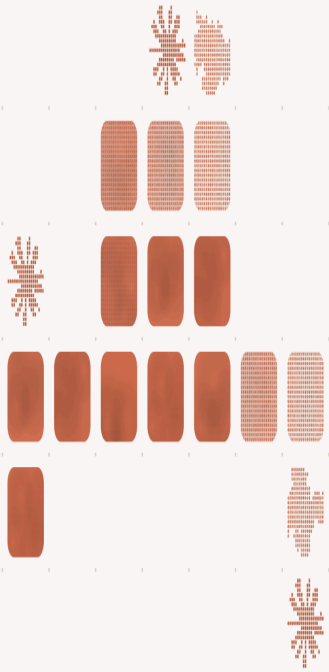
When AI builds itself

Our progress toward recursive self-improvement, and its implications.

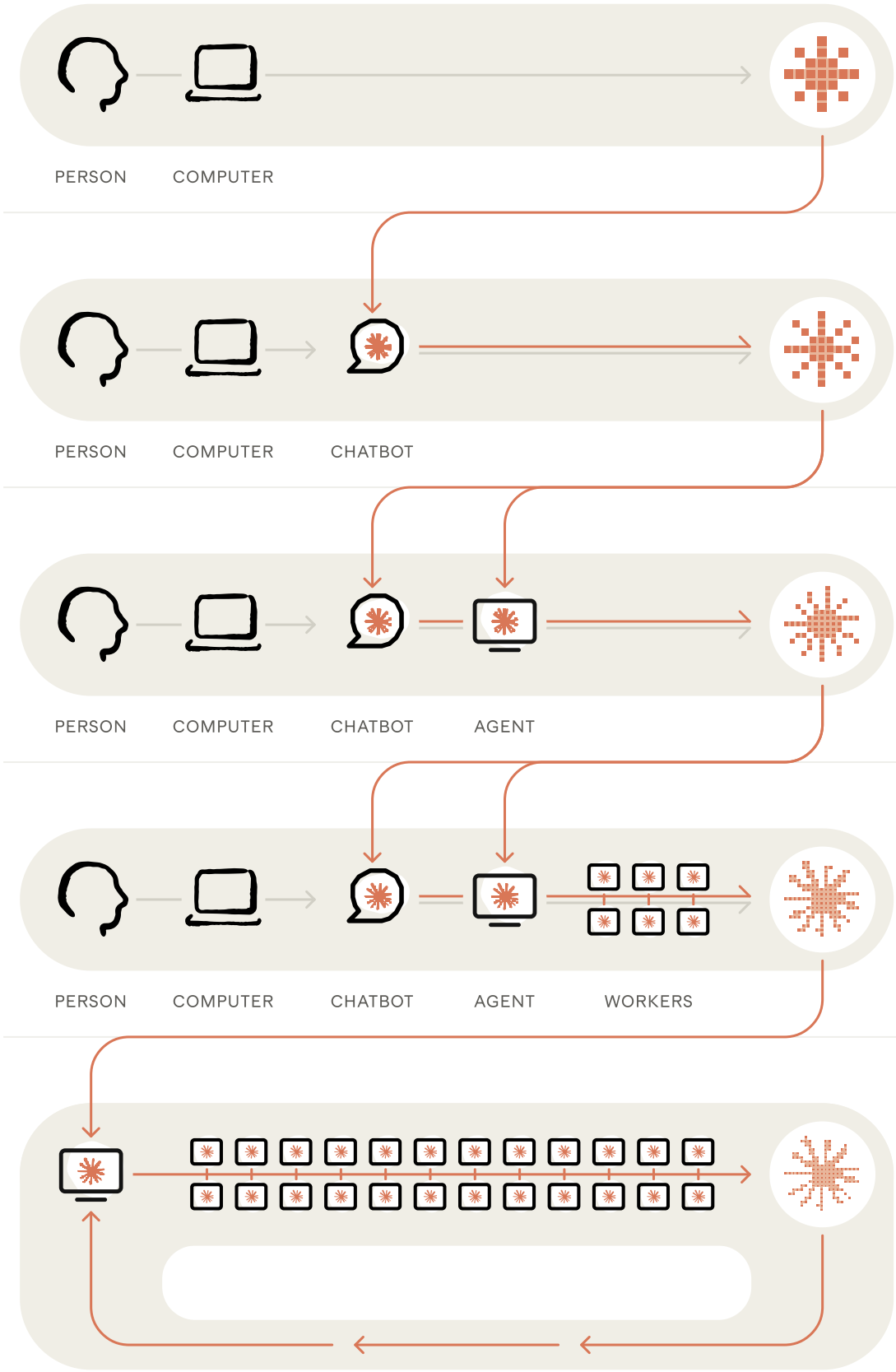


For most of AI's history, humans drove every step in its development cycle. But at Anthropic, we are delegating a growing share of AI development to AI systems themselves, which is speeding up our work.

Taken far enough, and given enough compute, that trend points to an AI system capable of fully autonomously designing and developing its own successor. This is called *recursive self-improvement*. We are not there yet, and recursive self-improvement is not inevitable. But it could come sooner than most institutions are prepared for.

