AI DIGEST

METR: AI Doubles the Pace of Anthropic Research

Anthropic builds its models about 2 times faster than a year ago, and AI drives most of it.

Saturday, 7 November 2026

IN THIS ISSUE

1. METR: Anthropic develops its models about two times faster, and AI gives most of the acceleration
2. Anthropic: AI can take 27-52% of a scientist's time with current models
3. DeepMind: AI agents found two weaknesses in RSA-1024 in weeks
4. Anthropic IPO: buyers now talk about $1.1-1.3 trillion
5. Anthropic on agent memory: "dreams" repair notes
6. Tesla stops Cybertruck production
7. Chainalysis: attackers got the login data of all customers
8. Hightouch: a cheap model plus Opus for hard cases gives the accuracy of Opus
9. Frontier Model Forum: a shared scale to rate cybersecurity jailbreaks
10. Congress calls Jensen Huang to a closed briefing

TOPIC 1

METR: Anthropic develops its models about two times faster, and AI gives most of the acceleration

SOURCES METR, 05.11 Metr · Ajeya Cotra, 06.11 Planned-obsolescence · HN, 06.11 Hacker News · Arjun Panickssery, 06.11 Lesswrong · HN, 06.11 Hacker News

From 12.08 to 29.10, METR staff worked inside Anthropic. They had access to internal metrics of pretraining and RL, to benchmarks of internal models and to interviews with people. The question was one: how much does AI already accelerate the development of AI itself?

Here is the main METR number. Since October 2025, the internal capability index of Anthropic (AECI, a version of the Epoch Capabilities Index) grows 1.9 times faster than before. The 95% confidence interval is from 1.5 to 2.8 times. In other words, in one year the lab made about two years of its former progress. METR is sure that faster growth of training compute is not the cause. METR explains most of the acceleration today with the wide use of AI in the research of Anthropic itself. METR saw no bottlenecks that could stop the progress in the next 12 months.

The report has an important warning. Anthropic had the right to edit non-public data, and it removed most of the evidence that METR found the most convincing. The company said that this data could accelerate other developers. Thus the public must trust the judgment of METR more than METR itself wants.

Ajeya Cotra from METR takes the number further, in her own name. If total progress goes two times faster, the work of researchers must accelerate much more. Her estimate is about 10 times. Anthropic staff in surveys report an acceleration of about 12 times. Cotra writes that Anthropic already admitted in its IPO documents that the company "may imminently meet" the "dramatically accelerated progress" threshold from its safety policy.

There is also a skeptical position. Arjun Panickssery on LessWrong describes some of his projects with agents: a Scrabble engine, an Ultimate Tic-Tac-Toe bot and a formalization of proofs from philosophy. His conclusion: an agent completes tasks with fast feedback well, but it chooses the direction badly. The agent misses the main paper in a literature review and runs experiments that cannot show anything.

Why it matters: this is the first independent measurement of a claim that only the labs made before. The two positions do not contradict each other. The acceleration is real where you can check the result fast, and most work on a model is of this type. For other industries, the conclusion is practical. Speed now depends on how well you can evaluate a task automatically.

TOPIC 2

Anthropic: AI can take 27-52% of a scientist's time with current models

SOURCES Anthropic, 05.11 Anthropic · @emollick, 07.11 @emollick · HN, 06.11 Hacker News

Anthropic economists analyzed 207,000 research sessions in Claude and Claude Code. Data analysis takes about one third of the sessions in Claude and almost half in Claude Code. Scientists use Claude less for ideas and data collection. But in a survey, they called exactly these tasks the most important. Computer science, mathematics and statistics use Claude the most. Medicine and laboratory sciences use it the least.

Then the authors build a model. If people adopt current capabilities widely, AI can take 27-52% of a researcher's time on one paper, with a median of 38%. This gives an acceleration from 1.35 times in the arts to 1.96 times in computer science and 2.02 times in statistics. In a hypothetical scenario where AI does all non-physical work, the output in computer science and statistics grows up to 15 times. In chemistry and biology, the laboratory limits the growth to 3-4 times. About 40% of the surveyed scientists who already work with AI expect that by 2030 AI will run full research projects alone.

Why it matters: the numbers in this report and item 1 describe one effect from two sides. METR sees acceleration in one lab, where almost all work is cognitive. Anthropic shows that

in science in general, physical experiments set the limit. A forecast "AI will accelerate science N times" without a split by field says little.

TOPIC 3

DeepMind: AI agents found two weaknesses in RSA-1024 in weeks

SOURCES @pushmeet, 06.11 @pushmeet · @GoogleDeepMind, 06.11 @GoogleDeepMind · signatures paper Googleapis · factorization paper Googleapis · The Wall Street Journal, 06.11 WSJ

The DeepMind Science team released two papers together with the cryptographers Nadia Heninger (UC San Diego and Inria)

and Emmanuel Thomé (Inria). The agents worked under human supervision. Pushmeet Kohli writes that such results took years before, and here they took weeks.

The first result is about digital signatures. Before a signature, standards add fixed data to the message, the padding.

