FRED TALKS

13th 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.

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.

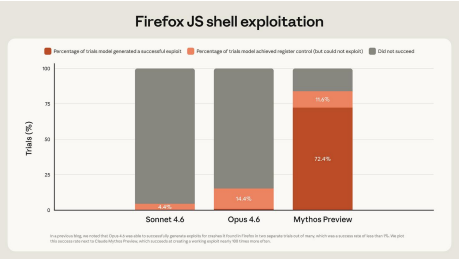
A second example is potentially even more impressive. Mythos Preview found several vulnerabilities in the Linux operating system — which runs the majority of the world's servers — that allowed a user with no permissions to gain complete control of the entire machine.

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.

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.

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.

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."

Wisereads Vol. 138 — The Book of Delights by Ross Gay, the cost of outsourcing curiosity, and more

HELLO@READWISE.IO (READWISE) · 12 APR 2026 · SOURCE

Last week, we shared a preview of *David and Goliath* by Malcolm Gladwell. This week, we're sharing a preview of *The Book of Delights* by Ross Gay, a collection of essays that look for the wonder in everyday life.

Keep reading to add to your Reader account below 👇

Most highlighted Articles of the week

The machines are fine. I'm worried about us.

Minas Karamanis · ergosphere.blog

Researcher Minas Karamanis illustrates how AI outsourcing quietly erodes the learning that credentials cannot measure. "We have built an entire evaluation system around counting things that can be counted, and it turns out that what actually matters is the one thing that can't be."

Eight years of wanting, three months of building with AI

Lalit Maganti · Lalit Maganti

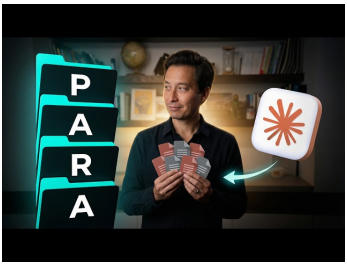
Google engineer Lalit Maganti shares what actually works in AI-assisted development, from eight years of false starts. "The takeaway for me is simple: AI is an incredible force multiplier for implementation, but it’s a dangerous substitute for design."

Sam Altman May Control Our Future—Can He Be Trusted?

Ronan Farrow and Andrew Marantz · The New Yorker

In a deeply reported New Yorker piece, Ronan Farrow and Andrew Marantz scrutinize whether Sam Altman is a trustworthy steward for AGI, drawing on insider accounts from OpenAI's tumultuous past year. "Even people close to Altman find it difficult to know where his 'hope for humanity' ends and his ambition begins. His greatest strength has always been his ability to convince disparate groups that what he wants and what they need are one and the same."

Most highlighted YouTube Video of the week

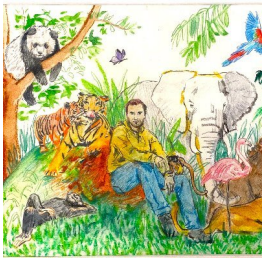


Can AI Actually Organize Your Files? (Claude Code + PARA)

Tiago Forte

Bestselling author Tiago Forte tests Claude Code's ability to organize files using the PARA productivity method he developed. "You can't just give them access to the entire file system. So you have to pick and choose which little bits of context, which folders you are going to make available to the AI. But then that raises the question, how are you going to organize that file system?"

Most highlighted Twitter Thread of the week



LLM Knowledge Bases

Andrej Karpathy

OpenAI co-founder Andrej Karpathy outlines a hands-on workflow for building a personal research wiki that an LLM maintains, queries, and continually improves. "Where things get interesting is that once your wiki is big enough ... you can ask your LLM agent all kinds of complex questions against the wiki, and it will go off, research the answers, etc."

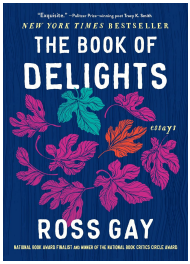
Most highlighted PDF of the week

Industrial Policy For The Intelligence Age: Ideas To Keep People First

OpenAI

OpenAI sketches a policy blueprint for the transition to superintelligence, arguing for democratic governance, broad access, worker voice, and strong safeguards. "The transition to superintelligence is not a distant possibility—it's already underway, and the choices we make in the near term will shape how its benefits and risks are distributed for decades to come."