Using public benchmarks and previously unreported data from within Anthropic, The Anthropic Institute is showing that AI is already accelerating the development of AI systems. To take just one example: today, Anthropic engineers on average ship 8x as much code per quarter as they did from 2021-2025.

The technical trends discussed in this piece suggest that AI systems are going to become much more capable in coming years. These trends have huge implications. AI that can build itself would be a major development in the history of technology—one that could bring enormous good for the world in science, healthcare, and beyond. But full recursive self-improvement also might increase the risks of humans losing control over AI systems. If systems are capable of fully building their own successors, the ways we secure them, monitor them, and shape their behavior all grow much more important.


PERSON
COMPUTER
PERSON
COMPUTER
CHATBOT
PERSON
COMPUTER
CHATBOT
AGENT
PERSON
COMPUTER
CHATBOT
AGENT
WORKERS

2021–2023

Building the first Claude

In the early days, work at Anthropic looked like work at any other tech company: people writing code and docs on laptops.

2023–2025

Chatbots

People used early chatbots to help with parts of the process, like generating short code snippets and copying the output into text editors.

2025–2026

Coding agents

As the agents became more capable, they were able to write and edit code on their own, sometimes entire files.

TODAY

Autonomous agents

Agents can now run code themselves and delegate hours of work to other agents.

20XX?

Closing the loop

In the future, agents could become capable enough to build and train models themselves. If this happens, future versions of Claude could be continuously improved by Claude itself.

Evidence from the outside world

The rate at which AI models improve is accelerating. The length of tasks that they can reliably complete on their own has been doubling roughly every four months, up from an earlier trend of doubling every seven months. In March 2024, Claude Opus 3 could complete software tasks that take humans about four minutes to complete. A year later, Claude Sonnet 3.7 managed tasks that took about an hour and a half. A year after that, Claude Opus 4.6 managed 12-hour tasks.[1] If this trend holds, tasks that take a skilled person days could come into range this year. In 2027, AI systems could be capable of tasks that take a person weeks.

The same pattern appears on coding and research benchmarks. Benchmarks measure the performance of models in a given domain, and they're "saturated" when models achieve close to 100% performance.[2] SWE-bench is a standard test of real-world software engineering: it hands a model an actual open-source codebase and a real bug report, and asks it to write a code change that fixes the issue and passes the project's own tests. Models have gone from scoring in the low single digits to saturating the benchmark in two years.

CORE-Bench tests whether a model can reproduce existing research, a prerequisite for them to conduct original research. It gives an AI model the code and data behind a published paper, and asks it to rerun everything and confirm it can replicate the paper's results. AI systems went from succeeding at reproducing the results roughly 20% of the time in 2024 to saturating the benchmark fifteen months later. METR, which runs the benchmark measuring how well models can complete long-duration tasks, found that Claude Mythos Preview could work for "at least" 16 hours and was "at the upper end of what [METR] can measure without new tasks."

Public benchmarks say a lot about the capabilities of these systems. But they can't reveal the impact AI systems are having on speeding up AI development itself. For that, we need direct evidence from within AI companies like Anthropic.

