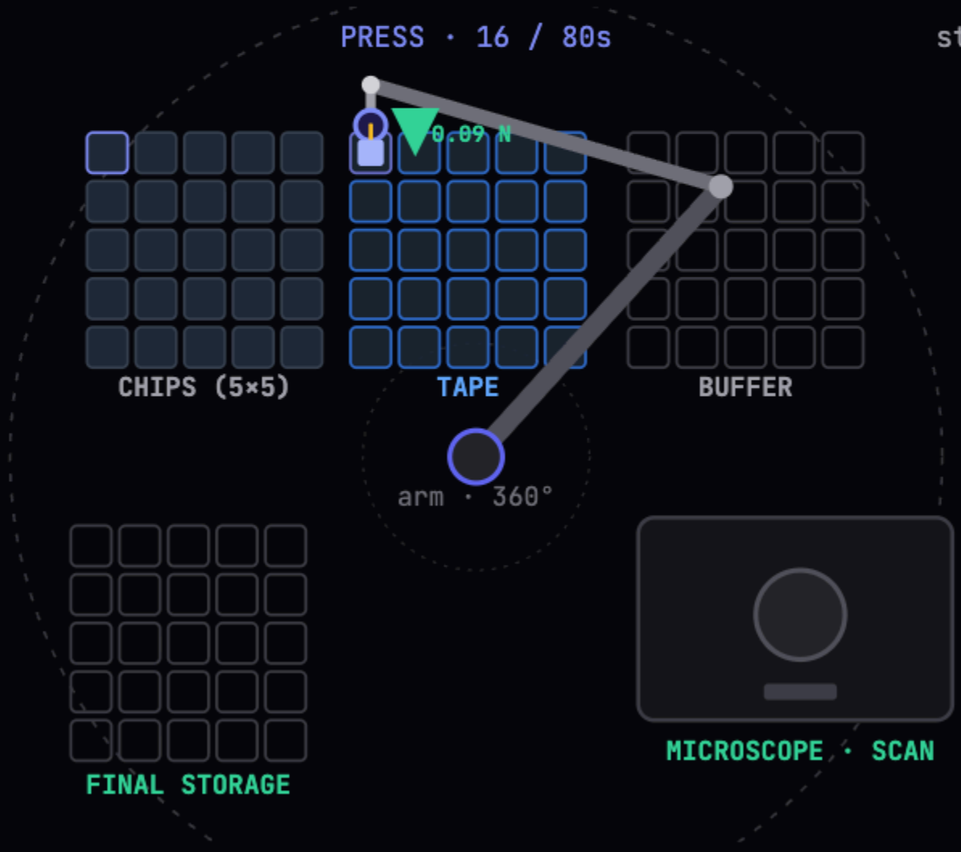
REF: Simon Willison — Code is cheap now ↗

CASE 01 · QUANTUM MATERIALS LAB

A robot arm that exfoliates 2D materials

Simulation of the real cell: pick, press, peel, scan.

Arm — pressing at set force (compliant tip + load cell)



Live telemetry — PRESS

force	peel	angle
0.09 N	0.12 mm/s	4.2°
dwell	**TCP (mm)**	**suction**
16 s	298,58	ON 35kPa



Run log, BO loop — trial 1/30, chip 0/25

Target: hBN 10–40 nm — Tape: 35 µm

awaiting first flake…

Optimiser: Condition 1 · rep 1/5 — running at 1.82 N…

CASE 01B · CUSTOM CLI + LITERATURE ENGINE

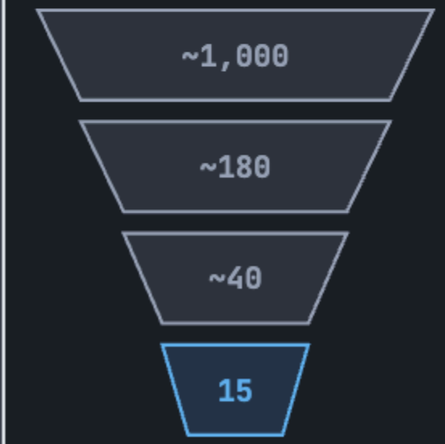
1,000 papers → the 15 that matter

Read everything relevant to your work.

