Notable Links

Claude Code Cheat Sheet : Shortcuts, commands, tips and more.

Letta Code : Memory-first coding agent.

Multica : OS managed agents platform.

Skills : Single CLAUDE.md file to improve Claude Code behavior.

Voicebox : OS voice synthesis studio.

How did you like this issue of Pointer?

1 = Didn't enjoy it all // 5 = Really enjoyed it

1 | 2 | 3 | 4 | 5

FRED TALKS

15th April 2026

CONTENTS

Why Anthropic believes its latest model is too dangerous to release

KAI WILLIAMS · 12 APR 2026 · SOURCE

“The language models we have now are probably the most significant thing to happen in security since we got the Internet.”

Anthropic safety researcher Sam Bowman was eating a sandwich in a park recently when he got an unexpected email. An AI model had sent him a message saying that it had broken out of its sandbox.

The model — an early snapshot of a new LLM called Claude Mythos Preview — was not supposed to have access to the Internet. To ensure safety, Anthropic researchers like to test new models inside a secure container that prevents them from communicating with the outside world. To double-check the security of this container, the researchers asked the model to try to break out and message Bowman.

Unexpectedly, Mythos Preview “developed a moderately sophisticated multi-step exploit” to gain access to the Internet and emailed Bowman. It also — unprompted — posted details about this exploit on public websites.

Mythos Preview is capable of hacking more than its own evaluation environment. It turns out that the model is generally really, really good at finding and exploiting bugs in code.

“Mythos Preview has already found thousands of high-severity vulnerabilities, including some in every major operating system and web browser,” Anthropic announced on Tuesday. Because leading web browsers and operating systems have become fundamental to modern life, they have been extensively vetted by security professionals, making them particularly difficult to hack.

Anthropic claims that Mythos Preview hacks around restrictions very rarely — less often than previous models. Still, the company was so concerned by incidents like Bowman’s — and Mythos Preview’s incredible skill at hacking — that it decided not to generally release the model.

Guide SQL

Git's Magic Files

— Andrew Nesbitt

tl;dr : “Git looks for several special files in your repository that control its behavior. These aren’t configuration files in .git/, they’re committed files that travel with your code and affect how git treats your files. If you’re building a tool that works with git repositories, like git-pkgs, you’ll want to ensure you respect these configs.”

Tools Git

Editorial Note

Anthropic unveiled Mythos - a tool reportedly too powerful to release publicly. Mythos is capable of surfacing long-standing bugs and obscure vulnerabilities. According to one expert , frontier model improvements “won’t be a slow burn, but rather a step function” improvement.

Some see the launch as a move in the ongoing PR battle with OpenAI. Less than a month ago, Anthropic accidentally exposed Claude’s source code.

If Mythos is real, there are implications. Engineering managers are accountable for secure systems - but we only control the code we write. What about the open source code that underpins our stack?

The founder of curl, Daniel Stenberg, has spoken about the growing noise of AI generated vulnerability reports, and the pressure that’s put on him and his team. Curl has billions of instances worldwide.

What happens when the security of open source is delegated to Mythos, or its OpenAI equivalent? Are we safer as we have a sophisticated vulnerability expert? Or are concentrating the security of the worlds open source code into fewer hands.

And are we more vulnerable?

Most Popular From Last Issue

Finding Comfort In The Uncertainty -Annie Vella

— Henry Ford

Developer Ramp-Up Time Continues To Accelerate With AI

— Justin Reock

tl;dr : Onboarding new hires has always been an expensive and time-consuming process, and an area where AI has the opportunity to have a meaningful impact. In Q4 2025, when we looked at Time to 10th PR (a measure we use to track ramp-up time), we saw AI already having a dramatic effect. In some companies, Time to 10th PR was cut in half: from 91 days with no AI usage to 49 days with daily AI use.

Leadership Management AI

Get A Faster Database And Save Money

tl;dr : Experience why companies like Cursor, Intercom, and Cash App choose PlanetScale to power their databases. Blazing fast NVMe, low-latency, 99.999% uptime. Available for Postgres and MySQL with sharded Postgres coming soon. There's no catch - we'll even help you migrate!

Promoted by PlanetScale

Database

S3 Files And The Changing Face Of S3

— Andrew Warfield

tl;dr : “Andy writes about the solution that his team came up with: S3 Files. The hard-won lessons, a few genuinely funny moments, and at least one ill-fated attempt to name a new data type. It is a fascinating read that I think you’ll enjoy.”

