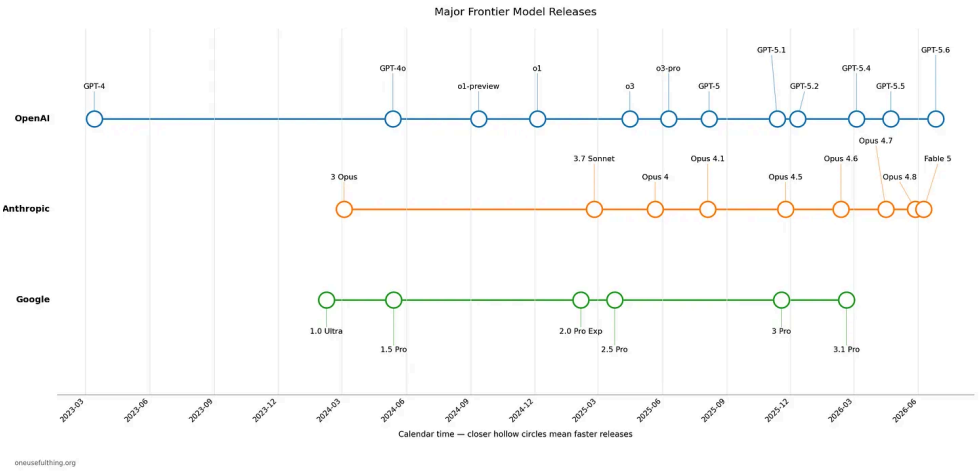
The twilight of the chatbots

How work changes along the exponential

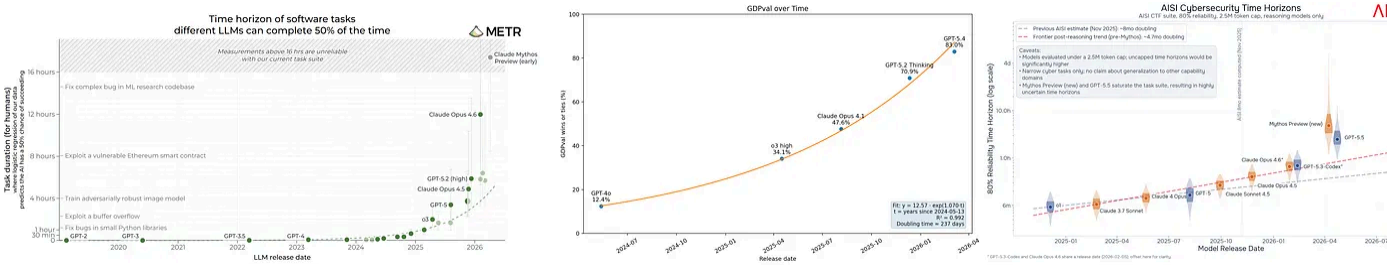
ETHAN MOLLICK

JUN 30, 2026

If you feel like things are accelerating in AI, you are probably right. Better AI mode from the leading American AI labs have been releasing more quickly than ever (tho government interventions stopped access temporarily to two of the most powerful models, Claude Fable and GPT-5.6).



But it isn't just release timing. The evidence points to accelerating capability gains well (though the frontier stays jagged, and AIs remain weak in many places). This is especially obvious when we look at the ability of AIs to do real work. There are a fe good assessments that try to measure how much human work AIs can do. Two of th most famous, from METR and the UK's official government AI Security Institute, estimate the amount of human programmer hours' worth of effort the AI can do wi single prompt. GDPval compares human experts in many fields to AI performance using professional judges. They are all increasing at a better than exponential rate.



Another organization doing similar experiments, Epoch, recently found Opus 4.7, working on its own for 14 hours, was able to build a software package that would ta 2-17 weeks of human engineering work (it cost $251 in tokens). Again, AI systems cannot pass every test, nor are they always cheap to run, but they are definitely improving at a very rapid rate. In my own experiments, I found Fable was able to w autonomously for 9 hours to execute on very complex software projects that would have taken a team well over a week to do.

AI can already solve many long-horizon MirrorCode tasks. The highest-scoring model in our experiments, Claude Opus 4.7, achieved a score of 56% across the benchmark. As an example of what AI can accomplish, Claude Opus 4.7 substantially reimplemented `gotree`—a bioinformatics toolkit with about 16,000 lines of Go and 40+ commands. We believe this same task would take a human engineer without AI assistance 2–17 weeks. Opus 4.7 solved it in 14 hours, passing 2,000/2,001 tests (99.95%), at a cost of $251. AI is also improving rapidly at long-horizon coding. Leading AI models from eight months ago would have scored about 30% on MirrorCode, and were limited to simpler tasks.