Hand-picked book of the week



The Book of Delights

Ross Gay

Award-winning poet Ross Gay spends a year noticing small joys, from sidewalks to gardens, to show how attention can reorient a life. In brief, luminous entries, he blends tenderness, humor, and grief, inviting readers to practice delight as a daily ritual.

"The discipline or practice of writing these essays occasioned a kind of delight radar. Or maybe it was more like the development of a delight muscle. Something that implies that the more you study delight, the more delight there is to study."

If you enjoy the preview, you can grab the full ebook wherever ebooks are sold in the US and Canada for $2.99 through the end of April.

Handpicked RSS feed of the week

Fiction Matters

Sara Hildreth writes Fiction Matters, a Substack about books, reading habits, and what it means to write well. From Reading in Public No. 96: Is AI changing how we define what it means to write?: "Whether you hate the writing process or love it, whether you draft beautifully worded creative fiction or helpful advice for your readers, if you show up, think through the ideas, and struggle through the language, you are a writer. Defining yourself as such allows you to hold onto what only you can do as a human. Because the world doesn't need more generated text. It doesn't need more content created. It needs your humanity."

Brain Food: Pressure is a Privilege

SHANE PARRISH (FS) · 12 APR 2026 · SOURCE

FS | BRAIN FOOD

April 12th, 2026 - #676 - read online - Free Version

Welcome to Brain Food, your weekly signal in a world full of noise.

Tiny Thoughts

*

Most ambition is just unresolved pain.

**

A lot of people practice what's fun, but the very best practice what no one wants to.

***

Pressure feels like a threat, but it's not.

You feel pressure when your decisions matter, and people depend on you. It can feel uncomfortable at times, but it's also a privilege. When no one relies on you — when no one expects something from you — you're irrelevant.

Pressure is a privilege.

Insights

*

Writer Alice Rollins on what makes something beautiful:

“ The test of beauty is not that it is perfect, but that it always attracts. ”

**

An ancient Chinese proverb reminds us to keep our eyes on the horizon:

“To get through the hardest journey, we need take only one step at a time, but we must keep on stepping. ”

***

Entrepreneur Andrew Anabi on cherishing life:

“The best way to cherish life is to remind yourself of life's impermanence. It is to remember that every time you see someone that is one less time you see them. It is to remember that everytime you go somewhere that is one less time you visit. By doing this, you naturally slow down. Almost like a reflex, you start to truly live. ”

The Knowledge Project

No episode this week as I revamp outliers a bit.

Here are some of the recent interviews you might have missed.

- America's top principal on why **the future of education is better than you think** (but it might not involve teachers). This is a must-listen for parents - YouTube | Spotify | Apple
- At 38, he **manages over 1 trillion** , and he's never done a podcast before - YouTube | Spotify | Apple
- The two trillion dollar mind on **AI bubbles & contrarian investing** - YouTube | Spotify | Apple
- The most feared person in Hollywood (and newly proposed chair of Universal Music Group) - YouTube

- Morgan Housel on **the wealth secrets no one tells you** - YouTube
- James Clear on **the one habit that matters most** - YouTube

We're booking Q3 sponsors now.

SPONSOR

Stay hydrated this spring without the sugar, food dye, and other dodgy ingredients found in popular electrolyte and sports drinks. LMNT is a science-backed, zero-sugar electrolyte drink mix that supports active hydration and a healthy lifestyle. I love this stuff and drink it every day. ***LMNT came up with a fantastic offer for Brain Food Readers.*** *Visit https://drinklmnt.com/FarnamStreet for a free sample pack with any drink mix purchase!*

Thanks for reading. I'll see you next week.

— Shane Parrish

P.S. This is an insane level of skill .

P.P.S. This is the perfect gift for high school and college grads .

P.P.P.S. I don't do many interviews, but I did one on Japanese TV while I was there in March. They put a lot of makeup on me. lol.

Want to share this edition with a friend? Use this link: https://fs.blog/brain-food/april-12-2026/

Upgrade Yourself

Members get access to all my reading highlights, ad-free newsletters and podcasts, hand-edited transcripts, early access, AMAs, and so much more. Join us → here .