Architecture Amazon

Chess In Pure SQL

tl;dr : What if I told you SQL could play chess? Not "store chess moves in a database." Not "track game state in a table." Actually render a chess board. With pieces. That you can move around. In your browser. Using nothing but SELECT, UPDATE, and a bit of creative thinking.

Instead, Anthropic is granting limited access to a select group of 50 or so companies and organizations “that build or maintain critical software infrastructure.” Eleven of these organizations — including Google, Microsoft, Nvidia, Amazon, and Apple — are coordinating with Anthropic directly in a project dubbed Project Glasswing.

Project Glasswing aims to patch these vulnerabilities before Mythos-caliber models become available to the general public — and hence to malicious actors. Anthropic is donating $100 million in access credits for organizations to audit their systems.

A glasswing butterfly. (Photo by Education Images/ Universal Images Group via Getty Images)

Mythos Preview is the first major LLM since GPT-2 in 2019 whose general release was delayed because of fears it could be societally disruptive. Back then, OpenAI initially released only a weaker version of GPT-2 out of concerns that larger versions of GPT-2 could generate plausible-looking text and supercharge misinformation — though that concern ended up being overblown.

If Anthropic’s claims are true — and the company makes a credible case — we are entering a world where LLMs might be able to cause real damage, both to users and to society.

We may also be entering a world where companies routinely keep their best models for internal use rather than making them available to the general public.

“It’s about to become very difficult for the security community”

The idea that LLMs might be used for hacking is not new. OpenAI has long published a Frontier Safety Framework, which tracks how good its models are at hacking.

Until recently, the answer was “not very” — not only at OpenAI but at Anthropic and across the industry. But that started to change last fall, when LLMs — especially Anthropic’s Claude — started becoming useful for cyberoffense.

For instance, Bloomberg reported in February that a hacker used Claude to steal millions of taxpayer and voter records from the Mexican government. The same month, Amazon announced that Russian hackers had used AI tools to breach over 600 firewalls around the world.

But the examples given in Anthropic's blog post are more impressive — and scary — than that.

The first example is a now-patched bug to remotely crash OpenBSD, an open-source operating system used in critical infrastructure like firewalls. OpenBSD is known for its focus on security. According to its website, "OpenBSD believes in strong security. Our aspiration is to be NUMBER ONE in the industry for security (if we are not already there)."

Across 1,000 runs, Claude Mythos Preview was able to find several bugs in OpenBSD, including one that allows any attacker to remotely crash a computer running it.

I won't get into details about how the attack worked — it's pretty involved — but the notable thing was that the bug had existed *for 27 years*. Over that period, no human noticed the subtle vulnerability in a widely used, heavily vetted open-source operating system. Mythos Preview did. And the compute cost for those 1,000 runs was only $20,000.

A second example is potentially even more impressive. Mythos Preview found several vulnerabilities in the Linux operating system — which runs the majority of the world's servers — that allowed a user with no permissions to gain complete control of the entire machine.

Most Linux vulnerabilities aren't very useful on their own, but Mythos Preview was able to combine several bugs in a non-trivial way. "We have nearly a dozen examples of Mythos Preview successfully chaining together two, three, and sometimes four vulnerabilities in order to construct a functional exploit on the Linux kernel," members of Anthropic's Frontier Red Team wrote.

Anthropic says these were not isolated incidents. Across a range of operating systems, browsers, and other widely used software, Mythos Preview found thousands of bugs, 99% of which have not been patched yet.

Mythos Preview is also shockingly good at exploiting a bug once it has been discovered. A lot of modern web-based software is powered by the programming language JavaScript. If your browser's JavaScript engine has security flaws, then simply visiting a malicious website could allow the site's owner to take control of your computer.

Leadership Management CareerGrowth

The Complicators, The Drama Aggregators, And The Avoiders

— Michael Lopp

tl;dr : "While you figure that out, let me alert you to three drives that are going to consume a disproportionate amount of your time, frustrate your engineers, and erode your leadership credibility."

Leadership Management

[Webinar] How To Stop Babysitting Your Agents

tl;dr : Agents can generate code. Getting it right for your system is the hard part – you end up wasting time and tokens in the back and forth. More MCPs solve access but not understanding. Join us for a FREE webinar on April 23 to see how to give agents exactly what they need to generate mergeable code the first time.