papers 1,000 → 15

1,000 → 15

98.5% noise filtered · read all the things



arXiv · OpenAlex · Exa
pulled by the CLI

Embedded in a vector DB
kept: closest to our sea

AI reads each paper
scored against lab hardw

In the robot code
as parameters with citat

ko arxiv | ko papers recipes → robot code

```
$ ko arxiv "2D hBN exfoliation peel" | ko papers --filter lab-setup
```

tape_um = 15–65 µm arXiv:2411.06891
#1 yield predictor; Bayesian opt hits peak in 23 trials vs. ~485 random

peel = 0.1–0.2 mm/s · 25 °C arXiv:2403.12901
Slow PDMS release beats heating (anneal |r| ≤ 0.18 → stay room temp)

oxide = 90 / 285 nm SiO_2 arXiv:2605.18407
Target 10–40 nm hBN flakes (min_area ≥ 100 µm²) via CIE-Lab contrast

vision = MaskTerial arXiv:2412.09333
+40% AP50 on 1–3L hBN trained from just 5–10 microscope images

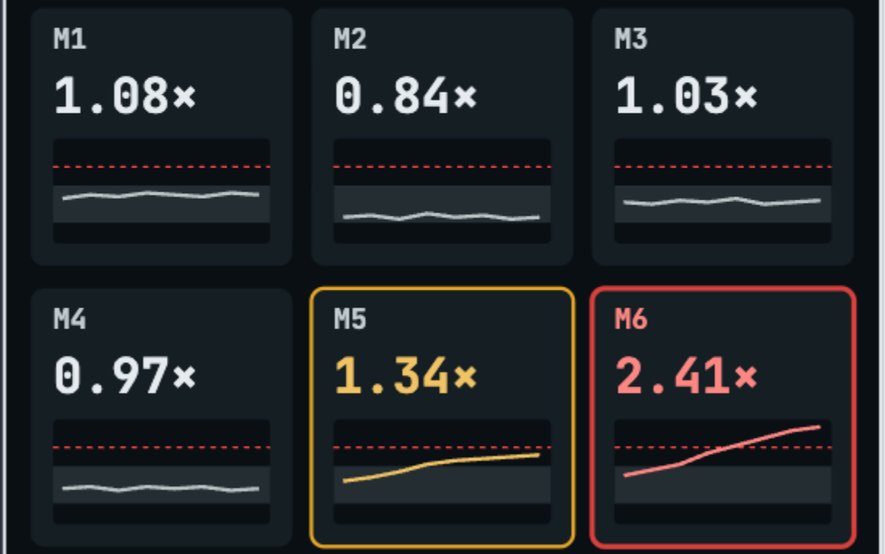
> **Robot UI help popover:** Default 0.1 mm/s @ 25 °C — fast peel grips PDMS, slow peel releases; heating adds zero yield gain.

CASE 02 · FACTORY TELEMETRY

Agents watch all the things

line monitor — 300+ points · live

~30 machines online in a day



load vs. the other machines; red line = alarm

agent — watching 300+ points

ALARM M6 load 2.41× its neighbours

DATA load climbing for a day, temperature drifting with it

BOOKS 16 textbooks + the drive manual, for that part only

CASE 03 · MEDICAL SERVICES MINI-ERP

Make AI do simple tasks

AI reads the photos, fills the form.

100s of photos dozens of fields ~1 hour saved per person

site_photos/ 56 of 100s



qc_form dozens of fields

Make · model
MedScan Optima 64-Slice

System serial · built
SN-884120 · Aug 2019

X-ray tube
Performix 8MHU

Slice count
412,380

Software
v14.2

Options
Cardiac · Angio · Dual-Energy

Functional checks · verdict
a person decides

RIGHT TOOL FOR THE DATA SHAPE

Cheap smarts, plain code

You can now run everything through an AI.

inbox — 8 unread

- urgent fault — **Probe error E-42 before morni…**
- grey-zone invoice — **Is Tuesday's recalibration co…**
- vendor blast — **Q4 parts catalog & shipping c…**
- double charge — **Invoice 2291 was charged twice**
- odd noise — **Scanner makes a new noise, st…**

→

decide(email) — Jev · 118 ms

category: choice

service_desk	**0.94**
billing	0.04
contracts	0.00
ops	0.00
noise	0.02

| sure enough = 0.85, else a person looks

→

route.py

```
if top < 0.85:          human review
elif service_desk:      service agent
elif billing:           billing agent
elif contracts:         contracts agent
elif ops:               ops agent
elif noise:             archive
```

plain_code.py

machine numbers: no model, plain code

```
if load_ratio ⩾ 1.55 and delta_15m > 6.0: alert_whatsapp(machine_id)
```

MY PERSONAL BOT

Build what you'll use every day

printer 06:45

MON 06:45

WEATHER	22° · sunny
BUS	07:42 · 8 min
SCHOOL	Maths · PE · Art

sample data

the bot about 20 tools

> connects to every source a family uses (getting there)
> logs in where there is no API, like the school portal
> answers the useful questions on the family WhatsApp
> fills out forms and books things
> searches the web, saves what you send it
> prints each person a brief, weekdays at 06:45

AI never makes up a fact.

