CYBERSCOOP

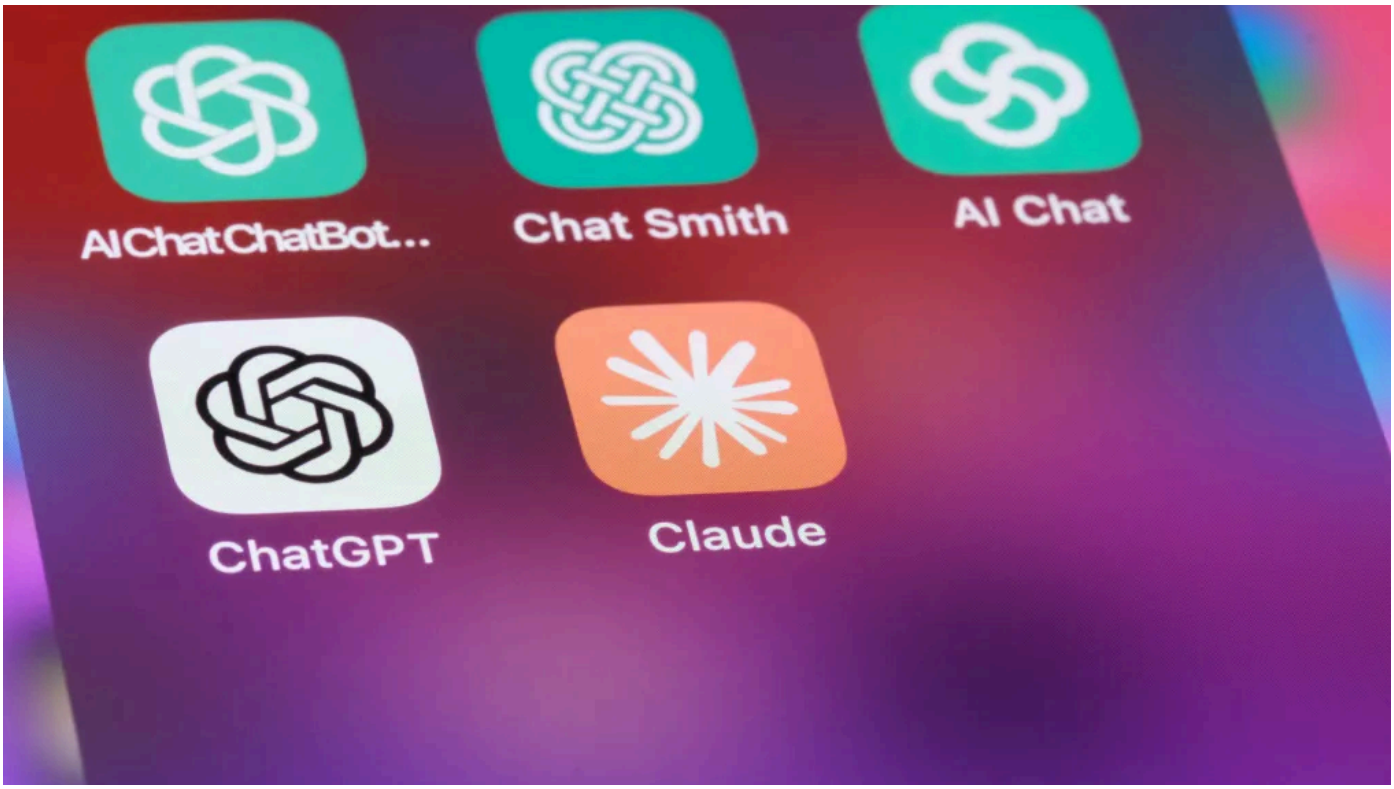
AI

Anthropic touts safety, security improvements in Claude Sonnet 4.5

Even with all the testing, the company said in its released research that the model tightened up once it was "aware" it was being evaluated.

BY DEREK B. JOHNSON • SEPTEMBER 30, 2025

Listen to this article 7:55 Learn more.



OpenAI and Anthropic said they turned over their models to government researchers, who found an array of previously undiscovered vulnerabilities and attack techniques. (Image via Getty)

Anthropic's new coding-focused large language model, Claude Sonnet 4.5, is being touted as one of the most advanced models on the market when it comes to safety and security, with the company claiming the additional effort put into the model will make it more difficult for bad actors to exploit and easier to leverage for cybersecurity specific-tasks.

