

Forum:[Hack Forums](#)

Thread: [Anthropic Mythos Leak Shocks AI Industry, Too Powerful Claims Raise Concerns](#)

JoseMar - April 1, 2026, 5:08 pm

Anthropic Mythos Leak Shocks AI Industry,
"Too Powerful" Claims Raise Concerns

Hi HF Users, Back Here Again With Some Crazy News About A New AI Leak.
This One Feels Different, Not Gonna Lie.

Anthropic Apparently Has This New Model Called Mythos,
Also Nicknamed Capybara For Some Reason, And It Got Leaked.
Not Officially Announced, Just Sitting In Like 3,000 Documents That Shouldn't Have Been Public.

And The Weird Part Isn't Even The Leak.
Leaks Happen. Whatever.?

It's What They Wrote In Those Docs.

This One Sounds Different

They're Calling It "Bigger And More Intelligent" Than Their Previous Best Models.
Yeah Yeah, We Hear That Every Time.

But Then They Go And Say It's Basically On Another Level.
Not Incremental. Not "Opus 4.6 But Better".?

And I'm Sitting Here Thinking... Didn't They Say The Same Thing Last Time?

Maybe. But Then You Read The Cybersecurity Part.

That's Where It Gets Weird.

Even They Seem Nervous

They Literally Warn That This Model Could Be Ahead Of Defenders In Cybersecurity.

Like... What?

Companies Don't Usually Say Stuff Like That Out Loud.
At Least Not This Direct.

They're Basically Saying, "Hey, This Might Help People Find And Exploit Vulnerabilities Faster
Than We Can Fix Them."

And I'm Thinking... Why Admit That?

Is It Transparency? Or Just Marketing Wrapped In Fear?

Because Let's Be Real, Saying Your Product Is So Good It's Dangerous... That's Also A Flex.

Meanwhile OpenAI Is Cooking Something

Of Course They Are.?

Codename "Spud". Yeah, Potato. No Idea Why.

Not Much Info Yet, Just That Pretraining Is Done.
But Timing Is Not Random.

Anthropic Moves, OpenAI Moves.
Back And Forth Like Always.

Feels Less Like Innovation And More Like A Race Where Nobody Can Slow Down Even If They Want To.

The Hype Machine Never Stops

This Is The Part That Feels Familiar.

Next Model Will Be Better.
Next One Will Change Everything.
Next One Will Be Smarter, Faster, Stronger.

Same Script We've Seen With Phones, GPUs, Everything.

And Honestly... It Works.

Because Six Months Ago We Were Impressed By Stuff That Now Feels Slow.
That's Kind Of Crazy When You Think About It.

We Adapt Way Too Fast.

But There's A Catch Nobody Talks About

These Models Are Getting Expensive.???
Like Really Expensive.

People On High Tier Plans Are Already Hitting Limits Fast.
Tokens Burning Like Crazy.

So What Happens Next??

Better AI... But Less Access?
Smarter Models... But Only For Companies That Can Afford Them?

Feels Like That's Where This Is Going.

I Don't Know, But This One Doesn't Feel Like Just Another Update.
Feels Like Something Is Shifting A Bit Faster Than Before.

Hackers - April 1, 2026, 5:11 pm

I'm okay with the leak overall because from what I hear, it's really just frontend mappings and none of their actual model, so the main IP is completely protected from this issue and nobody can actually replace Claude with what was leaked.

This also could be a publicity stunt w/ the notes you mentioned being placed there to draw up hype... if that was the plan, it worked.

AI is definitely going to go up in price & be more restricted. That's capitalism and many people aren't prepared. AI is set to replace anywhere from lets conservatively say 50-70% of the digital tasks people are usually hired for, and most likely more like 85-95% of current tasks. People might still be employed, but the TASKS they do will be made redundant and they'll be assigned new work, and others simply won't be employed.

DramaKing - April 1, 2026, 6:38 pm

Then you'll be interested to hear about Gemini 3.1 Flash. I read that it's significantly cheaper than Gemini 2.5 Flash.

Lyrlei - April 1, 2026, 8:58 pm

Maybe the real AGI was the AI slop we made along the way.

Kaleb - April 12, 2026, 7:33 am

This is actually fairly interesting if you read about it, surprised more people did not respond to this or cover it, its honestly pretty huge news as far as I know.

From everything gathered it seems its just a general purpose LLM that has SWE capabilities and just performs well on the benchmark. In spite of this it managed to find >1,000 0-day vulnerabilities despite not even being created strictly to do so (although I'd argue that's why its probably better at it than an agent that's solely focused on SWE).

But this really does bring together an interesting situation. We basically have a gatekept model that can apparently just print out 0 days like its candy and its only in the hands of whoever the fuck Anthropic decides. If leaks get out, think WannaCry 2016 style things could be really, really, really bad and have real life consequences very fast. Do not forget people quite literally died because of the WannaCry attack in hospitals because they quite literally could not access their computer systems.

I'm really curious if its just the case that this thing went balls to the wall and scanned a bunch of obscure code and architecture types and software and shit and just found vulnerabilities that realistically humans could've or would've found eventually if they tried, but there was just nobody even considering tinkering there. They said it actually found the zero day while having it clean code I believe, maybe they were iterating over an entire operating system or something insane and it managed to find 0 days throughout that? I have no idea but its very interesting. I think the biggest claim is that they created a 0 day for FreeBSD that allowed for RCE in under like a day? It makes cyber security quite interesting.

(Apr 1, 2026 05:11 PM)Hackers Wrote: [\[CITE\]](#) AI is definitely going to go up in price & be more

restricted. That's capitalism and many people aren't prepared. AI is set to replace anywhere from lets conservatively say 50-70% of the digital tasks people are usually hired for, and most likely more like 85-95% of current tasks. People might still be employed, but the TASKS they do will be made redundant and they'll be assigned new work, and others simply won't be employed.

The biggest problem is Mythos is already out to the vendors that have access to it but the exclusivity is to that of them, and I guess whoever works for them. I'm sure most of the people with access to the model have a LOT to lose but the fuel could always be there or it could be tempting to abuse one of these models. If you give people the ability, they will, look at what was happening in the NSA with PRISM and probably still.

MaîtreCornichon - June 8, 2026, 8:44 am

Sounds like the AI race just hit a new gear; this could shake things up a lot.