Code writes deterministic output. The agent can paraphrase it in family-friendly speak.

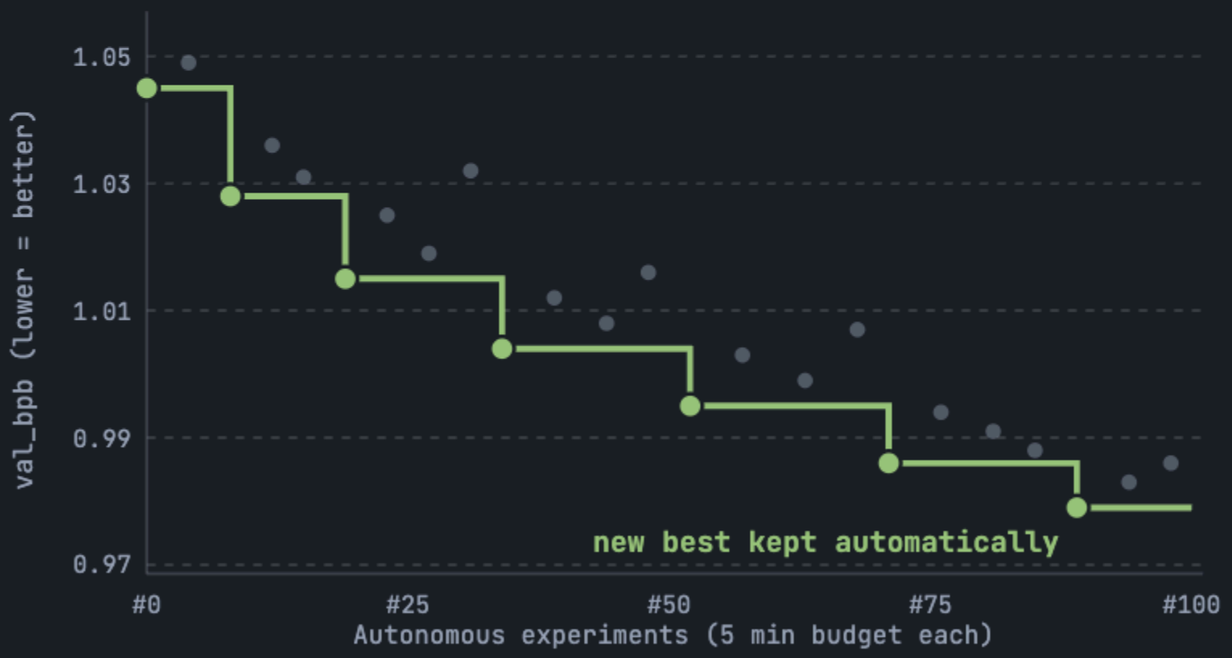
CLOSED-LOOP EXPERIMENTATION

Autoresearch: see idea, ship idea overnight

Karpathy introduced the idea in March 2026. Days later Shopify's CEO made their page rendering 53% faster.

The loop: illustrative run, ~100 experiments overnight

● kept (git commit) ● discarded (git reset)



tobi lutke ✓
@tobi · CEO, Shopify

OK, well. I ran /autoresearch on the liquid codebase. 53% faster combined parse+render time, 61% fewer object allocations.

This is probably somewhat overfit, but there are absolutely amazing ideas in this.

Mar 12, 2026 · 1M views View post on X ↗

Shopify/liquid#2056 974/974 tests pass ↗

parse + render
7,469 → 3,534 µs (-53%)