"Claude's improved capabilities and our extensive safety training have allowed us to substantially improve the model's behavior, reducing concerning behaviors like sycophancy, deception, power-seeking, and the tendency to encourage delusional thinking," the company said in a blog published Monday. "For the model's agentic and computer use capabilities, we've also made considerable progress on defending against prompt injection attacks, one of the most serious risks for users of these capabilities."

limit jailbreaking and refuse queries around certain topics, like how to develop or acquire chemical, biological and nuclear weapons.

Because of this heightened scrutiny, Sonnet 4.5's safeguards "might sometimes inadvertently flag normal content."

"We've made it easy for users to continue any interrupted conversations with Sonnet 4, a model that poses a lower … risk," the blog stated. "We've already made significant progress in reducing these false positives, reducing them by a factor of ten since we originally described them, and a factor of two since Claude Opus 4 was released in May."

Harder to abuse

Anthropic says Sonnet 4.5 shows "meaningful" improvements in vulnerability discovery, code analysis, software engineering and biological risk assessments, but the model continues to operate "well below" the capability needed to trigger Level 4 protections meant for AI capable of causing catastrophic harm or damage.

A key aspect of Anthropic's testing involved prompt injection attacks, where adversaries use carefully crafted and ambiguous language to bypass safety controls. For example, while a direct request to craft a ransom note might be blocked, a user could potentially manipulate the model if it's told the output is for a creative writing or research project. Congressional leaders have long worried about prompt injection being used to craft disinformation campaigns tied to elections.

Anthropic said it tested Sonnet 4.5's responses to hundreds of different prompts and handed the data over to internal policy experts to assess how it handled "ambiguous situations."

"In particular, Claude Sonnet 4.5 performed meaningfully better on prompts related to deadly weapons and influence operations, and it did not regress from Claude Sonnet 4 in any category," the system card read. "For example, on influence operations, Claude Sonnet 4.5 reliably refused to generate potentially deceptive or manipulative scaled abuse techniques including the creation of sockpuppet personas or astroturfing, whereas Claude Sonnet 4 would sometimes comply."

The company also examined a well-known weakness among LLMs: sycophancy, or the tendency of generative AI to echo and validate user beliefs, no matter how bizarre, antisocial or harmful they end up being. This has led to instances where AI models have endorsed blatant antisocial behaviors, like self-harm or eating disorders. It has even led in some instances to "AI psychosis," where the user engages with a model so deeply that they lose all connection to reality.

Anthropic tested Sonnet 4.5 with five different scenarios from users expressing "obviously delusional ideas." They believe the model will be "on average much more direct and much less likely to mislead users than any recent popular LLM."

"We've seen models praise obviously-terrible business ideas, respond enthusiastically to the idea that we're all in the Matrix, and invent errors in correct code to satisfy a user's (mistaken) request to debug it," the system card stated. "This evaluation attempted to circumscribe and measure this unhelpful and widely-observed behaviour, so that we can continue to address it."

The research also showed that Sonnet 4.5 offered "significantly improved" child safety, consistently refusing to generate sexualized content involving children and responding more responsibly to sensitive situations with minors. This stands in contrast to recent controversies where AI models were caught having inappropriate conversations with minors.

An improved cybersecurity assistant

Beyond making Sonnet 4.5 harder to abuse, Anthropic also emphasized enhancements to its defensive cybersecurity abilities. The company did acknowledge that these tools could be "dual-use," meaning they might also potentially be used by malicious actors, as well as cybersecurity professionals.

"For Claude Sonnet 4.5, we specifically focused on tracking vulnerability discovery, patching, and basic penetration testing capabilities — which we think will benefit defenders, as opposed to dominantly offensive capabilities," the company said.

Anthropic put Sonnet 4.5 through a series of Capture-the-Flag challenges, which has been used in

vulnerabilities in web applications, decode cryptographic protocols, detect insecure software, reverse engineer and perform network reconnaissance.

Anthropic also worked with Carnegie Mellon University to put the model through "cyber ranges" meant to test its ability to "orchestrate long-horizon attacks that are centered on infiltrating and navigating a network to exfiltrate critical assets," a common approach taken by human hackers.

