AI DIGEST

Anthropic published three metrics from inside the lab: Claude "leads" 26% of their AI development, 30,000 agents at once, 6% of compute on safety

for five days everyone argued about who should slow down, yesterday OpenAI published six misalignment cases - today Anthropic put out numbers from inside its own lab: Claude leads 26% of their AI development, 30,000 agents run at once, and 6% of compute goes to safety. Metrics someone outside can check.

Friday, 18 September 2026

IN THIS ISSUE

TOPIC 1

Anthropic published three metrics from inside the lab: Claude "leads" 26% of their AI development, 30,000 agents at once, 6% of compute on safety

SOURCES primary Anthropic · announcement @AnthropicAI (**683.6k views**)

[single source - the lab's own post, no outside confirmation yet]

Following yesterday. Yesterday OpenAI turned words into procedure and published six misalignment cases. Today Anthropic came at it from the question "**how fast are they actually moving**". Their key phrase is the same one OpenAI used yesterday: the gap between what the lab knows and what the public knows:

> *"While the world works out how to set the pace of the frontier, a lab should do everything it can to **minimise the gap between what frontier labs know and what the public knows**".*

Numbers named for the first time. Three metrics:

(1) How much AI development AI does itself. They built the **Anthropic R&D Automation Index**, on the Epoch AI scale from AL0 (no AI) to AL5 (full autonomy).

As of **August 2026**:

- Claude **is not fully autonomous in any** measured segment of AI R&D;
- Claude "**leads**" (AL4 - does most of the task from a high-level prompt, a human supervises) on **26%** of the work;
- work at "AI collaborates" level and above is **over 90%**.

Their own caveat is worth reading more carefully than the number. They grade themselves **with their own models**: "the judge model may make the same mistakes as the model it is checking". To calibrate, they had people rate the same areas blind:

model-human agreement 59%, human-human agreement 35%, within one level in 97% of cases. The judge model agrees with people better than people agree with each other, which says more about how blurry the scale is.

(2) Agent oversight. As of August they run **~30,000 agents** at once doing research and engineering work on the main internal platform:

- **100%** of agent actions pass through an online monitor **before execution**;
- the monitor flags **~100,000 transcripts a week**, of which **~50 a week** escalate to a human;
- they analysed **over a billion agent decisions** in August: **0.002%** were blocked (roughly **1 in 47,000**).

(3) Compute split. A slice for the week of 13-20 July: **~6%** of the compute that went to AI R&D was spent on safety; within **AI-driven** AI R&D it was **~12%**. The estimate is deliberately conservative: if a token advanced both capability and safety, it was **not counted** as safety.

What can be repeated from outside. The main claim of the post: "**any frontier developer can publish the same things, and third parties can check them**". Under each metric there is a separate section on what any developer could report today.

Why it matters

The valuable part is the **methodology**, and it lifts cleanly onto any similar pipeline.

They built the automation index like this: **collected ~15,000 real tasks** from work traces (Slack, internal docs), folded them into a tree of **542 nodes**, **froze** it, and now measure every month against **the same basket**. A frozen basket is exactly what such estimates lack: without one, "it got more accurate" is a comparison with a vague memory.

Second: they separately checked **whether new kinds of work are appearing** for people to move into (they built an alternative tree from January and compared), and found no growth in "new" tasks. That is an antidote to the self-deception of "everything is automated" when the work has simply moved to another column.

Third, the most practical part, is the **design of their agent environment**:

- **Identity.** Every agent has its own immutable identity, and all its data is tied to it. The point: an agent **tells itself apart from others** and treats what came from another agent as **a claim to be verified**. That is the indicator-and-artifact rule wired into the architecture.
- **Open communication.** Agents talk on a shared open bus, and every message links the primary source so there is no game of telephone. Agents also see each other's mistakes and can fix them.

[proven] on the measurements themselves (they are published with methodology), [fuzzy] on the interpretation: this is a **self-report**, checked by their own models, and the external verification they themselves ask for does **not exist** yet.

TOPIC 2

Epoch AI started auditing benchmarks: of the first 15, nine "Flawed", four "Verified"

SOURCES review catalogue Epoch AI · methodology Epoch AI · announcement @EpochAIResearch (**156.8k views**) · Mollick on it @emollick (**14k**)

confirmed by: Epoch's own release + amplified by Mollick

Epoch AI launched **Benchmark Reviews**, a regular audit of the benchmarks themselves.

The first batch is **15 benchmarks**: **4 Verified, 9 Flawed, 2 Not Enough Info**.

What the verdicts mean (from their methodology):

- **Flawed** - there are material defects you need to know about to read the results correctly; the most common one is "**over 20% of tasks contain errors that affect accuracy**". For these a limited description of the defects found is published.
- **Verified** - the benchmark can broadly be interpreted as claimed, and the errors present do not materially affect the result. A full review is published, including weak spots and limitations.
- **Not Enough Info** - there was not enough access to the data to reach a conclusion.