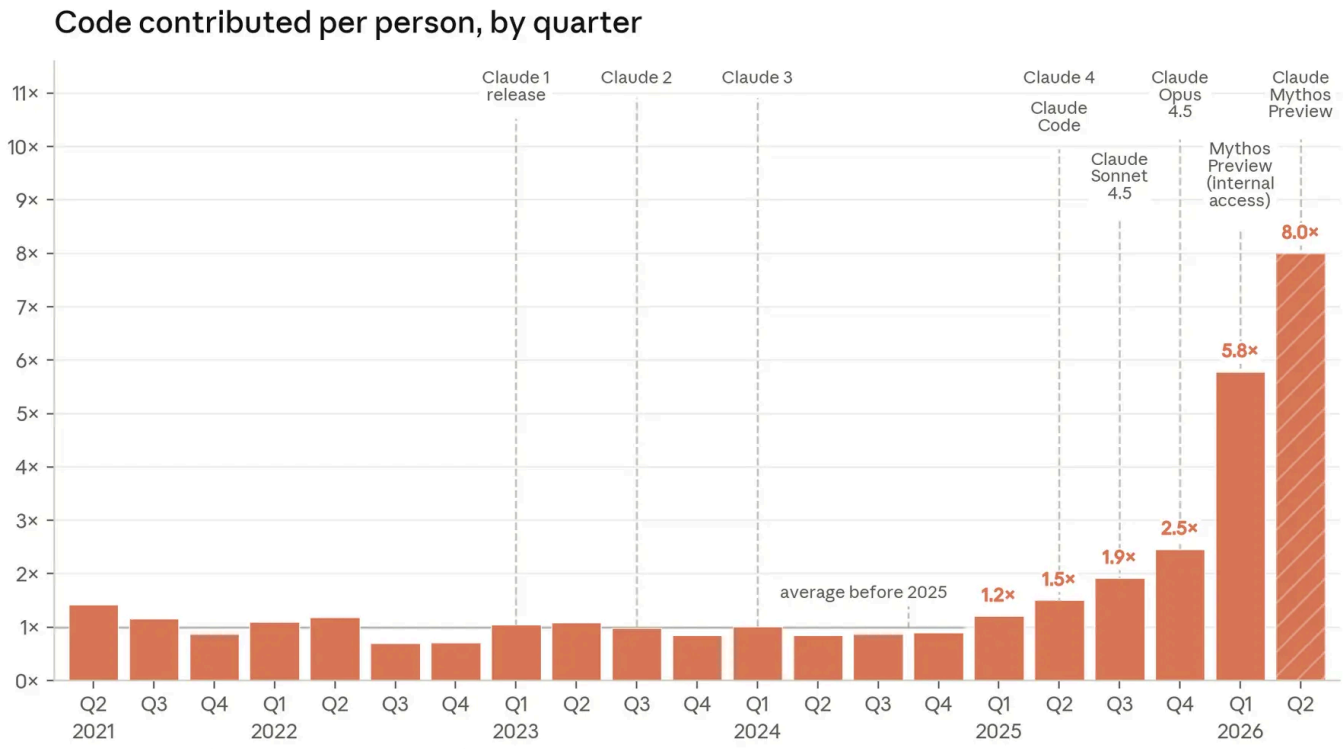
Evidence from within Anthropic

Building a frontier model takes two broad categories of work. There is *engineering*: writing the code, standing up the infrastructure, and overseeing the model training. And there is *research*: deciding what experiments to run, interpreting what comes back, and figuring out which ideas to try next.

Across both engineering and research, the picture is consistent. In engineering, Claude can be handed an underspecified problem and figure out how to solve it; humans supply the goal, but they no longer need to supply the method. In research, Claude can already match or outperform skilled humans at executing a well-specified experiment. However, large performance gaps persist when it comes to Claude exercising judgement in choosing goals in both engineering and research. That's the gap between AI today and a future system that could autonomously design its own successor.

It's common for employees at Anthropic to receive more open-ended and important tasks as they gain more experience. Early on, they execute a task someone else specified, like, *"The export button isn't working, please fix it."* With experience, they're handed a goal and design the approach themselves, such as, *"Investigate why the network slows down under heavy load."* At the most senior levels, they are deciding which problems are worth working on at all: *"What should the team build next quarter?"* We can use internal Anthropic data to see how far Claude has come in being able to handle these different kinds of tasks.

Claude writes a significant proportion of Anthropic's code. As of May 2026, more than 80% of the code we merge into Anthropic's codebase was authored by Claude.[3] Before Claude Code launched in research preview in February 2025, this number was in the low single digits. That shift also shows up in the amount of output per engineer. Lines of code merged per engineer per day stayed constant through Anthropic's first four years (2021-2024), then began to climb upward in 2025 when Claude began to run code rather than just suggesting it for an engineer to copy and paste. The slope steepened again in 2026 when models began to work autonomously over longer time horizons. These two inflection points are shown in the chart below. In the second quarter of 2026, the typical engineer was merging 8× as much code per day as they were in 2024.[4] This is because much of the code is written by Claude, with the engineer directing and reviewing, rather than typing it themselves.



Each bar is the average, over the days in that quarter, of lines of code merged per active contributor — shown as a multiple of the pre-2025 average. The hatched final bar is a partial quarter: it averages only the days observed so far, not a full quarter. Dashed lines mark public announcement dates. Per-PR line counts are capped at the 99th percentile; "active contributor" means a distinct author in the trailing twelve months.