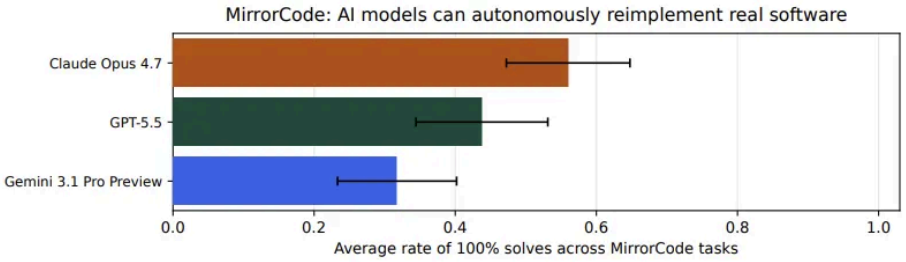
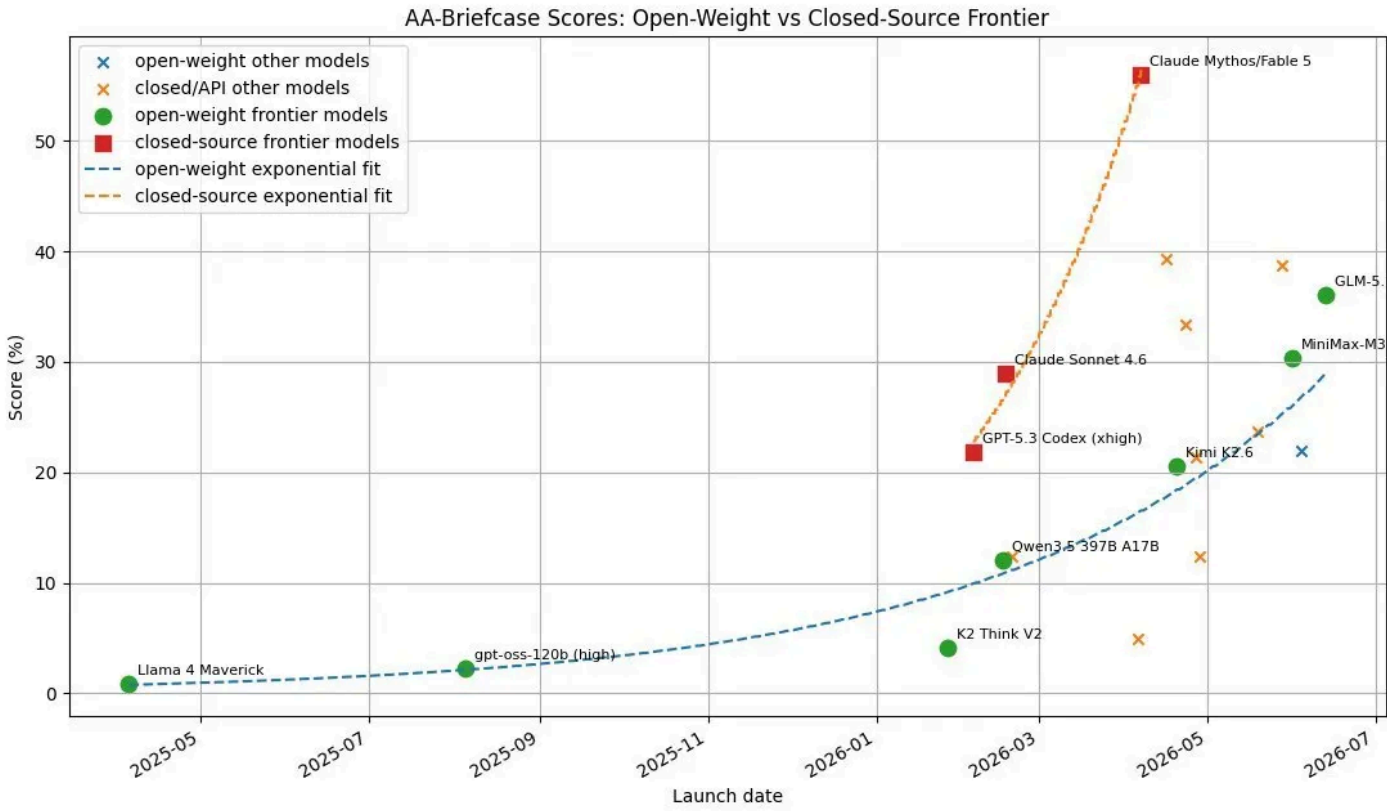