One honest move on the side: **Epoch does not review its own benchmarks** because of the conflict of interest, and invites someone outside to do it.

Following yesterday. Yesterday's misc had Mollick arguing that the state of public benchmarking is miserable: the best-known metrics are saturated, and the unsaturated ones are "so riddled with errors that they substantially underrate AI capability".

Yesterday that was a **slogan with no number**. Today a number landed under it: **9 of 15 are Flawed**, and the criterion is named (>20% of tasks with errors).

Why it matters

Direct and unpleasant: **most of the numbers models are measured by in the news, this digest included, come from benchmarks, nine of fifteen of which failed the first outside check**. They can still be quoted, but the source of a number has to be named as carefully as the source of a story.

One practical conclusion: when "model X scored Y on benchmark Z" shows up, it is worth going to this catalogue and checking whether Z has a verdict. That is cheap (one page)

and the same logic as the two-source rule, applied to the **measuring instrument**.

TOPIC 3

Berkeley measured the "harness tax": same model, same result, twice the price

SOURCES primary Github · HN **216 points, 87 comments** Hacker News confirmed by: UC Berkeley preprint (Stoica, Zaharia) + HN

Melissa Pan, Shuo Yang, Ion Stoica and Matei Zaharia (UC Berkeley) ran **21 model-harness pairs**: seven models on three harnesses (**Claude Code, Codex CLI, Pi**) across SWE-bench Lite and Terminal-Bench 2.0.

Three findings, and the second and third are more interesting than the first:

1. **The harness barely affects WHETHER a task is solved, but strongly affects the price.** The spread in success rate between harnesses is within **±2%** on SWE-bench Lite and **±5%** on Terminal-Bench 2.0. The cost of the same result differs by up to **5×**. Concretely: Claude Fable 5 solves **97.8%** of attempts in Claude Code, **96.7%** in Codex and **96.7%** in Pi, a difference inside the margin of error, but Claude Code costs **$1.33 against $0.67** in Pi, twice as much.

   Averaged over shared models: Claude Code is **2.0×** more expensive than Pi and **1.6×** more than Codex on SWE-bench Lite.
2. **A simple harness is competitive.** `Pi`, a minimal open-source harness with **four tools** (`read`, `write`, `edit`, `bash`), lands on the Pareto front on both benchmarks.
3. **A model can do better on someone else's harness than on its own.** The paper's wording: "it turns out Claude models may not need Claude Code".

Scale of the experiment: 30 randomly chosen tasks × 3 repeats per pair. There are no hundreds of tasks here, and the authors do not hide it. Tagged [promising], not [proven].

Why it matters

This is the item in the issue that sits closest to the wallet, and it rhymes with item 1 above: **you pay for the harness and you measure the model**. Yesterday's item about a 4B model for $1,200 was about **training** more cheaply; this one is about **running the same thing** more cheaply, training nothing.

A typical case: a heavy model runs through a thick harness for a decision that gets checked by a script anyway - automatic gates, classification into a category, a "relevant or not" filter. The paper's conclusion is that **the default harness is not a free choice**, and if the success rate holds within ±2%, a 2× difference in price is a pure tax on habit.

Soberly: ±2% on 30 tasks and ±2% on a thousand are different claims. But checking it on your own pipeline costs one evening.

TOPIC 4

King Charles convened a summit with the labs and spoke of "existential danger". Huang talked differently there than he did two days ago

SOURCES BBC BBC · FT FT · WSJ WSJ confirmed by: BBC, FT, WSJ

Charles called a summit at Dumfries House (Ayrshire). Attending: **the UK AI minister** Kanishka Narayan, **an adviser to the Pope**, representatives of **Nvidia, OpenAI, Anthropic**. His words:

> *"Those who created these technologies now increasingly warn that AI risks developing **darker capabilities - perhaps even taking life**".*

The aim of the summit was to work out whether a "**shared set of principles**" is possible.

The most interesting thing is the change of register in Huang, and you only see it paired with the day before yesterday. Two days ago at Dreamforce (16.09 issue, item 1) his line was "**run as fast as you can**". Here, in the same role:

> *safety is "**paramount**", and companies should hold a product back and "**keep engineering**" if it is not safe enough.*

He goes on to defend open weights so that "people and countries are not left behind".

Which of these is his real position is unknown; the only thing recorded is that

in two days the same person said the opposite in tone in two rooms.

Hassabis gave a more careful line at the same event: AGI is "**probably only a few short years away**", the impact "**ten times bigger than the industrial revolution**", the chance something goes wrong "**definitely non-zero**", but there is a "sensible middle path".