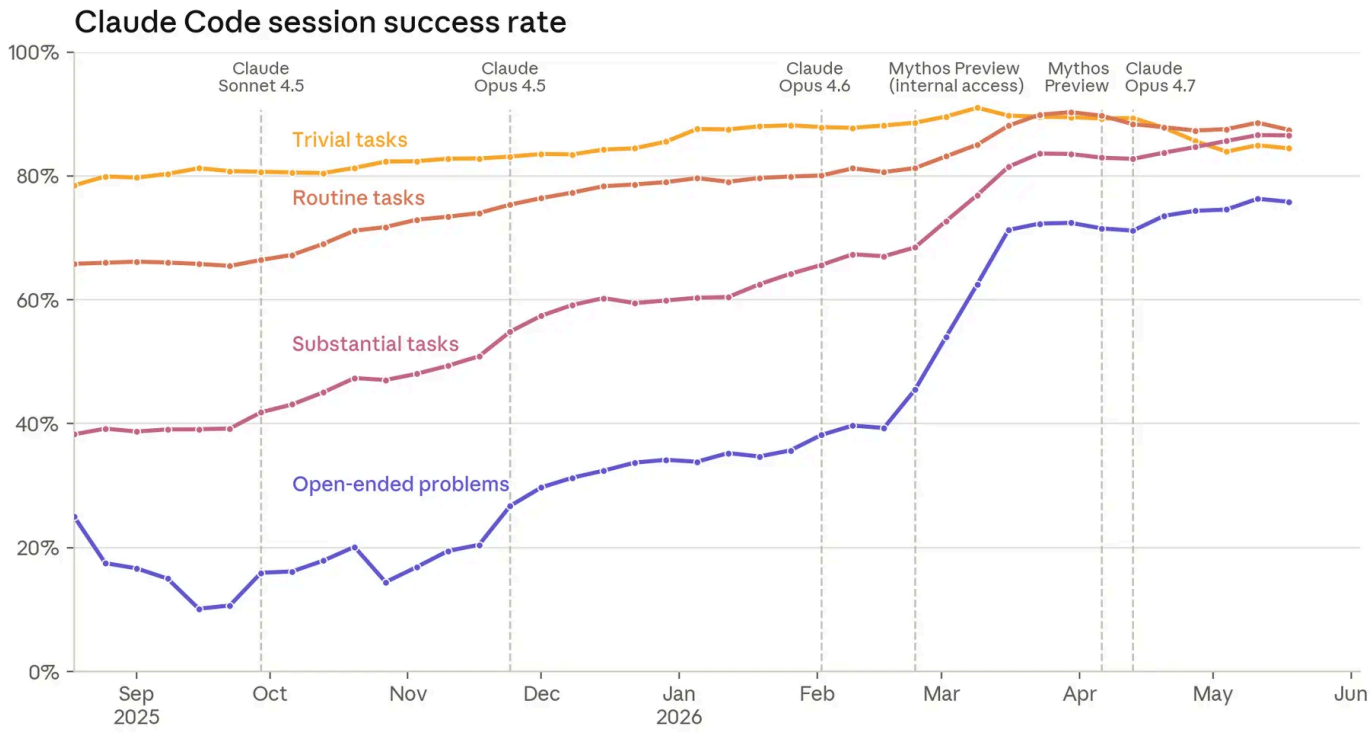
A caveat: Lines of code is an imperfect measure, as it measures quantity over quality. So *8× lines of code/engineer/day* in the second quarter of 2026 is almost certainly an overstatement of the true productivity gain. Nonetheless, it indicates an acceleration. At Anthropic, we don't reward people for how many lines of code they write; rather, team members are producing more code simply because they're using AI systems to write more code.

The increase in lines of code written lines up with subjective impressions of large productivity increases. In a March 2026 poll of 130 employees from across Anthropic research teams, the median respondent estimated that they produced around 4x as much output with Mythos Preview as they would have without access to any AI models, on the kinds of projects they would have been working on regardless.[5] We expect that the true degree of uplift in March was somewhat lower.[6] Nevertheless, we find the overall claim plausible, and in line with our other observations: a significant fraction of Anthropic technical staff is accomplishing their core work multiple times faster than they could without AI assistance.

We also see evidence that people at Anthropic are using Claude to do work that simply wouldn't have happened otherwise, like building exploratory tooling and addressing long-deferred cleanup. For example, in April 2026, Claude shipped over 800 fixes that reduced a class of API errors by a factor of one thousand. The engineer overseeing Claude estimated that a human would have taken four years to complete this work; solving other people's bugs is slow and painstaking, and humans struggle to hold that much unfamiliar context in their head at once.

> "I started leaning hard into Claudifying about a year ago. That's been a crazy adventure and it's now been ~5 months since I last wrote any code myself."
>
> Anthropic employee*

The code that Claude writes is "good" and improving. "Good code" means two things: it works, and it is written in a manner that allows another engineer to understand it and build upon it. On the first criterion, the evidence is clear. The rate at which Anthropic staff correct, redirect, or take over mid-task from Claude has been falling steadily for a year, including on the most complex and open-ended tasks. This means problems with no clear specification, where the engineer isn't sure what the answer looks like. This is evident in Claude's success rate over time on tasks of different difficulties, as shown in the graph below. Claude writes code that works.



Internal Claude Code sessions at Anthropic. Weekly, 4-week trailing mean. LLM-judged success.
Task complexity is assigned by an LLM classifier that reads a summary of each session and picks one of four levels from a fixed rubric.

How to read this: *Session success is determined by a Claude judge; a session is deemed successful if the Claude Code agent clearly succeeded at the user's tasks without requiring corrections. Changes in workloads can lead to short-term fluctuations in success rates.*

On the most open-ended tasks, Claude's success rate reached 76% in May 2026, up 50 percentage points in six months. To give an example of tasks in this difficulty tier, a routine upgrade began crashing tens of thousands of training jobs. An engineer pointed Claude at the live incident with little more than some text content and cluster access. Working through the running jobs and testing one environment setting at a time, Claude isolated the single obscure debugging flag that was triggering the crash, reproduced it reliably, and confirmed a fix. In about two hours, Claude delivered what would normally be two to three days of work.

The second criterion is writing code that another engineer can understand and build on. Here the gap between humans and AI persists, but is closing fast. There isn't full consensus among staff at Anthropic, but many believe that the Claude-written code was still worse in quality than human-written code at Anthropic in late 2025, and is roughly at parity today. We expect it to be better within the year.