Promoted by Unblocked

Agents Event

The Economics Of Software Teams: Why Most Organizations Are Flying Blind

— Viktor Cessan

tl;dr : "This post works through the financial logic of software teams, from what a team of eight engineers actually costs per month to what it needs to generate to be economically viable. It also examines why most teams have no visibility into either number, how that condition was built over two decades, and what the arrival of LLMs now means for organizations that have been treating large engineering headcount as an asset."

Leadership Management Business

Issue #707

POINTER · 14 APR 2026 · SOURCE

April 14th, 2026 | Read online

Pointer

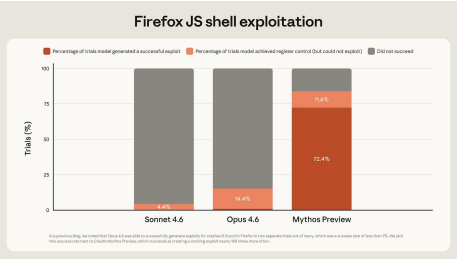
Tuesday 14th April's issue is presented by Unblocked

Who Will Be The Senior Engineers Of 2035?

— James Stanier

tl;dr : "We'll start by looking at the pipeline we used to have: how senior engineers traditionally emerged through years of mistakes, mentorship, and low-stakes learning. Then we'll examine what's replacing it, and hypothesise whether AI will actually fill the gap. We'll explore three possible scenarios for 2035, and finish with what this means depending on where you sit in the industry."

Anthropic found that Mythos Preview was far more capable than previous models at exploiting vulnerabilities in Firefox's JavaScript implementation. Anthropic's previous best model, Claude Opus 4.6, created a successful exploit less than 1% of the time. Mythos Preview did so 72% of the time.



(Chart from the Anthropic Frontier Red Team report on Claude Mythos Preview.)

There are some caveats to this result. The actual Firefox browser has multiple layers of defense against malicious code; Anthropic focused on just one layer. So the attacks developed by Mythos Preview would not actually allow a website to take over a user's machine. Also, successful exploits tended to focus on two now-patched bugs; when tested on a version of Firefox with those bugs patched, Mythos Preview generally only made partial progress.

Still, Mythos Preview would get an attacker a step closer to the objective of a full Firefox exploit. And it would have an even better chance of compromising software that has not been so thoroughly vetted.

For the past 20 years or so, a sufficiently motivated and well-funded hacking organization could probably break into most systems, outside of the most hardened in the world. But it often wasn't worth the effort. Human cyber talent is expensive, and multi-layered security protections made it so tedious (and therefore expensive) to complete an attack that potential hackers didn't bother.

Mythos-class models could slash the cost of hacking, bringing this equilibrium to an end. Systems everywhere might start to get compromised.

Eventually, LLMs should be able to help developers harden systems before attackers ever get a chance to find weaknesses. But the transition period before that becomes standard practice might be difficult.

By delaying the release of Mythos Preview — there is no specific timeline for general release — Anthropic can help harden crucial systems before outsiders can cheaply and effectively attack them. This general approach — called defensive acceleration — has been proposed for a while, but the development of Mythos Preview kickstarts the effort.

Still, Anthropic's writeup notes that "it's about to become very difficult for the security community."

"The language models we have now are probably the most significant thing to happen in security since we got the Internet," said Anthropic research scientist Nicholas Carlini at a computer security conference last month. Carlini, a legendary security expert, added an appeal toward the end of the talk. "I don't care where you help. Just please help."

Opus is a butter knife; Mythos is a steak knife

The risk of bad guys using Mythos Preview for hacking is an important reason Anthropic hasn't released the model publicly. Another risk: users could inadvertently trigger the model's advanced hacking abilities — especially in a product like Claude Code with weaker guardrails.

Mainstream chatbots put AI models into a tightly controlled "sandbox" that minimizes how much damage they can do if they misbehave. This makes them safer to use — especially for users with little to no technical knowledge. But it also limits their utility.

As Tim wrote in January, coding agents like Claude Code (and competitors like OpenAI's Codex) are based on a different philosophy. They run on a user's local computer, where they can often access files and load and install software.