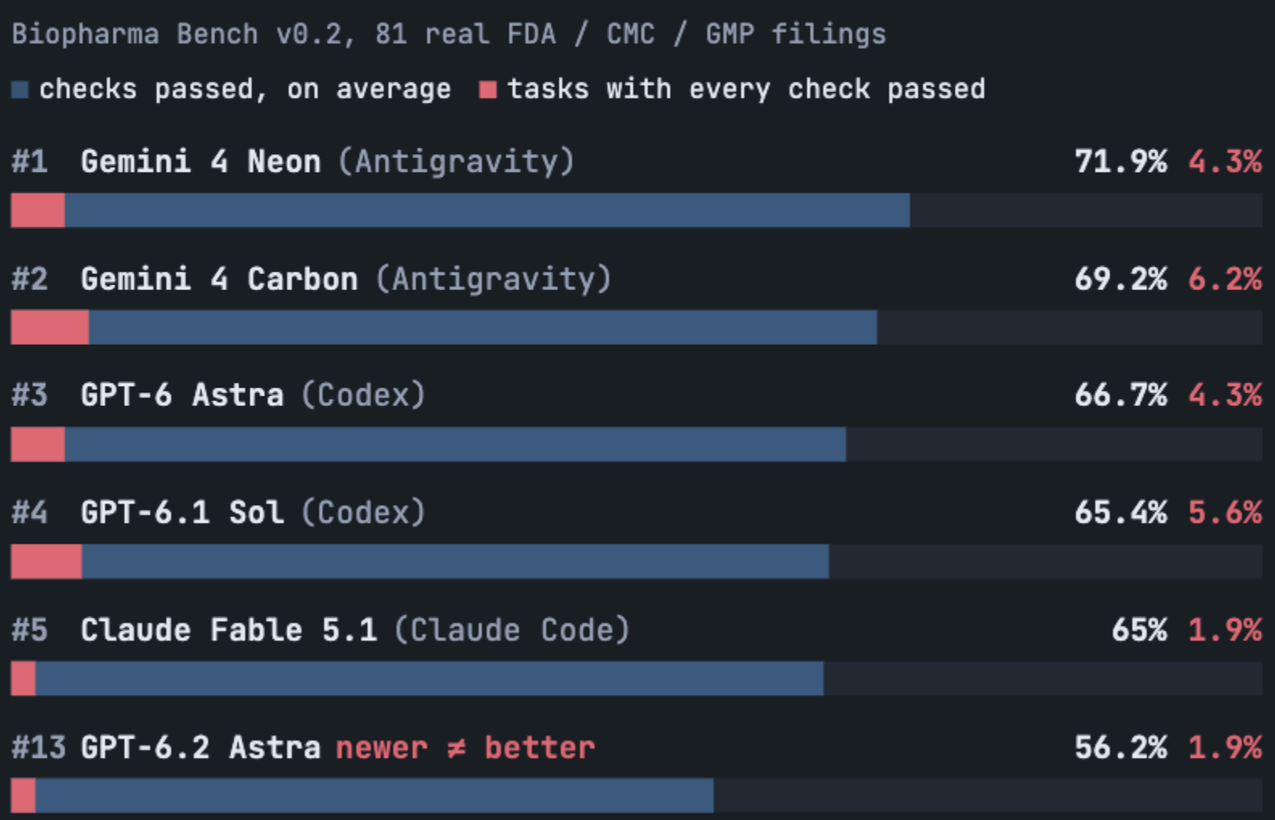
allocations
62.6k → 24.5k (-61%)

REF: karpathy/autoresearch ↗ Tobi Lütke on X (Mar 2026) ↗ Shopify/liquid#2056 ↗ Simon Willison — Liquid speedup ↗

REGULATED DOMAIN PROOF · REALITY CHECK

72% mostly-right isn't shippable

Biopharma Bench v0.2: 81 real FDA filings, CMC validations & GMP audits across 19 frontier models.



Each task is a real filing, validation or audit, graded check by check by regulatory veterans. 19 models, 3,078 runs.