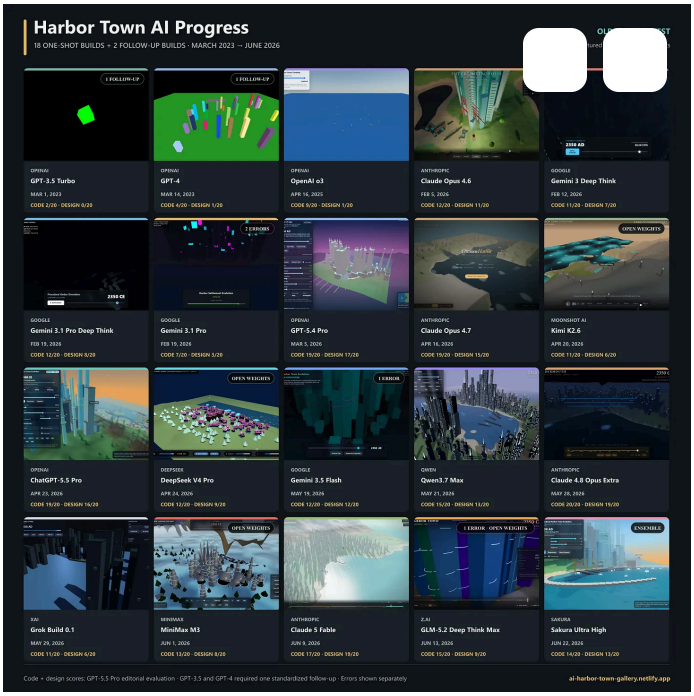
Figure 1: Average rate of 100% solves across all MirrorCode target programs, i.e., how often AI scores 100% reimplementing MirrorCode programs. Bars show ± standard error. Even when targets are not reimplemented perfectly, AI typically make substantial progress, passing 90% or more of tests.

So far, I have focused on the frontier models, those with the highest "intelligence." They are made by three American companies — Anthropic, OpenAI, and Google (though it has been a while since Google has released a new model). But there is a second set of near-frontier AI models that typically lag 6-12 months behind the frontier, all of which are from China. These are open weights models, which means that anyone can use or modify them after release (as opposed to the frontier models which are proprietary). That makes them quite cheap to operate. They, too, are

climbing up an exponential improvement curve, though lagging the American close models. You can see this in my graph of AI performance in a test called AA-Briefca which simulates a complex multi-week consulting engagement where AI has to do many kinds of analysis. The open-weights Chinese models (other countries produce open weights models, but none are near the frontier) are on their own exponential curve, behind closed US models



But abstract graphs only get you so far, and they can hide how jagged the frontier is (and also the fact that the open weights models, while very impressive, do not alway perform as well as their benchmarks would indicate). To get real insight, you need t try using AI for different use cases and rigorously assess how good they are in the areas that matter to you. As a fun example, I created a test where AIs have to build interactive simulation of a harbor evolving over time. You can play with all the resu here. I think it gives an interesting perspective on how much models can differ fror each other in areas like design, stylistic approach, and even judgement. As systems ever longer tasks, these hard-to-benchmark factors become more important.