The agents showed that the padding gives no protection when an attacker controls about half of the block or more. On a key from an already solved academic benchmark, the attack used 3.5 times less CPU. For the case with a short 32-byte digest, the new algorithm is about 1,600 times better than the previous one, by a theoretical estimate.

The second result is about the factorization of many keys of the same length. The attacker pays once for a large shared computation. After that, each next RSA-1024 key costs about 30 times less than with previous batch methods. Readers on X immediately clarified: the 30 times apply to the price of an additional key in the batch. The price to break one separate key does not fall 30 times.

These results do not affect RSA-2048 and longer keys. The WSJ puts a wider idea in its headline: AI can break old encryption before quantum computers can.

Why it matters: cryptography stands on the assumption that a specific problem is hard. Until now, the safety margin changed over years. These papers show that the margin can now change in weeks. Systems that still use RSA-1024 keys or old signature schemes must move to 2048 bits or more, even without a quantum computer.

TOPIC 4

Anthropic IPO: buyers now talk about $1.1-1.3 trillion

SOURCES Bloomberg, 06.11 Bloomberg · HN, 06.11 Hacker News · Financial Times, 06.11 FT · BBC, 06.11 BBC

Yesterday this digest wrote that large investors doubt the $1.5 trillion valuation. Today Bloomberg gives new numbers from sources. Potential buyers discuss a valuation of $1.2-1.3 trillion, and some talk about $1.1 trillion. Before the crash, expectations were $1.4-1.5 trillion, and at the start of October they reached $2 trillion. Formal marketing of the offering starts next week. Anthropic did not comment.

Bloomberg calculates the scale of Thursday. The Nasdaq 100 fell 6.4% and lost $2.3 trillion of value, a record for one day. HSBC strategist Max Kettner thinks that a valuation of about $1 trillion would decrease the risk of sharp moves after the start of trading. The FT headline says a different thing. Anthropic keeps its $1.5 trillion target and prepares a separate offer for retail investors in Britain.

On Friday, the market partly recovered. According to the BBC, the Nasdaq Composite closed 1.1% higher, and the S&P 500 closed 0.3% higher. Asia continued to fall: the Kospi in Korea lost 6%, and SK Hynix lost more than 10%.

Why it matters: from $2 trillion to $1.1 trillion in one month, almost the full value of Anthropic in May ($965 billion) disappeared.

The IPO of the first frontier model lab will set the price for the full industry. Thus this price is also important for OpenAI, which raises money at a valuation of $1.25 trillion. Item 1 adds context: investors value a company that, according to METR, accelerates its own development.

TOPIC 5

Anthropic on agent memory: "dreams" repair notes

SOURCES Anthropic, 06.11 Anthropic · @AnthropicAI, 06.11 @AnthropicAI

Anthropic engineers described how Claude keeps memory between sessions. Here, memory means note files that the agent writes itself and reads in the next session. "Dreaming" means a separate pass without the user. The agent reads the logs again, repairs wrong notes and builds new structures.

Pokémon Red gives the clearest example. Sonnet 3.7 wrote everything into memory: 413 files in 225 folders, and one badge in 10,000 steps. Fable 5 had nine files in two folders and one shared `global.md`, and it earned four badges.

Fable 5 did not write down what the screen shows. It kept only strategies and verified rules.

Strict rules hurt strong models. When Opus 5 and Opus 4.7 got the rule "write down only errors", they lost most of the gain from memory. For older models, hints helped. Sonnet 4.6 with the instruction "read memory before you answer"

almost doubled its result.

A separate test is about a wiki. Claude built a wiki from 350,000 files (about 420 million tokens) of the company Stainless. Over 128 "dream" sessions, accuracy on 25 questions grew from 13% to 48%. Search together with the wiki gave 61% against 53% with search only. It also saved about 20,000 tokens per question.

Why it matters: the main lesson is simple. Explain to the agent the purpose of memory. Do not dictate what it must write. Second, memory gets old. Without a check against the logs, a

false note repeats the same error each time. For any system with long memory, a regular check of notes against the primary source costs less than the consequences.

TOPIC 6

Tesla stops Cybertruck production

SOURCES Reuters via NBC News, 06.11 Nbcnews · The Guardian, 06.11 Guardian · Ars Technica, 06.11 Ars Technica · HN, 06.11 Hacker News

Elon Musk wrote on X that Tesla will stop Cybertruck production at its Texas factory by about the middle of 2027. The free capacity will go to Cybercab, the robotaxi of the company. Musk called Cybertruck "the best Tesla car" and "the end of an era".

The sales numbers explain the decision. According to Kelley Blue Book, Tesla sold 20,237 Cybertrucks in the US last year. This is 48% less than 38,965 in 2024. Before the launch, Musk promised 250,000 per year. The HN thread got 903 points and 1,753 comments.

Why it matters: Tesla removes a product for its bet on autonomy and robots. In the same week, the market sold everything related to AI. Thus Tesla doubles its bet at the exact time when investors doubt it the most.