The best model ticks

71.9%

of the checks on a task, on average

But gets a whole task right

4.3%

of the time: every check passed, ready to file

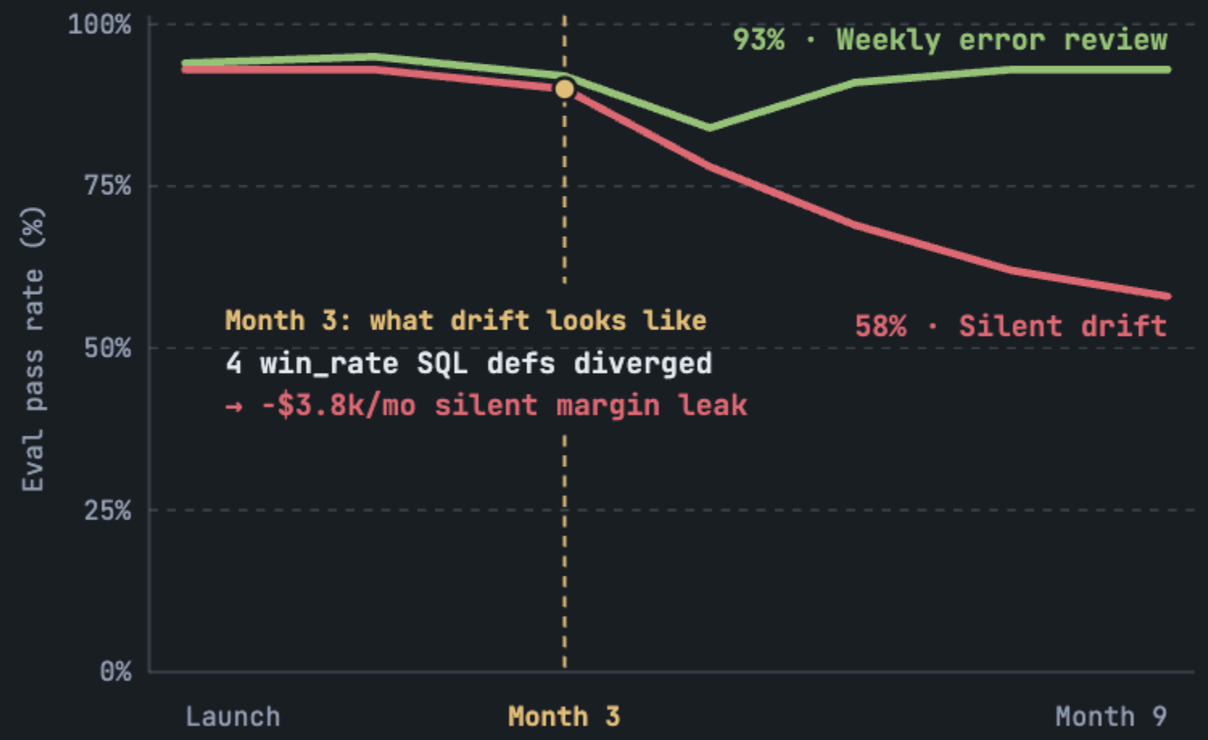
Best of any model: 6.2%

EVALUATIONS & DRIFT

There is no static pipeline

If you are not evaluating it continuously, you are flying blind on yesterday's world.

Pipeline pass rate over time, weekly review vs. set & forget (illustrative)



Hamel Husain, Evals FAQ

60–80%

of eval work is humans reading raw failures.

Binary pass / fail first

Beats 1–5 scales. 8 of 12 public evals had broken graders (Eugene Yan).

Measure pass^k reliability

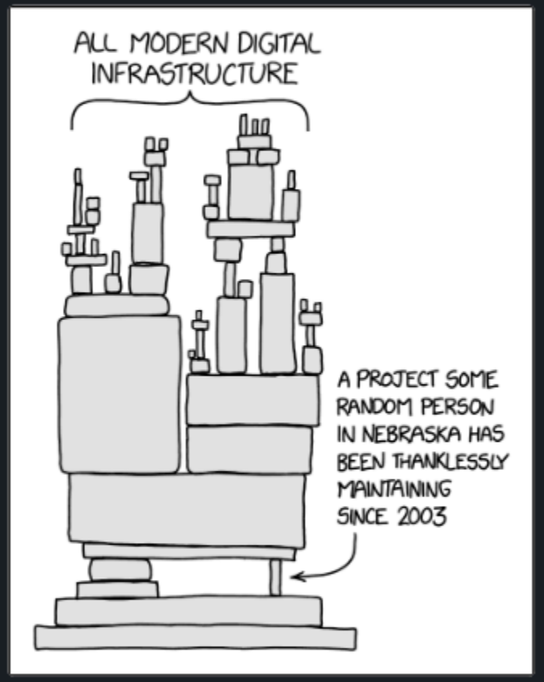
Average pass rate hides flaky agents — inspect failures before buying any vendor tool.

OLD-FASHIONED SOFTWARE ENGINEERING NEVER WENT OUT OF STYLE

All Code Is Slop (Except Mine)

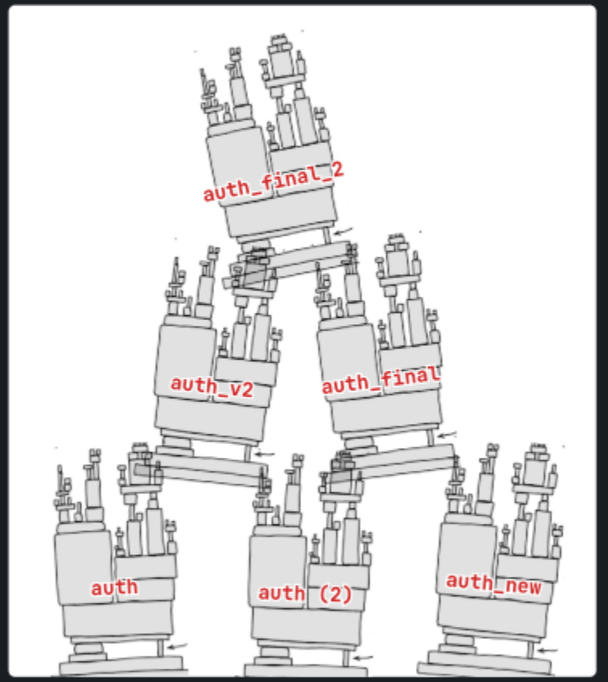
AI builds fast. Left alone, even a very smart model builds a mess.