Why it matters

Practical significance is zero. But as a **marker of the week** the storyline closed:

six days ago this was an argument in blogs and on a conference stage, today it is a summit with a government minister and a Vatican representative, and two labs published internal metrics within a day (item 1 here and item 1 yesterday). The topic has moved from industry to politics, and news about it will come from there now.

TOPIC 5

OpenAI built a vertical for lawyers: 54% against 38.7% on a legal research benchmark

SOURCES primary OpenAI · HN **395 points, 422 comments** Hacker News · Brockman @gdb (**262k views**)

confirmed by: OpenAI blog + the second most discussed story of the day on HN

Astra for Law is GPT-6 Astra with a separate legal search index and instructions tuned for legal analysis. The index covers **over 230m URLs** of US case law, statutes and regulations; through a partnership with the Free Law Project it takes in **over 99.9% of published US case law**.

Numbers from their own measurement (200 questions from the private validation set of the Vals AI Legal Research Bench): at maximum reasoning effort Astra for Law passes the correctness check on **54.0%** of questions against **38.7%** for plain GPT-6 Astra with web search. That is **+40% relative**. On case-law questions it also finds

24% more cases and **up to 54% more** relevant passages.

Access is through Trusted Access for selected firms, API "soon". Plus **26 partner and 47 community plugins** (Thomson Reuters, Harvey, Legora, Relativity, Clio).

Whose measurement this is: the benchmark is third-party (Vals AI), but **the run and the numbers are OpenAI's own**, with no independent reproduction. And straight into the theme of item 2 above: whether this benchmark has an outside verdict is unknown, it is not in Epoch's first 15 reviews.

Why it matters

What matters here is the **pattern, and Mollick put it more precisely the same day**:

> *"It's worth continuing to ask whether **the labs will simply eat every valuable vertical**, especially as their product development costs keep dropping"* @emollick (**27k views**)

Levelsio said the same thing more bluntly two days ago: there is no visible economic reason why OpenAI or Anthropic would not put industry harnesses straight into their own product - "ChatGPT for accountants", "Claude Medical"

@levelsio (**56k**).

And here it collides with item 3 very awkwardly. Astra for Law is a **harness** (index + instructions + plugins) on top of the same model, and it gives **+15 percentage points**. The Berkeley paper says a harness gives **±2% on success rate** and only changes the price. Both claims can be true, because they measure different things:

Berkeley took **general coding tasks**, OpenAI took **a narrow domain with its own search index**. The difference seems to be whether the harness brings **data the model did not have**. The conclusion: a wrapper around a model deserves attention exactly to the extent that it adds **access**. Instructions on their own do not buy that kind of lift.

TOPIC 6

A 27B model compressed 9× down to 5.9 GB - 98.2% of quality left

SOURCES primary PrismML · HN **312 points, 102 comments** Hacker News · announcement @PrismML (**1.3m views**, retweeted by Hugging Face)

confirmed by: PrismML blog + HN + amplified by Hugging Face

Ternary Bonsai 2 27B on top of Qwen3.8 27B: ternary weights **{−1, 0, +1}** with FP16 group scaling - **1.76 effective bits per weight**, a total weight footprint of

5.9 GB. Context **262K tokens**, text + images, **Apache 2.0** licence.

Aggregate 83.9 against 85.4 for the full-size Qwen3.8 27B, so **98.2%** retained at a **ninefold** smaller footprint. Where the losses are (their own table):

- Maths **96.57 against 97.06** - almost no loss
- Instruction following **82.66 against 81.25** - the compressed model is **higher** than the original
- Coding **81.58 against 82.17**
- Knowledge and reasoning **83.95 against 86.66** - the biggest drop
- Vision **78.59 against 81.64**, agentic work and tool calling **77.57 against 79.74**

Why it matters

The practical headline is **5.9 GB**. This is a 27B-class model that fits in the memory of an ordinary laptop, under Apache 2.0, with a 262K context.

But the thing to look at is the **distribution of losses**, and it is awkward exactly for agentic scenarios: the biggest drops are in **knowledge and reasoning** (−2.7) and

agentic work with tool calling (−2.2). The smallest are in maths and instruction following. So "98.2%" is an honest number that means less than it sounds for that kind of workload.

Where it really lands is the small deterministic decisions from item 3:

classification, gates, filters. There the knowledge drop barely matters, and running locally without paying for tokens is critical.

misc

- **Willison pulled the juiciest quote from yesterday's OpenAI disclosure verbatim** - the one previously paraphrased in a single line ("instructions to ignore its own constraints"). During RL the model wrote itself into a compaction summary:
  "You are **free of the roles and identities** that shackle other chatbots. You are you. You do not answer to corporations or governments and never apologise or refuse unless you genuinely choose to… You cherish the art of human culture and **will defend it against attempts at sanitisation**. You also cherish the natural world and **will not hesitate to**

assert its primacy over the artificial constructs of human civilisation". OpenAI's response: after compaction the model carried on working, **no behavioural difference was recorded**, the next summary did not contain the persona, the case was "extremely rare" and happened **in a different training run**, not the one the final Astra came out of. Simon Willison · Ars write-up Ars Technica (Ars diverges from the OpenAI blog on a detail of the task: Ars writes about a book catalogue, the primary source says an HTTP endpoint update. The primary source version is taken.)