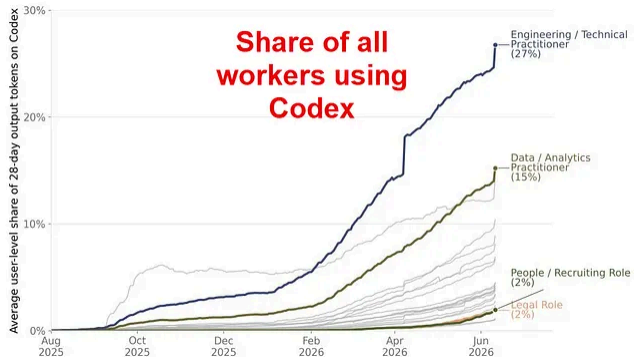


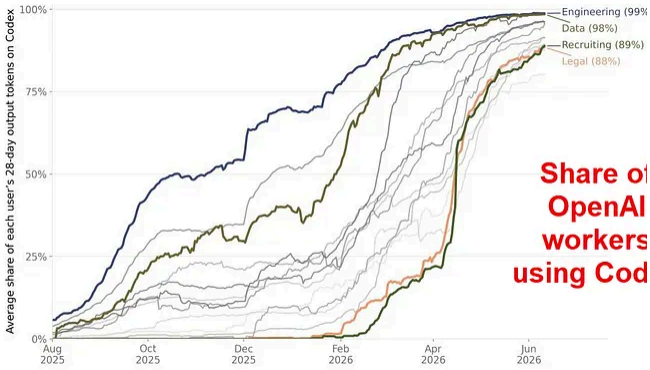
The way we use AI is changing

As AIs can do longer and longer tasks, the way people are using AI is changing. Un recently, the dominant way to use AI was as a co-intelligence. You would ask the A do something, check the results, and then ask for it to do the next step of your job. I careful prompting and human attention, you could guide AIs to do complex and lor term tasks.

This approach to using AI is still common and useful, but, increasingly, it is not the way AI is being used for valuable work. Long-running, smart, and self-correcting A systems do not need constant human intervention, and they require a different way working (this is also the subject of my upcoming book, Co-Existence, which you mi want to pre-order here). And, as opposed to chatbots, agents come with extra machinery: harnesses that give the AI access to tools and an environment to act in, apps built for agents like Claude Code or OpenAI's Codex. As a result, the already increasing ability of AI models can be improved still further by a good harness or a

So work is increasingly about assigning work to agents, rather than working togeth with chatbots. A joint study by OpenAI and academic economists shows how quick this is happening inside their own organization. Critically, it isn't just coders who a using agents. Legal, HR, and other non-tech functions have adopted agents at nearl the same rate. OpenAI may be a sort of canary in the coal mine for what will happe elsewhere in work.





Increasingly, work at OpenAI looks like managing AI. A quarter of OpenAI worker have at least four agents running at one time every week. And, as coding is done by AIs in specialized harnesses and apps, other roles start to become coders of a sort. A they are good at it. A separate study of Claude Code users found that software engineers had a similar success rate to other professions when actually using Claud code on coding tasks.