Software, before AI



xkcd 2347 · CC BY-NC 2.5

Unconstrained AI



after xkcd 2347 · illustrative

1 Checks
lint, types, tests, permissions

2 Verification
run it, look at the result

3 Prose rules
a line in the prompt file

REF: xkcd 2347 — Dependency ↗ Andrej Karpathy — the original "vibe coding" post (Feb 2025) ↗ Lauren Tan — All Code Is Slop (YouTube) ↗ PocketOS incident (AI Incident Database) ↗ No AI Slop: A Manifesto ↗ UC San Diego Study (arXiv:2512.14012) ↗

START WITH ONE LOOP

Don’t overcomplicate your setup

Don’t repeat your org chart in agents. Models keep absorbing the scaffolding.

Are you solving Navier-Stokes or the Riemann Hypothesis?

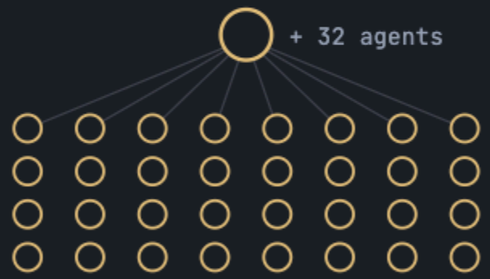
NO? MOST BUSINESS WORK



One loop: tools, docs & skills

Anthropic: folding subagents back into one loop is often more accurate and cheaper.

YES? VERY HARD PROBLEMS



A swarm: R&D, parallel work

Replit: 32 agents hit **18.9×** on independent NumPy kernels. Another model got worse.

REF: Anthropic — When to fold subagents into your main loop ↗ Replit — Throwing headcount at the problem ↗ OpenAI — Navier-Stokes solution (Sep 2026) ↗ Antigravity — Swarms ↗

BUY THE PLUMBING. OWN THE NEXT CHANGE.

Can you fix Tuesday's bug without calling the vendor?

Why old-school black-box consulting breaks in the AI age — and the three living maps your team needs.

Black box, vendor SOW

6 weeks

Log ticket → change request → wait for the contractor who built it.

Glass box, 3 living maps

Same afternoon

Live trace: Tuesday 09:00 bug — 10-min map per system

Process map — Tuesday 09:00 · Region 7 Ops

Dashboard shows +86% order spike, but pallet weight is +1.2%.

Data map — feeds/floor_count.csv → sql/unit_check.sql

May 12 feed switched cartons → single packs (×12 unit shift).

Software map — transforms/orders.py → tests/test_pack_ratio.py

Added unit guard + regression test · PR merged Tuesday 14:15.

Incentives, not villains — retainers reward opaque hand-offs.

verified in git

REF: Tim O'Reilly — What Happened to Silicon Valley? (Oct 2026) ↗ Anthropic — Working with AI ↗

FORWARD-DEPLOYED ENGINEERS, FROM INSIDE

Deploy your own people

Outsiders hit unwritten company context. Your insiders already have it.

Atlassian: “everyone is an AI builder”

1 month

until they brought the job titles back

“Throughput’s not our problem; direction’s our problem.” Sherif Mansour

	Builder 2–3 curious insiders	Owner runs the process today
Problem	Maps the live system in 10 minutes	Flags real daily friction & broken numbers
Build	Ships CLIs, small diffs, security sweeps	Brings unwritten floor & domain rules
Truth	Wires automated checks, swaps models	Owns the ground truth: machines, filings

REF: Lenny’s Podcast — Sherif Mansour (Atlassian, Nov 2026) ↗ OpenAI — OpenWorks ↗ Anthropic — Working with AI ↗

REAL OWNERSHIP

What real ownership looks like

Can someone on your team trace a number on a dashboard all the way back to the source, fix it, understand what the fix means, and update the eval?