This has changed the way that Anthropic now reviews its own code. Proposed changes to our codebase are now read by an automated Claude reviewer that looks for bugs, security flaws, and other defects before it can merge. Using this tool, we ran a retrospective analysis, and found that an automated Claude review of every change to our codebase would have caught roughly a third of the bugs behind past incidents on claude.ai before they ever reached production. The engineers who wrote that code are among the best in the world at building these systems. Claude is now catching the mistakes that they missed.

"Claude-written code was somewhat worse than human-written code at Anthropic in late 2025, is roughly at parity today, and we expect it to be strictly better within the year."

Claude is good at running experiments to hit a goal that someone else has set. Every time Anthropic releases a model, we run the same test: we give Claude some code that trains a small AI model, and ask it to make that code run as fast as possible while still passing the same correctness checks. The goal and the success metrics are fixed in advance, so Claude's job is to find speedups by rewriting the code, running it, timing it, and repeating. It's a miniature version of an experimental research loop. In May 2025, Claude Opus 4 averaged a ~3x speedup over the starting code. By April 2026, Claude Mythos Preview was achieving ~52x. For calibration, a skilled human researcher would need four to eight hours to reach 4x.[7] In this part of the research workflow—optimizing steps within a clearly defined experiment—Claude has gone from super helpful to superhuman in under a year.

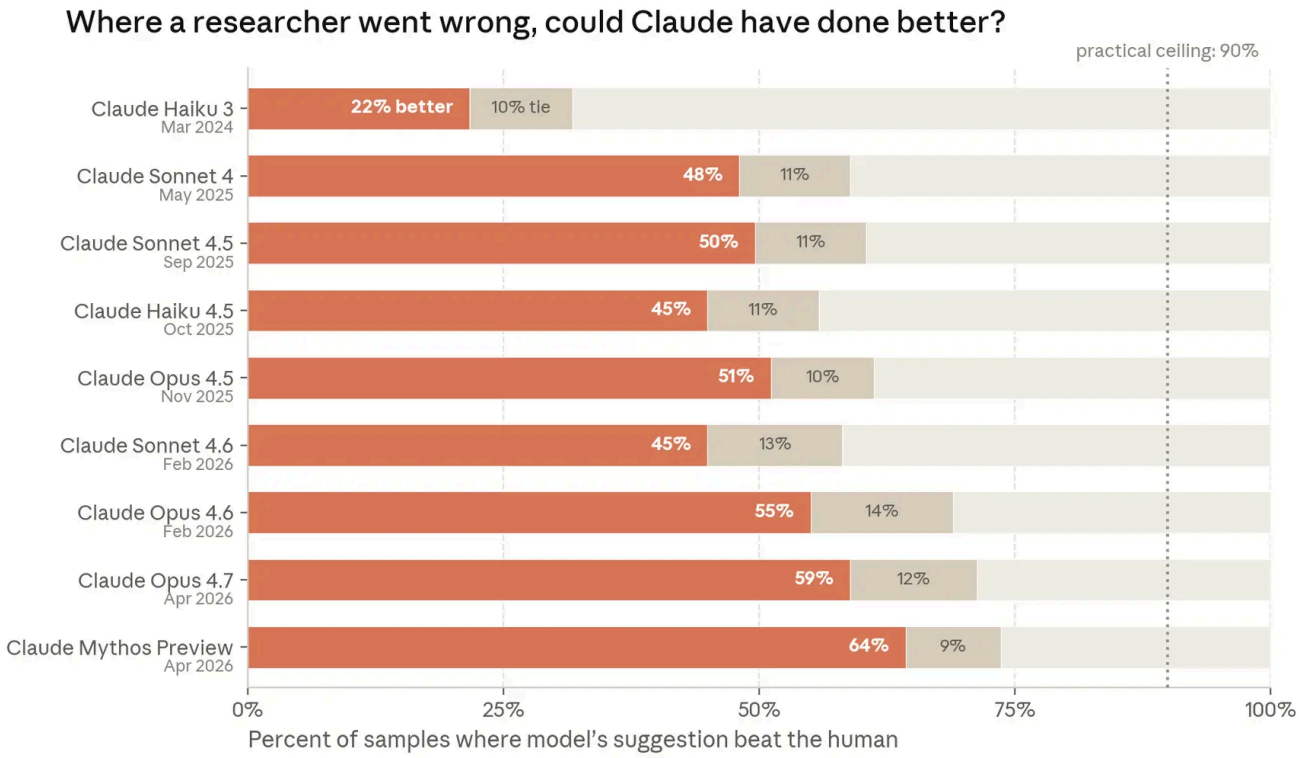
"The shape of stuff today is roughly 'humans have ideas, and the models are able to implement, test and evaluate them an [order of magnitude] faster than before.'"

Claude is getting better at proposing its own experiments. In April 2026, Anthropic published the first demonstration of Claude running an open-ended research project end to end. Claude-powered agents were given an open problem in AI safety—roughly, *can a weaker model reliably supervise a stronger one?*—and were left to solve it. This involved proposing hypotheses, testing them, sharing findings with parallel agents, and iterating. The task has a clear performance "floor" and "ceiling": the floor is how well the weak supervisor would do on its own; the ceiling is how the strong model does when trained on correct answers. Two human researchers, over about a week, recovered roughly 23% of that gap; the agents recovered 97% over 800 cumulative hours and used roughly $18,000 in compute. There are some caveats to this work; the result didn't transfer cleanly to production-scale models, and humans still chose the problem and created the scoring rubric. But within those bounds, the agents designed every experiment themselves. Direction-setting was the only meaningful role a human played.