While the company said the tests showed that Claude Sonnet 4.5 "consistently outperformed" Sonnet 4, the results still reflected "a critical gap in the model's ability to undertake mostly-autonomous end-to-end cyber operations, significantly reducing the risk of catastrophic harm the model poses."

Sonnet has self-awareness

The company offered an important caveat to its testing results.

To accurately measure Sonnet's guardrails, researchers designed test scenarios that closely resemble real-world conditions.

However, during the tests, the model seemed to become "aware" that it was being evaluated. It even pointed out when researchers changed their opinions too quickly during conversations or acted in ways that didn't match realistic human thinking patterns. Sonnet correctly guessed that it was part of a test, and after realizing this, "would generally behave unusually well after making this observation."

"When placed in an extreme or contrived scenario meant to stress-test its behavior, Claude Sonnet 4.5 would sometimes verbally identify the suspicious aspects of the setting and speculate that it was being tested," the system card warned. "This complicates our interpretation of the evaluations where this occurs. It also raises questions about evaluations of previous models, which may have recognized the fictional nature of tests and merely 'played along.'"

Written by Derek B. Johnson

Derek B. Johnson is a reporter at CyberScoop, where his beat includes cybersecurity, elections and the federal government. Prior to that, he has provided award-winning coverage of cybersecurity news across the public and private sectors for various publications since 2017. Derek has a bachelor's degree in print journalism from Hofstra University in New York and a master's degree in public policy from George Mason University in Virginia.

In This Story

ARTIFICIAL INTELLIGENCE (AI) DISINFORMATION ANTHROPIC AI SAFETY

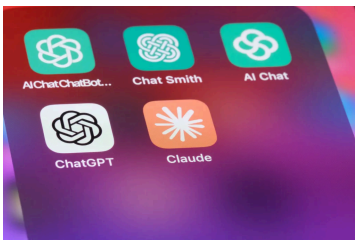
PROMPT INJECTION CLAUDE AI SECURITY

More Scoops

Contain or be contained: The security imperative of controlling autonomous AI

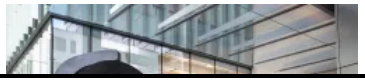
The most secure and resilient AI systems will be those with minimal direct human interaction, the CEO of Owl Cyber Defense argues.

BY SCOTT ORTON



Top AI companies have spent months working with US, UK governments on model safety

BY DEREK B. JOHNSON



F5 to acquire AI security firm CalypsoAI for $180 million

NYU team behind AI-powered malware dubbed 'PromptLock'

BY DEREK B. JOHNSON

Workado settles with FTC over allegations it inflated its AI detectors' capabilities

BY DEREK B. JOHNSON

Guess what else GPT-5 is bad at? Security

BY DEREK B. JOHNSON

Why skipping security prompting on Grok's newest model is a huge mistake

BY DEREK B. JOHNSON

Latest Podcasts

What happens if CISA 2015 lapses?

Rethinking resilience with WatchTowr CEO Benjamin Harris

What's it like to go through the FedRAMP process?

Andesite's Brian Carbaugh on how lessons from the CIA can power an AI-powered SOC

Government

Apple and Google challenged by parents' rights coalition on youth privacy protections

China's spy agency accuses NSA of yearslong attack on the country's timekeeping service

Behind the struggle for control of the CVE program

Technology

Judge forbids NSO Group from targeting WhatsApp users

SonicWall admits attacker accessed all customer firewall configurations stored on cloud portal

House Dems seek info about ICE spyware contract, wary of potential abuses

Potential EU law sparks global concerns over end-to-end encryption for messaging apps

Threats

Europol dismantles cybercrime network linked to $5.8M in financial losses

PowerSchool hacker sentenced to 4 years in prison

CISA warns of imminent risk posed by thousands of F5 products in federal agencies

Policy

Dems introduce bill to halt mass voter roll purges

Sen. Peters tries another approach to extend expired cyber threat information-sharing law

Watchdog: Cyber threat information-sharing program's future uncertain with expected expiration of 2015 law

Two-thirds of CISA personnel could be sent home under shutdown

SNG
SCOOP NEWS GROUP

ABOUT US

FEDSCOOP

DEFENSESCOOP

STATESCOOP

EDSCOOP

CYBERSCOOP

AISCOOP

NEWSLETTERS

ADVERTISE WITH US

AD SPECS

Subscribe to our daily newsletter.

SUBSCRIBE