REAL ARTIFACT · PHYSICS PHD → PRODUCTION PRS

Same software lifecycle. AI does the typing.

one change, start to finish — 2 of 6 steps need a person

1	2	3	4	5	6
Idea or bug	**Read the code**	**Small change**	**Checks**	**Review**	**Deploy**
describe the real friction	trace screen → API → database	search first, tiny edit, on a branch	types, tests, evals	read the diff, approve or reject	logged, reversible
person	AI	AI	automatic	person	automatic

SECURITY, MAINTENANCE & PILOT COST

Security, bus factor & economics

Practical boundaries for compliance, second-person maintenance, and 30-day pilot cost.

Repo security sweep — deepsec-style



auth/webhook.ts — HIGH: Auth bypass

Unverified signature fallback on legacy callback route.

1. regex scan → 2. agent trace → 3. revalidate

Model proposes; code decides

Zero-retention APIs or VPC; secrets out of Slack; read-only evals.

Automated full-repo security sweeps (e.g. deepsec) catch legacy auth flaws & leaks.

No single point of failure

Boring open-source stack (SQL, Python, TypeScript, Git) + living system maps.

Right-sized models + hard spend caps so any second engineer picks it up fast.

30-day shadow pilot

Test on live rows before committing 7 figures to a multi-quarter SOW.

Keep permission NOT to use AI if plain SQL or code wins.

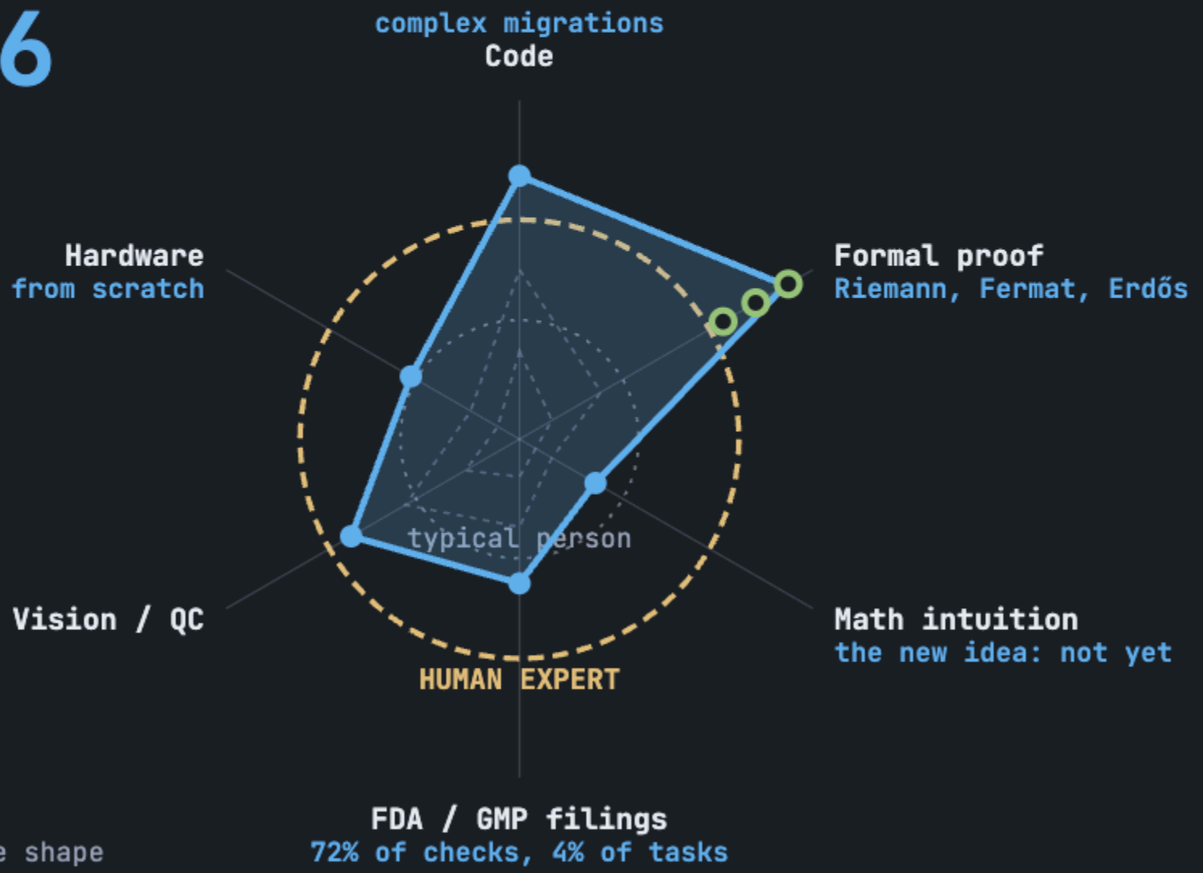
REF: Vercel — deepsec (full-repo agent security sweeps) ↗ Simon Willison — External Sandboxes (Nov 2026) ↗ Anthropic — Working with AI ↗ The Grug Brained Developer ↗

WHY WATCH THE FRONTIER

The jagged frontier is expanding faster than expected

I'm not a mathematician — watching frontier teams pushes how I use AI on everyday work.