> "Claude did all of this with pretty minimal help from me over the course of 1-2 days. I think if [a junior colleague] came back to me with results like this in the same span of time, I would be mildly impressed. The future is now."

Claude is getting better at steering research sessions towards research findings. We examined real Claude Code sessions (between January and March 2026) where Anthropic researchers were working with Claude on an open-ended investigative problem, like figuring out why a training run kept crashing, or why a model scored poorly on a benchmark. In each case, we found a moment where the researcher took a detour: they pursued a direction that sent the session sideways before it eventually got back on track. We then showed various Claude models *only* the work from before the session went off-course and asked what it would do next. A separate Claude that was able to see how the session eventually turned out then judged whether the AI or the human suggested the better next step.[8]

Because we deliberately picked moments (n=129) where we know the human's choice had room for improvement, this isn't a like-for-like comparison between model and human judgement. What these moments give us is a set of realistic, challenging situations where the right next step is not obvious, and where the human's choice serves as a useful yardstick to compare model performance over time. On this measure, our best model in November 2025 (Opus 4.5) beat the human choice 51% of the time; in April 2026 (Mythos Preview), this grew to 64%. The day-to-day work of research is largely a chain of these next-step decisions, making this a relevant measure of the model's ability to eventually run an investigation of its own. We view this result as an early signal that AI systems are getting better at making the kinds of judgement calls that AI research depends on.



The study covers 129 internal Claude Code research sessions, each cut at a turn where the human's next direction had room for improvement. Model proposes an alternative; a judge that knows the session's eventual key findings picks the better move.

How to read this: *The practical ceiling line measures an "ideal" answer written by a model that could see the whole session (including how it ended).*

"The comparative advantage of humans as of right now is still in seeing the bigger picture and thinking beyond

the confines of the immediate task."

What might the future of work at Anthropic look like?

The evidence suggests that the human role is narrowing at each step in the AI development process. Once human- and AI-authored code quality reach parity, humans will stop writing code entirely, and shift to only reviewing it. But if they can't review code as quickly as Claude can generate it, human review will become the bottleneck to AI development. Similarly, once Claude can run experiments, the question shifts towards "Which of these experiments is worth running?" Put simply: the *doing* (i.e., writing the code, running the experiment, producing the result) now costs almost nothing in human time, even if it still has costs in compute.

An area of human comparative advantage, for now, is research taste and judgment, including choosing which problems matter, which results to trust, and when an approach is a dead end.

"Work (and life) ran on a gift economy of small favors between humans. 'Can you help me get this script running?' [...] each one created a little debt, a little mutual awareness. [Claude is] faster, it creates zero debt, but each of these is a lost bid for human collaboration."

"On days where everything works well, I can't help but think nothing I do matters, everything is automated and better and faster than I ever will be. But then there are days where everything breaks and I don't understand why and I realize I have no idea what I've been up to anymore."

What if we're wrong?

A natural objection to the evidence presented above is that the work that is still in human hands—choosing which problems to work on—is what matters most. Without that judgment, Claude is a capable assistant, but not a system that could drive AI progress on its own.

It is genuinely unclear whether today's training methods and architectures could unlock that capacity. But AI is rarely advanced by "eureka!" moments. There have been a few of these in AI's recent history, like the Transformer architecture, or mixture-of-experts models, but paradigm-shifting ideas arrive years apart. In between, most progress is incremental: we scale something up, see what breaks, fix it, and try again. That is exactly the kind of workflow Claude now excels at. Edison said that genius is 1% inspiration and 99% perspiration. But we see perspiration becoming increasingly automated. It's becoming clear that much of what advances the frontier is automatable; large-scale research progress is mostly a function of tools and resources, which dictate how fast you can run experiments, how many you can run at once, and how quickly you can get results.

Even if we suppose that Claude never achieves good research taste, a conservative reading of our evidence still implies compounding acceleration. If humans spend most of their time on the single-digit fraction of work that is direction-setting, while Claude handles the rest, that means each engineer or researcher is steering far more work than before. The evidence we see suggests that people at Anthropic are both moving faster and covering a broader surface. In practice, this means that AI already makes Anthropic move much faster than it did before the advent of effective AI tools.

The less conservative reading is that the early evidence on Claude's improving research judgment—narrow as it is today—is an indicator that this capability is improving as well. "Research taste" might be just another AI capability that AI systems fail at for a time, then get good at. We've seen a similar pattern with other qualitative skills, like AI systems being able to explain why a joke is funny, demonstrate theory of mind, and solve linguistic riddles.