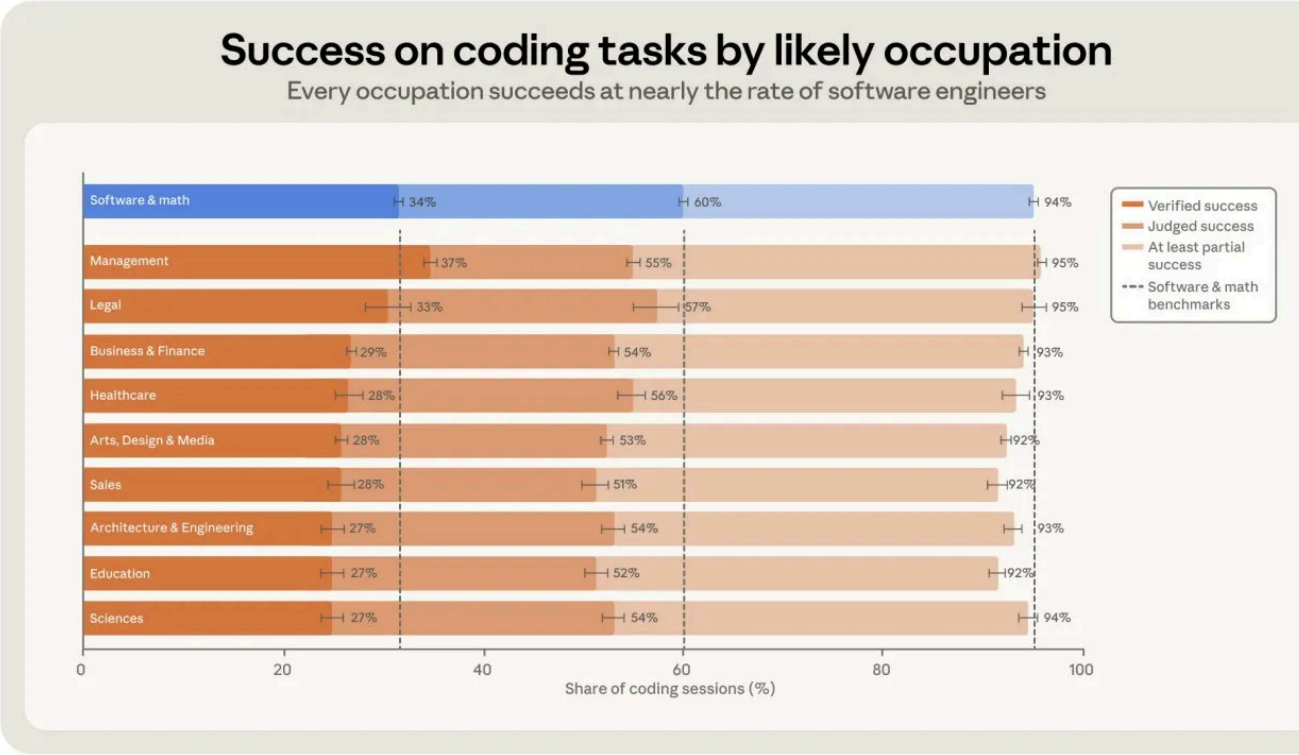


Figure 6: Success rates in coding sessions by inferred occupation
Share of sessions meeting strict definitions of success—judged success and verified success—among sessions that add or change at least one line of code, by the user's inferred occupational group, for the ten largest groups. Every group is within seven percentag points of software/math users (SOC Code Computer and Mathematical Occupations). Error bars are 95% confidence intervals computed on distinct accounts.

What actually mattered was not the profession of the user, but their expertise. The more domain experience someone had, the more successful they were in using Clau Code in that domain. And, even more interestingly, the more useful output they got from Claude from each prompt.