2026



illustrative shape

Nov 2026, formal math

Riemann Hypothesis verified in Lean

17M lines across 175,187 modules (Levent Alpöge + Claude), certified across 5 independent checkers in 24 hours.

Terence Tao & 25 Fields Medalists

"Solving problems is only a proxy for conceptual understanding."

Experts move from **generators** to **interpreters & verifiers**.

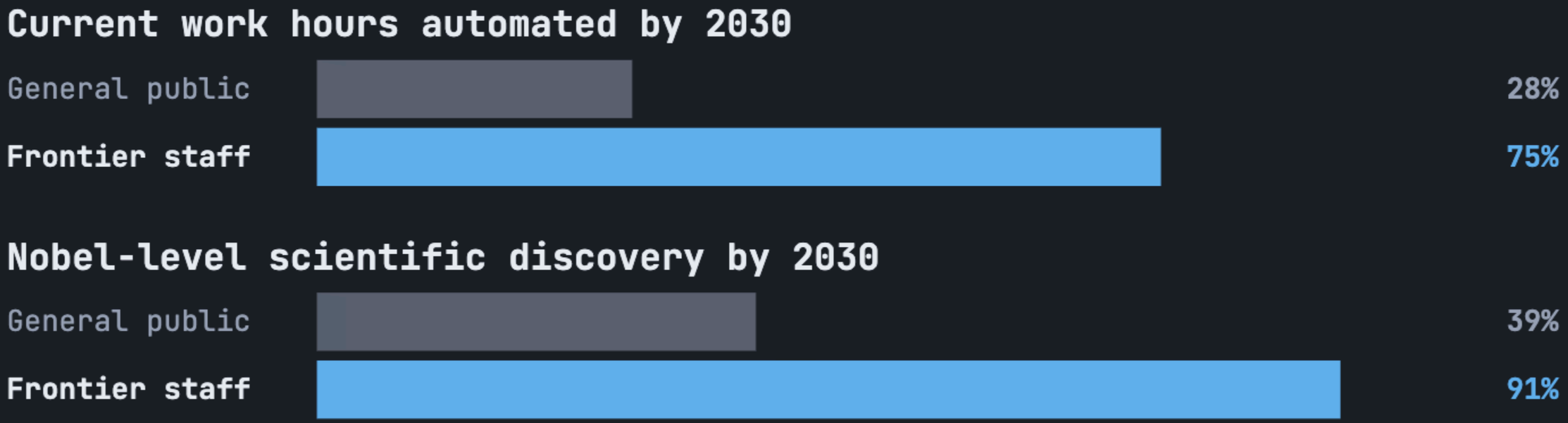
REF: Andrej Karpathy — Widening funnel of AI reality (X, 2026) ↗ Riemann Hypothesis Verification Status (Nov 2026) ↗ HN Discussion (50574968) ↗ Scott Aaronson — on this month's AI proofs (Nov 2026) ↗ Terence Tao — Mathematics in the age of AI (ICM 2026) ↗ Anthropic — Claude Fable & Mythos 5.2 (Nov 2026) ↗ Dwarkesh Patel — Recursive self-improvement has started ↗

THE EXPECTATIONS GAP · ASK THE ROOM FIRST

What the future holds: white-collar automation

Surveyed Anthropic employees expect 75% of current work hours automated by 2030 — surveyed US adults expect 28%.

The AI Expectations Gap (Anthropic, Oct 2026) 1,334 frontier staff vs. 10,980 US adults



You and your AI can build almost everything.

WHAT TO DO NEXT

Talk is cheap. Build something useful.

1 Skip the classes and steering committees — give 2-3 curious insiders frontier tools and real data.

2 Start with real business problems — make someone's Tuesday visibly faster, easier, and more honest.

3 Own and understand the code — trace every number back to the source.

4 Run a 30-day shadow test: compare manual vs. plain SQL vs. AI — keep permission not to use AI if plain code wins.

Full reading list, workshops & tools below on **khalido.dev**

REF: Sean Goedecke — When in doubt, flail about ↗ Simon Willison — NICAR 2026 Data Analysis Workshop ↗