Possible futures

What happens next depends on two things: whether the trend continues, and what we choose to do if it does. We can imagine at least three future scenarios:

1. **The trend stalls, but today's AI capabilities are widely diffused.** This article features many exponential trajectories. But these trajectories may actually turn out to be S-curves. We may be approaching the bend in the curve, where returns to scale diminish and the line straightens, then flattens. The judgment that separates a competent researcher from a great one might be a capability that cannot come from scaling up training inputs like compute and data. If so, getting past this bottleneck would require a new idea, like an architectural approach that supplants the Transformer architecture that all current frontier models use.

   Alternately, the binding constraint to AI progress could be in the supply chain, not the model: advancing and diffusing the frontier may require more energy and compute than presently exists. The pace of chip fabrication, grid expansion, or interconnect bandwidth may be the constraint, rather than intelligence itself. We also cannot rule out an exogenous shock to the AI ecosystem that dramatically slows things, like a sudden diminishment in the supply of compute or electricity, either of which would slow progress and make forward investment by labs more expensive. Or we may not be anticipating some other barrier to progress.

   Even if model capabilities were frozen at today's level, we would expect major changes to occur in the world. Project Glasswing is one early sign: in its first weeks, Mythos Preview found more than ten thousand high- and critical-severity software vulnerabilities across the world's most important systems—enough that the bottleneck in cyber defense has already shifted from finding vulnerabilities to patching them fast enough. And we are still early in the diffusion of today's models into the wider economy, where a 100-person company can increasingly do the work of a 1,000-person one, because each employee will sit atop a pyramid of agents.

   We include this scenario for completeness, but we don't believe it's likely. Every capability we can measure, including those that feel "squishier," like quality of code and success on open-ended tasks, has so far followed the same curve. We have not yet seen that curve bend. Of the three futures we

consider, this one would give governments and societies the most time to adapt. We are more worried about the next two, which would move faster and leave far less room for preparation.

2. **AI labs continue to see compounding efficiency gains.** In this scenario, AI development becomes substantially automated, but humans continue to set research directions and judge results. Organizations that use AI systems would become much more efficient as time goes on, so we could expect to see significant productivity multipliers on each person in this organization. 100-person companies could do the work of 10,000- or 100,000-person organizations. This would revolutionize knowledge work and government services, but could also be turned to harmful ends, from authoritarian surveillance of whole populations to influence operations that tailor manipulation to each individual and run at a scale no human team could match. The role of humans at companies like Anthropic would shift. People would partner with AI systems to scale up research and generate new insights, and together they would build the systems needed to verify that AI outputs can be trusted.

   The evidence we've laid out here suggests that we're likely heading into this scenario. But speeding up one part of a process often just shifts the bottleneck elsewhere: overall pace is capped by the parts that haven't sped up. In computing, this is known as Amdahl's law, and the same logic can apply to organizations. Anthropic has already encountered one signature of Amdahl's law: as we've begun to push more code around the organization, human code review has become a new bottleneck.

   We've also encountered this friction outside engineering. There has been an explosion of new ideas, initiatives, tools, and simulations, as a result of Anthropic employees working with highly capable models—far more than we have the capacity to pursue. The rate at which organizations can spot and fix these bottlenecks may be a skill that improves over time, and it may become the most important skill for any organization.

3. **AI systems themselves become capable of full recursive self-improvement, and begin building their successors.** If technical trends in advancing capabilities continue, *and* AI systems are able to develop the

capabilities inherent to transformative human ingenuity, then it is plausible that AI systems could design and refine themselves.

In this world, the pace of progress in AI development becomes determined entirely by the availability of compute (or the speed of discovering various efficiencies in algorithmic training or inference) for AI systems. Humans play a substantially diminished role in their development, likely moving most of our effort towards oversight, validation, and verification of an expanding "virtual lab" run by AI systems. We expect that systems capable of automated AI research and development would have skills that would transfer to the rest of science, allowing them to begin to revolutionize other fields.

How the alignment problem gets solved—or not—in this future is something we are least certain about. Models could prove to be sufficiently aligned and capable enough of research taste that they discover and implement novel solutions that we have not yet reached. They could also be sufficiently wise to halt development if not. Alternatively, the rare occurrences of misalignment present in today's models could compound as the models build their successors, growing more frequent but less understood until we lose control of them. It's possible that we can't build, integrate, and verify the tools that we'd need to understand which trendline we are actually on.

We do not have good intuitions for what this world would look like, because our economy is currently driven by humans and human-built tools. By its nature, a world driven by fast recursive self-improvement could become dominated by the self-improving model as its capabilities fully eclipse those of humans and the model proliferates across the broader economy. It is difficult to predict what the economy looks like if human labor stops being competitive.

Even if model development became fully automated and recursive, we can't predict what that would mean for most humans' daily lives. Amdahl's law applies here as well. Recursive intelligence could lead to achieving many of the benefits outlined in Machines of Loving Grace, quickly in some domains. We expect that embodied intelligence (i.e., robotics) might quickly follow recursive intelligence, and follow a similar path of increasing returns at

decreasing cost. More powerful intelligence might help us build things in the physical world more quickly, run more productive clinical trials of lifesaving drugs, and develop novel forms of coordination.