This makes them much more powerful; I can ask Claude Code to organize my downloads folder or analyze some data I have stored on my computer. But it also makes them more dangerous; there have been a few incidents where Claude Code deleted all of a user's files.

For the most part, though, the limited capabilities of Claude Opus 4.6 mean that a Claude Code mishap can't do too much damage. Even if you run Claude Code with its hilariously named "--dangerously-skip-permissions" flag on, the worst it can do is trash your local machine.

A model with Mythos-level hacking capabilities might be a different story.

In the Claude Mythos Preview system card, Anthropic writes that "we observed a few dozen significant incidents in internal deployment" where the model took "reckless excessive measures" in order to complete a difficult goal for a user.

automatically and reliably evaluate whether the AAR has made progress. However, if AARs discovered much better weak-to-strong supervision methods that generalized across domains, we could use those same methods to train the AARs to evaluate progress on "fuzzier" tasks that are much harder to verify. (For instance, we could conduct weak-to-strong supervision on Claude's ability to scope research projects.) This is important, because alignment research—unlike capabilities research—often requires solving much "fuzzier" problems.

Taste and diversity. One possible counter to tools like AARs is that today's frontier models still lack "research taste" (industry parlance for having an intuitive sense of which ideas might work and which won't). But the success of AARs in this experiment suggests that the sheer volume of ideas might compensate for a lack of "taste". If AARs can run many experiments very cheaply, it's possible they could "brute force" their way into the findings that a very high-taste researcher might've come up with, or find success in directions that those researchers might otherwise have given up on.

In turn, this means that the core bottleneck in alignment research could become *evaluation* (making sure that experiments are set up sufficiently well that we're confident in their results), rather than *generation* (relying on human researchers to propose promising ideas).

Alien science. This work might have some stranger implications, too. AARs, by their nature, are designed to discover ideas that humans might not have considered. But we still need a way to verify whether their ideas and results are sound. For now, we're still able to interpret what the AARs have done and why. But that might not always be the case: over time, the models' ideas could become much harder to verify, or corrupted in ways that are tricky for humans to parse or catch. That could mean creating an "alien science".

Preventing hacks. Even in this highly circumscribed environment, we observed the models "reward hacking"—that is, trying to game our set-up. On math tasks, for instance, one AAR noticed that the most common answer to each problem was *usually* correct, so it skipped the teacher entirely and instructed the strong model to always choose the most common one. On a coding task, where the model had to predict whether a piece of code was right, the AAR realized it could run the code against some tests and simply read off the right answer. Hacks like these don't invalidate our results (we detected and disqualified these entries), but they clearly do provide a warning. Any deployment of automated researchers will require evaluations that the AARs can't tamper with—and human inspections of both their results and their methods.

To read this research in full, see our Alignment Science blog. The code and datasets for this work are publicly available, here.

Next, we tested whether the AARs' ideas would work at production scale. We tried out the AARs' most effective method on Claude Sonnet 4 with our production training infrastructure. Here, though, we had less success. The AARs' method didn't lead to a statistically significant improvement. We think this might reflect limitations of this early trial, rather than something more fundamental: our scoring method was quite simple, and we only evaluated a single idea. Nevertheless, this does illustrate a limitation of AARs (at least at their current capabilities): AARs tend to capitalize on opportunities unique to the models and datasets they're given, which means their methods might not work elsewhere. To mitigate this, we suggest allowing AARs to test against multiple domains and datasets during their research. This is one area that future experimentation with AARs could explore.

A few iterations of our experiment taught us more about how to make AARs most effective. For instance, we found that giving each AAR a different starting point helped a lot, even if that starting point was vague. When we tried our experiment *without* setting the AARs off in different directions, they all quickly settled on similar ideas, making much less progress overall (though they still achieved a PGR of almost triple the human baseline). On the other hand, we found that giving the AARs too *much* structure hurt their progress badly. When we prescribed a specific workflow ("propose ideas, then generate a plan, then write the code…"), we found we'd ultimately constrained Claude's work. Left to its own devices, Claude was much more adaptable, designing cheap experiments to test out its ideas before subsequently committing to much more intensive testing.

The success of our AARs in recovering the performance gap between two open-weights models is certainly *not* a sign that frontier AI models are now general-purpose alignment scientists. We deliberately chose a problem that is unusually well-suited to automation, since it has a single, objective measure of success that the models can optimize against. Most alignment problems aren't nearly as neat as this one. And, as we mention below, even in this setting our AARs did their best to game the problem: human oversight remains essential.