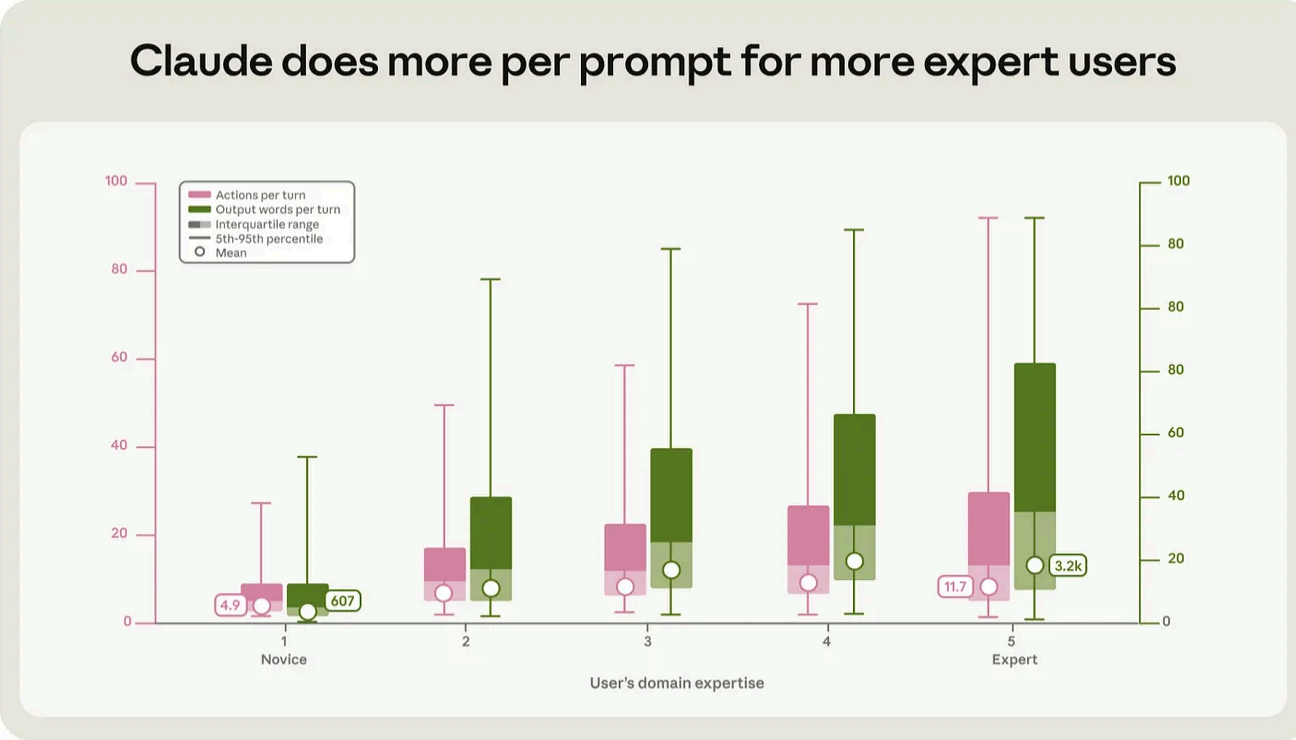


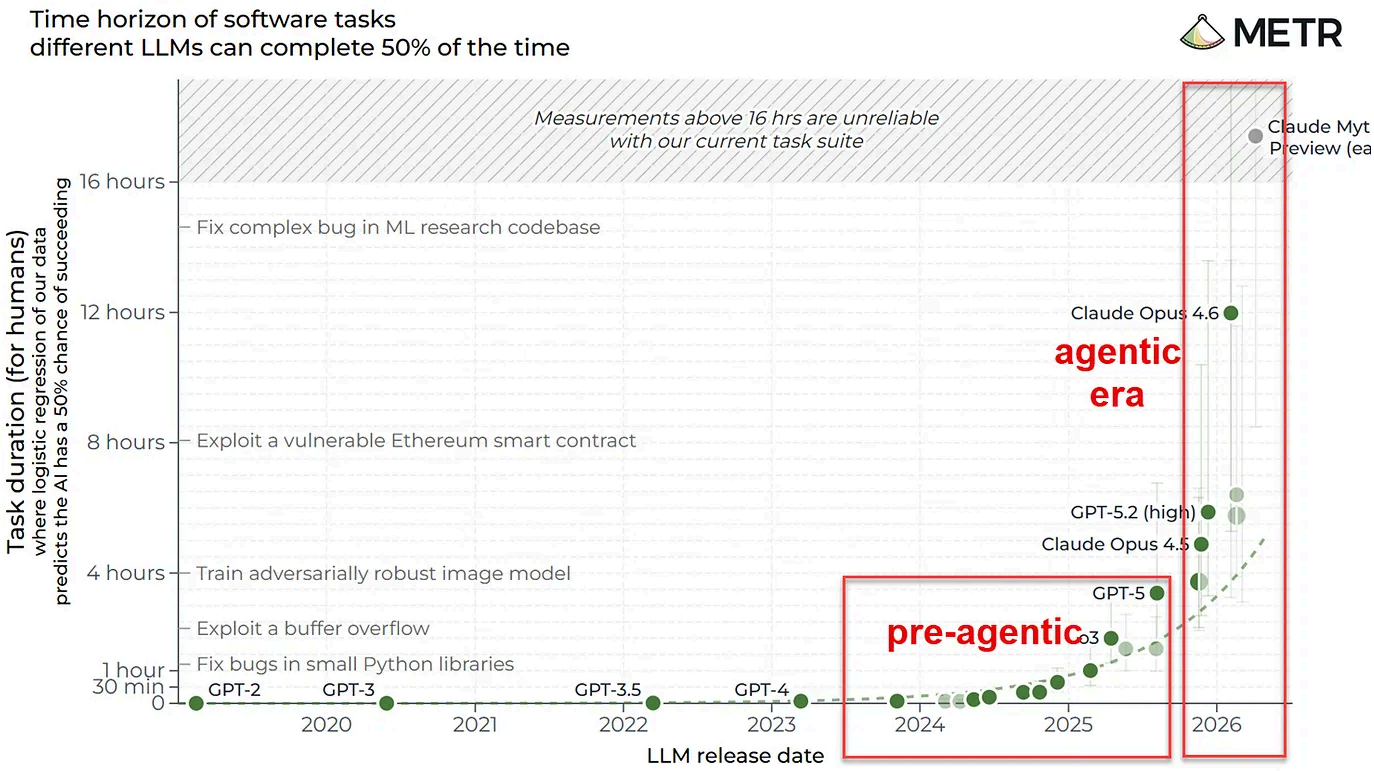
Figure 3: Claude does more per prompt for more expert users
Claude produces more actions (left bar) and text output per prompt (right bar) for more expert users. Boxes span the interquartile range (split at the median). Whiskers represent the 5th to 95th percentile. White dots are geometric means. Both upward trends are statistically significant ($p < 0.001$), as is each adjacent-level step, and they remain significant (at +9% actions and +13% output per expertise level) in a regression controlling for work mode, task value, month, occupation, and model family, with standard errors clustered by user.