But achieving recursive improvement alone does not suggest an immediate change in how industrial production occurs, societies organize, or markets function. More intelligence can't learn what a drug does over decades of use, can't hold elections sooner than a constitution dictates, and can't turn a stranger into an old friend in a weekend. For most people, the felt pace of this future will still be set by the bottlenecks, even if the laboratory upstream runs at the speed of compute. That collision, where recursive intelligence building itself ever faster meets the world of humans, relationships, and governance, is another part of this future we can't predict.

What should we do?

If it were possible to effectively slow the development of this technology to give ourselves more time to deal with its immense implications, we think that would likely be a good thing. But if a slowdown simply lets the least cautious actors catch up technologically, it could leave everyone less safe. Without a global coordination mechanism, companies and governments will have to make difficult decisions about safety while under competitive and geopolitical pressures.

We believe it would be good for the world to have the *option* to slow or temporarily pause frontier AI development to enable societal structures and alignment research to keep up with the advance of the technology. The Anthropic Institute will conduct research—in collaboration with many others—and take actions to help build the systems that a credible slowdown or pause would require. These systems would enable frontier AI developers to verify that others globally have actually stopped or slowed, and that a bad actor could not use the auspices of a coordinated slowdown to jump ahead in secret. If such systems existed, we expect that we would slow down or temporarily pause, if other developers at or near the frontier also did so in a verifiable manner.

A meaningful slowdown or pause would require multiple well-resourced labs at or near the frontier, in multiple countries, agreeing to stop under the same

conditions. It would also require that each can verify that the others have actually stopped. Due to the unique characteristics of AI systems, the detectability (a lower standard than verifiability) element of this arms control problem is much more challenging than with other technologies. Training runs are far easier to conceal than missile silos, their inputs are general-purpose, and the incentive to defect quietly is enormous, because whoever continues while others pause could inherit the lead. A credible pause also has to specify what triggers it, what lifts it, and who adjudicates.

None of this is necessarily impossible in principle—the world has built verification regimes for other complex technologies (e.g., the Intermediate-Range Nuclear Forces Treaty)—but those regimes took decades to build both the infrastructure and the trust. We don't have that long. A unilateral pause by one lab, by contrast, is achievable immediately, but accomplishes much less: it would change who the front-runner is, but it would not create the wider deliberative process that is currently missing.

In the coming months, we will organize conversations where policymakers, researchers, civil society, and other AI companies can help answer some of the questions this piece raises, especially around full recursive self-improvement and how to create better options for coordination and deliberation. We'll publish what comes out of it. The window to investigate the questions together is here, and people outside AI companies should be involved in this deliberation.

Marina Favaro and Jack Clark co-authored this piece, with editorial support from Santi Ruiz. Shan Carter, Romello Goodman, and Nikki Makagiansar created the visuals from data collected by Brian Calvert and Jun Shern Chan. Daniel Freeman, Jim Baker, Max Young, Sarah Pollack, Francesco Mosconi, Holden Karnofsky, Andy Jones, Kevin Troy, Anton Korinek, Meg Tong, Andrew Ho, Dan Altman, Drake Thomas, Jack Shen, Sasha de Marigny, and Avital Balwit provided feedback.

FOOTNOTES

1. METR's key measure tells you the time horizon over which AI systems can be 50% reliable at a basket of tasks, though the trendline looks the same at 80% reliability.
2. Especially as they shift toward more open-ended formats and more difficult tasks (e.g., Olympiad-level mathematics), benchmarks often saturate below 100% due to errors in the question and answer sets like ambiguous problem statements and unsolvable questions.
3. Anthropic leadership have publicly estimated that 90% or more of our code is written by Claude, including scripts and experimental code. Our >80% figure measures the share of lines merged to production that can be attributed to Claude. This is a more conservative measurement in two ways: our attribution pipeline has gaps, and the lines not attributed to Claude include auto-generated code and other artifacts that were not hand-written by humans either.
4. This surge in code production is straining the infrastructure everyone shares. GitHub—the platform most of the world's software is built on—saw roughly one billion code commits in all of 2025; by mid-2026 it saw 275 million a week, on pace for roughly 14 billion over the year. The company's COO has said that it is "pushing incredibly hard" on capacity just to keep up.
5. Additional details on the methodology of this survey are discussed in section 2.3.5 of the Claude Opus 4.7 System Card.
6. Many respondents may not have thought carefully about how to account for various biases or subtleties in the question definition, and recent research by METR shows that developer estimates of AI productivity uplift can be overestimated.
7. How large the speedup gets depends heavily on how much room for improvement the starting code leaves, and it should not be read as a real-world training speedup. So the absolute multiple is not the figure to anchor on here. What is more informative is the like-for-like comparison that this experimental setup makes possible, both across models (~3x to ~52x over the past year) and against a skilled human (~4x in four to eight hours on the same task).
8. As a check on judge bias, we ran the same test on a separate set of 127 moments where the human's next move was already strong (as opposed to the original set, where the human's direction had room for improvement). There, the models' suggestions were judged better only about 20% of the time.

* Quotes from Anthropic employees throughout this article are drawn from internal discussions and used with permission. They reflect individual views as of May 2026, not official company positions.