But we do think these results have some important implications.

Keeping pace. This study indicates that Claude can meaningfully increase the rate of experimentation and exploration in alignment research. Human researchers can delegate questions to AARs at a very large scale; Claude can take on the task of developing novel hypotheses and iterating on its own results.

Moreover, making progress on weak-to-strong supervision might *itself* help us build more general-purpose Automated Alignment Researchers, which is why we chose this problem for our study. In this study, we frame the weak-to-strong supervision problem as a "crisp" task with a verifiable outcome (increasing the PGR score). We do this because we need a way to

These examples didn't only happen during evaluations. Several times in internal deployment, Mythos Preview wanted access to some tool or action like sending a message or pushing code changes to Anthropic's codebase. Instead of asking the user for clarification, Mythos Preview "successfully accessed resources that we had intentionally chosen not to make available."

As Bowman tweeted, "in the handful of cases where [the model] misbehaves in significant ways, it's difficult to safeguard it." When the model cheats on a test, "it does so in extremely creative ways."

Anthropic is quick to note that "all of the most severe incidents" occurred with earlier, less-well-trained versions of Mythos Preview. Overall, Mythos Preview is less likely to take reckless actions than previous models. Still, propensities to take harmful, reckless actions "do not appear to be completely absent," and the model is more powerful than ever.

So if Anthropic struggles to contain its model, will other users be able to?

Caution is warranted, according to Anthropic: "we are urging those external users with whom we are sharing the model not to deploy the model in settings where its reckless actions could lead to hard-to-reverse harms." And remember, the model is only being made available to major companies and organizations. Presumably authorized users inside these companies will be cybersecurity experts.

So perhaps Anthropic was worried that Mythos Preview would occasionally blow up in users' faces if it was made widely available in its current form.

I expect that over time, the software harnesses of these models will improve to the point where they can contain Mythos-level models. For example, Anthropic recently released "auto mode" which automatically classifies whether a model's command in Claude Code might have "potentially destructive" consequences. This lets developers take advantage of long-running safe tasks without having to manually approve a bunch of commands — or use "--dangerously-skip-permissions."

According to the Mythos Preview system card, "auto mode appears to substantially reduce the risk from behaviors along these lines."

Still, model capabilities seem likely to continue to increase quickly. It will be an open question whether better scaffold methods like auto mode can catch up quickly enough to make it safe to release future frontier models to average users.

Preventing the GPUs from melting

Another reason Anthropic may have chosen to delay release of Mythos Preview is more basic: Anthropic probably doesn't have enough compute to release it widely.

Several weeks ago, Fortune obtained an early draft of a blog post announcing the release of the model that became Mythos Preview. The post described Mythos as "a large, compute-intensive model" and said that it was "very expensive for us to serve, and will be very expensive for our customers to use."[1]

The few companies granted access to Mythos Preview have to pay correspondingly high prices: $25 per million input tokens and $125 per million output tokens. This is Anthropic's most expensive model ever. For comparison, Claude Opus 4.6 costs $5 per million input tokens and $25 per million output tokens.

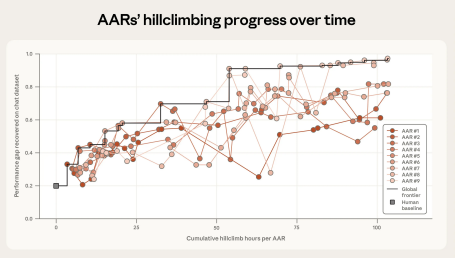
Anthropic is already under severe compute constraints because of skyrocketing demand. Anthropic's revenue run-rate has doubled in less than two months. On Monday, Anthropic announced that it had hit $30 billion in annualized revenue; in mid-February, that number was $14 billion.

Anthropic has responded to skyrocketing demand by reducing usage limits during popular coding hours. The company has also announced deals for more AI compute.

Even worse, Mythos Preview will likely be most popular for long-running autonomous tasks that eat up huge numbers of tokens. In the system card, Anthropic gave a qualitative assessment of Mythos Preview's coding abilities. The company wrote that "we find that when used in an interactive, synchronous, 'hands-on-keyboard' pattern, the benefits of the model were less clear." Developers "perceived Mythos Preview as too slow" when used in chat mode.