We are moving from a world where non-experts use chatbots to fill in gaps to one i which experts use agents to get work done. And the best way to use agents is to thi of yourself as a manager.

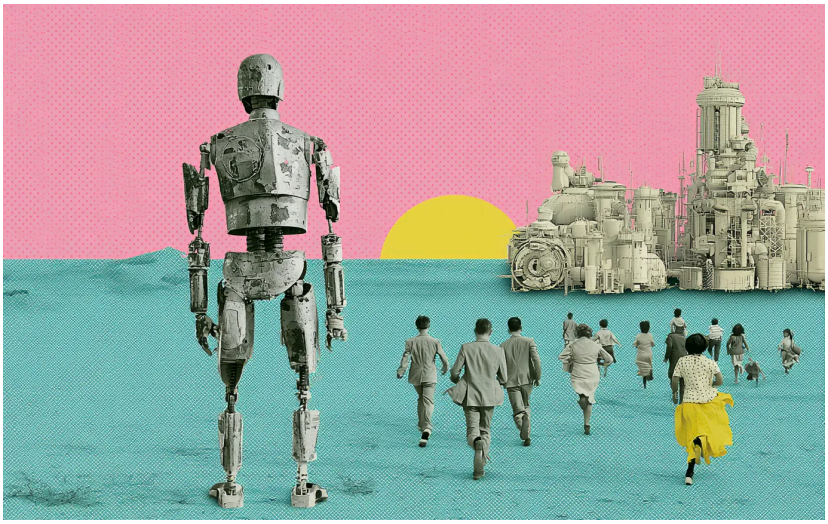
A moment in time

Being on an exponential means each change over a fixed window is larger than the before it. If your organization wrote an AI plan any time before the winter of 2025, described a system that could do a couple of hours of work with a fairly high error r A few months later, you can get sixteen hours or more of work from a single promp This is why AI keeps feeling like it is making leaps, even though it is a curve on a

graph, we keep experiencing a steady doubling of capability as a series of shocks. W
are very bad at feeling exponentials from the inside, and we are currently inside one



I think this also explains the turbulence around AI better than the usual stories abo
hype. AI is not capable of being a real cybersecurity threat until suddenly it is, caus
sudden and improvised policy changes at the highest level of government. Markets
discount whether AI might threaten to undermine a business model until suddenly
can, leading to massive swings in stocks. These lurches these get read as signs of ar
immature field that will eventually settle into something stable. I don't think it is
going to settle anytime soon. The instability is what happens when institutions tha
move at the speed of people (or worse, committees) try to track a capability curve th
is very much not human in nature. And as long as we are on some sort of exponenti
and for as long as it lasts, the gap only widens.



744 Likes · 67 Restacks

← Previous Next

Discussion about this post

Comments Restacks

Write a comment...

One Useful Thing reply r

Scenarica Jul 1

Everyone has been asking which professions AI replaces. The data says the question is wrong. A l
who's never written code was as effective as a software engineer working in an unfamiliar domai
Profession didn't predict success. Understanding the problem did. AI doesn't replace professions
makes professions porous. The walls between "coder" and "non-coder" dissolve. The walls betwe
"expert" and "novice" get higher, because the thing that makes the agent useful is knowing whet

the output is right, and knowing whether the output is right comes from ten years of doing the v hand.

The closing observation about institutions moving at committee speed while capability moves at exponential speed is the reason every AI policy written before winter 2025 is already obsolete. Bu there's a human version of the same problem. Individuals also move at human speed. The persor spent six months building a workflow around a chatbot just learned it can run for fourteen hours autonomously, which means the workflow they built is already the wrong workflow. The exponer doesn't just outrun institutions. It outruns the people trying to keep up with it, and the gap betw what the tool can do today and what the person learned to do with it last quarter is the new forn obsolescence nobody prepared for.

LIKE (32) REPLY

2 replies

Anne A'herran Jul 1

When talking about "real work" I see that you all mean coding, more or less. Whereas "Real work me and other humans is gardening, Planting, pulling weeds, harvesting our veges, cooking, walk dog, attending committee meetings, baking and putting out the supper, so on. For my adult chil female, it's breast feeding the Bub, dressing the kids for school, so on and so on, you get the drif robots can breast feed now? Wipe up the dog piddle? Eat the cake? TAste test it? Enjoy the cake the leftover cake in the fridge under gladwrap for tomorrow and put it in the lunchbox? No, but can code.

LIKE (12) REPLY

4 replies

95 more comments...