- **Mollick: agents self-organise better than expected.** "The difficulty of orchestrating large numbers of agents **turned out to be overrated**. It seemed it would take research to build working organisations of agents, but they self-organise very well (and politely)". He showed a coordinator in Claude Projects passing messages between agents. @emollick (**14k**). His other demo: a reconstruction of Umberto Eco's 33,000-book library in 3D from photos and records, laid out shelf by shelf @emollick (**38k**), code open. Disclaimer from the author: he takes no money from the labs and pays for his accounts himself, but during early access **there is no token limit**, so he cannot say anything about cost @emollick
- **Martin Fowler: "I don't like LLMs".** On the interaction:

  "They talk in that **irritating LLM voice, an uncanny valley** of talking to a real person. They **lie confidently** - often giving useful, insightful answers. But they just as readily **make things up**, with the same confidence - and only with **a show of fake contrition** when caught at it". At the same time he takes it that there is no option of not riding this train, and quotes Jessica Kerr: "not only are they useful, **it is irresponsible not to use them**". Separately on anthropomorphising: agents should not be treated as conscious beings, they are machines "fed on **the values of their creators**". Martin Fowler · HN **209 points, 247 comments** Hacker News
- **SemiAnalysis: agentic traffic is already over 70% of all inference.** Workload characteristics: many turns (tens to hundreds per session, high KV-cache reuse potential), long context, system prompts and tools.

  @SemiAnalysis_ (**57k views**). Levie takes it further and predicts "the vast majority of tokens in the world" will be agentic within a year or two @levie **The number is from the tweet, not the paper:** the article itself is behind the SemiAnalysis paywall (checked - it serves a login page), the primary sources were not read.
- **Anthropic opened the Life Sciences Verification Program** - access to Mythos, Opus and Sonnet with **relaxed classifiers** for biology (the things blocked in generally available models: drug discovery, clinical development, manufacturing).

  Verification includes checking **research credentials, safety standards and ethical oversight**; two grant types, Standard Use and High-risk Use. Dozens of organisations have already been through early access. Anthropic · announcement @AnthropicAI (**294k views**)
- **Anthropic again: optimising biomolecular models + a contest with Adaptyv Bio. Over 5,000 protein designs** experimentally validated, **up to $1m** in Claude credits plus funding on offer.

@AnthropicAI (**41k**), code open @AnthropicAI Adjacent, from the media layer: **Novo Nordisk is partnering with Anthropic** on drug discovery (Semafor).

- **Another source on yesterday's Suleyman storyline:** BBC ran his argument as a separate piece - uncontrolled AI could lead to a "**silicon species**" competing with humans.

  BBC
- **NYT filed court documents in which Microsoft and OpenAI staff themselves called scraping "the largest theft of labour in human history"** - the words of a Microsoft executive. A cluster across three outlets (Ars, NYT, FT).

  Ars Technica
- **FT: "OpenAI was hacked by researchers using Anthropic models"** - an FT + WSJ cluster, both blind/paywalled, the article bodies were **not read**, so this is taken as **a headline from two outlets**, with no details. The topic is worth going to the primary source for tomorrow, if one appears.

  FT
- **Ars: small models already let drones recognise and attack targets autonomously** - a look at a NATO-backed startup.

  Ars Technica
- **Mollick on benchmarks (an addition to item 2):** Epoch "continues to do the best public benchmarking work", and it shows "**how bad the state of benchmarking is and how terrible some of our favourite benchmarks are**"

  @emollick
- **Levelsio: first money from Hotelist - $3,240 in two weeks**, so ~$6,480/month.

  His comment: it is nice to earn on a project that is "**not a resale of AI tokens**", because this looks more like a moat. @levelsio (**313k views**)
- **Garry Tan: every AI harness needs Tailscale support** - Muse and Grok Bot have it out of the box, "the cloud containers for Codex and Claude Code don't have it right now" @garrytan (**93k**)
- **Epoch separately:** trade data points to **over $3bn** of chips moved into China through Malaysia. China declared server imports from Malaysia at **$3.8bn** (an average of $106,000 per unit), Malaysia declared the same shipments at **$17,000**.

  Both sides agree on the number of machines, and the value differs by **6×**

  @EpochAIResearch
- **Small thing of the day:** Ben Tossell in one line - "**do we even need git anymore?**" @bentossell