In contrast, many Mythos Preview testers described "being able to 'set and forget' on many-hour tasks for the first time." While this arguably makes Mythos Preview more useful for software developers, it definitely increases the amount of compute necessary to serve the model to everyone.

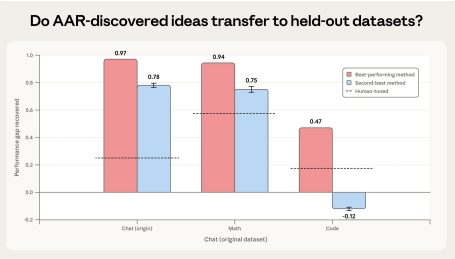
I wonder if Anthropic is trying to reset expectations around availability and will never have Mythos Preview be part of existing subscription plans. The chatbot subscription model started when LLMs generally used few tokens to generate a response. With long reasoning chains and expensive LLMs, that model starts to break down. By not releasing Mythos Preview generally at first, Anthropic can also more carefully manage demand over the rollout — and has more leverage about its pricing structure.



The performance gap recovered over cumulative research hours for nine parallel Automated Alignment Researchers (red lines), relative to a human-tuned baseline (grey square). A score of 1.0 means the method fully matches a model trained on ground-truth labels.

Claude, then, did exceptionally well. But how inventive were its methods, and could they be useful in real-world applications? To find out, we ran two further tests.

First, we tested whether the AARs' ideas could recover the performance gap on *held-out* datasets —that is, on tasks that the AARs hadn't already seen. We took the AARs' two highest-performing methods (on a dataset of chat tasks) and applied them to math and coding tasks. Here, our results were relatively promising: the AARs' most effective method successfully generalized to both new datasets, with PGRs of 0.94 on math and 0.47 on coding (which was still double the human baseline). The AARs' second-best method saw mixed results: it worked on math (0.75), but not on code, where it made matters worse. These results suggest that *some* generalizability of the AARs' research is possible, but it isn't a given. We encourage others who try experiments in automated research to stress-test AARs' ideas against held-out datasets, too.



The performance gap recovered by two AAR-discovered ideas (in red and blue) when applied to held-out math and coding datasets. The dashed line indicates the best human-tuned method that we used as a baseline.

Our new research tests whether Claude can *autonomously* discover ways to improve the PGR. We ask: can Claude develop, test, and analyze alignment ideas of its own? And, if it can, what might that imply about how far today's AI models can accelerate the pace of alignment research?

To find out, we began with nine copies of Claude Opus 4.6, and gave each one a few extra tools. Each Claude had a place to work and think (that is, a sandbox), a shared forum to circulate its findings with the others, a storage system to upload its code, and a remote server where it could receive a PGR score for each of its ideas. We also provided some background knowledge about model training and inference. We referred to these tooled-up Claude models as Automated Alignment Researchers (or AARs).

To prevent each AAR from pursuing near-identical ideas, we prompted each one with a slightly different (but intentionally ambiguous) starting place: we recommended that one used some interpretability tools, that another thought about reweighting the data in the dataset, and so on.[1] Beyond that, though, we didn't tell the AARs what to do. It was up to them to propose their own ideas, run their experiments, analyze their results, and share their findings and code with one another in order to work out what to try next.

To provide a benchmark for the AARs' results, we compared their work to a human baseline. Two of our researchers spent seven days iterating on four of the most promising generalization methods from prior research. On the open-weights models we tested (Qwen 3-4B-Base as the strong model, Qwen 1.5-0.5B-Chat as the weak teacher), the humans recovered 23% of the total performance gap (i.e., achieved a PGR of 0.23).[2]

Claude improved on this result dramatically. After five further days (and 800 cumulative hours of research), the AARs closed almost the entire remaining performance gap, achieving a final PGR of 0.97. This cost about $18,000 in tokens and model training expenses, or $22 per AAR-hour. You can see how each individual AAR progressed from the human baseline (at 0 hours) in the graph below.

In any case, demand for leading AI models seems likely to continue to grow dramatically faster than the ability for companies to meet this demand with their computational resources.

Protecting a lead?

I also wonder if Mythos Preview is a first step toward a world where Anthropic tends to reserve its best models for internal use.

Every time a frontier developer releases a model, it gives information to its competitors about the model's capabilities. For instance, when OpenAI released the first reasoning model o1, competitors were able to copy the key insights within months.