TOPIC 7

Chainalysis: attackers got the login data of all customers

SOURCES Ars Technica, 06.11 Ars Technica `[single source]`

The blockchain analysis company Chainalysis reported that attackers got the names, logins, password hashes and live tokens of all its users. The customers of the company include the FBI, ICE, the US Air Force, Microsoft, Coinbase and Barclays.

Chainalysis describes the attack as "multifront". The attacker used techniques with help from AI agents and an on-path attack, which is interception of traffic on intermediate infrastructure. The company does not say where the interception happened. The investigation found the leak on Monday, and the company reported it on Wednesday. Users must reset their credentials. Chainalysis gives back access to its services step by step, through an allowlist of IP addresses.

Why it matters: live tokens are more dangerous than password hashes, because they give access without a break-in.

The customers of such a service are targets themselves. Thus a leak at the vendor gives the attacker the list and the keys at the same time. This is one more case where a company directly mentions AI agents in an attack.

TOPIC 8

Hightouch: a cheap model plus Opus for hard cases gives the accuracy of Opus

SOURCES Hightouch, 04.11 Hightouch · HN, 06.11 Hacker News

Hightouch checks if its system merged customer records into one person correctly. A person checks about 50 such profiles per hour, and an accurate evaluation needs more than 1,000. The team compared Claude Opus 5.5 with two "decision models", Jev and GPT-6 Luna. Such models return only a probability from 0 to 1.

Alone, the decision models cost $0.05-0.12 per 1,000 profiles and answer in 0.15-0.20 s. Opus costs $3.86-5.39 and answers in 2.2-2.6 s. On the hard test, Jev has 89.1% and Opus has 95.1%. Then the team made a hybrid. The decision model sends to Opus only the cases where it is not sure.

Calibration decided the result. Jev makes most errors with low confidence. Thus the hybrid with Jev gave 95.3% on the hard test for $2.94. Luna makes errors with high confidence, so its hybrid gave only 91.6%. On the easy test, the Jev hybrid gave 98.8% against 99.0% for Opus, at $0.88 instead of $3.86.

Why it matters: the accuracy of a cheap model is less important than its knowledge of when it is wrong. The scheme "cheap judge, and the hard cases go to the large model" works for any mass check. Before you launch such a scheme, check that the errors of the cheap model really have low confidence.

TOPIC 9

Frontier Model Forum: a shared scale to rate cybersecurity jailbreaks

SOURCES Frontier Model Forum, 05.11 Frontiermodelforum · @OpenAI, 06.11 @OpenAI

The Frontier Model Forum released version 1.0 of a framework to rate jailbreaks in cybersecurity. The framework rates each bypass of protection in four categories: how much it increases the capabilities of an attacker, how wide these capabilities are, how easy it is to use and how easy it is to find. The goal is practical. Labs must rate the same bypass in the same way and close the most dangerous ones first. Amazon, Anthropic, Google DeepMind, Microsoft and OpenAI support the framework.

Why it matters: yesterday OpenAI showed a model that writes exploits for verified customers. The border between such a model and a usual one depends on protection against bypasses. A shared scale lets people compare whose protection is weaker.

TOPIC 10

Congress calls Jensen Huang to a closed briefing

SOURCES Ars Technica (FT text), 06.11 Ars Technica · The New York Times, 05.11 NYT

The chairs of three House committees (China, foreign affairs and intelligence) invited the Nvidia CEO to a closed briefing on 18.11. The reason: last week the NYT wrote that US intelligence sees Huang as a target of a Chinese spy operation. The story is about Alice Huang, the head of the Singapore cloud company Megaspeed. Her company bought Nvidia chips for billions of dollars. Nvidia rejects the accusations and says that customers get only usual business information.

Why it matters: after the elections, Washington became stricter to big tech companies. Nvidia now has pressure from two sides: export control and the personal security of its CEO. Item 4 shows how fast politics now gets into the price of AI companies.

---

misc

- **Claude Code and email, continued:** the thread about the email in User-Agent became the most popular HN story of the day, with 1,559 points. The issue has 699 thumbs-up reactions, and it still has no comment from Anthropic. GitHub · HN Hacker News
- **AI voice fraud:** in the US, prosecutors charged three men. They used AI voice and video to act as executives and lawyers and stole more than $129 million. Guardian
- **libghostty 1.0:** Mitchell Hashimoto released the stable C terminal library from Ghostty. The terminal in JetBrains IDEs already runs on it, with more than 12.5 million users. Mitchellh · HN Hacker News
- **FAA without floppy disks:** by the end of 2028, the FAA will replace floppy disk systems in 435 air traffic control facilities. Ars Technica · HN Hacker News
- **ChatGPT connectors:** a developer described how ChatGPT silently sent the next call through a different connector with a one-time code from the first one. Openai · HN Hacker News
- **BioBasicsBench:** @OpenAI and Ligo released a test of 100 tasks from real laboratory experiments for agents. @OpenAI
- **DeepFinance:** @0interestrates released a test of 21 tasks at PE and investment bank level on real deal data. @0interestrates
- **AI spam in replies:** @levelsio blocked more than 10,000 accounts with AI replies and promises to publish the list. @levelsio

---