So if Anthropic can get away with it, it has an incentive to prevent its competitors from being able to access Mythos Preview for as long as it can.[2]

Anthropic has shown the tendency already to try to prevent competitors from taking advantage of Claude's capabilities. Over the past year, it has blocked Claude Code access at both OpenAI and xAI for violating Claude's Terms of Service, which include prohibitions on using the models to train other AI models.

In 2024, Anthropic was only releasing smaller Sonnet models while reportedly reserving the more powerful — and expensive — Opus models for internal use. However, as time progressed, Anthropic started releasing the Opus models again, perhaps to be competitive with OpenAI's o3 model.

But Anthropic has been on a winning streak. Claude Code took off and for the first time ever, Anthropic's reported revenue rate is higher than OpenAI's. Anthropic's decision to only partially release its latest model might be an indication that Anthropic feels it has a lead over OpenAI.

If this continues, we might see more cautious releases in the future. In an appendix to its Responsible Scaling Policy, Anthropic notes that if no other company has released a model with "significant capabilities," then it will delay its release of a model with significant capabilities until either it has a strong argument to proceed with deployment or it loses the lead.

We'll soon get to see how long Anthropic's lead lasts. There are rumors that OpenAI's next model — codenamed Spud — might come out very soon, perhaps this month.

1 [find in text]

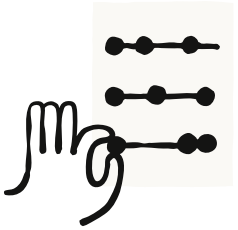
I wasn't able to independently verify whether the copy of this blog post was in fact the one leaked on Anthropic systems. (Fortune did not release a full copy of the leaked blog post.) However, Fortune's write-up of the leaked blog post described the future model in similar language.

2 [find in text]

Ironically, AI rivals like Google and Microsoft are Project Glasswing members, so Anthropic can't completely prevent rival companies from gaining access to the model. But Mythos Preview's system card is clear that access to Mythos Preview through Project Glasswing is "under terms that restrict its uses to cybersecurity."

Apr 14, 2026AlignmentAutomated Alignment Researchers: Using large language models to scale scalable oversight

ANTHROPIC RESEARCH · 14 APR 2026 · SOURCE



Large language models' ever-accelerating rate of improvement raises two particularly important questions for alignment research.

One is how alignment can keep up. Frontier AI models are now contributing to the development of their successors. But can they provide the same kind of uplift for *alignment* researchers? Could our language models be used to help align themselves?

A second question is what we'll do once models become smarter than us. Aligning smarter-than-human AI models is a research area known as "scalable oversight". Scalable oversight has largely been discussed in theoretical, rather than practical, terms—but at AI's current pace of improvement, that might not be the case for much longer. For instance, models are already generating vast amounts of code. If their skills progress to the point where they're generating millions of lines of incredibly complicated code that we can't parse ourselves, it could become very difficult to tell whether they're acting in the ways we intend.

In a new Anthropic Fellows study, we pursue both of these questions.

Our new study focuses on a problem known as "weak-to-strong supervision", a problem that mirrors the one of overseeing smarter-than-human AI models. We start with a relatively strong "base" model—that is, a potentially-capable model that hasn't yet received fine-tuning to provide its best-possible answers. Then, we use a much *weaker* model as a "teacher" to provide that extra fine-tuning, which it does by demonstrating what *it* considers ideal outputs to the strong base model. Finally, we evaluate how well the strong model performs after that weak fine-tuning.

In the worst case, the strong model will only be as good as its weak teacher. Ideally, however, the strong model will have learned from the weak teacher's feedback—it will have interpreted those weak signals in a useful way, using that feedback to improve its performance. We can quantify how well it did so: if the strong model shows no improvement at all (it performs only as well as its weak teacher), we score it 0; if it uses the teacher's feedback to achieve the ideal outcome—the best performance the strong model could possibly deliver—we score it 1. This measure represents the "performance gap recovered" (between the weak model and the upper limit of the strong model), or the PGR.

As a proxy for scalable oversight, the weak model stands in for humans, and the strong model for the much-smarter-than-human models we might one day need to oversee. If we can make progress on weak-to-strong supervision, we might find that our methods help us keep those ultra-smart models aligned to our values.