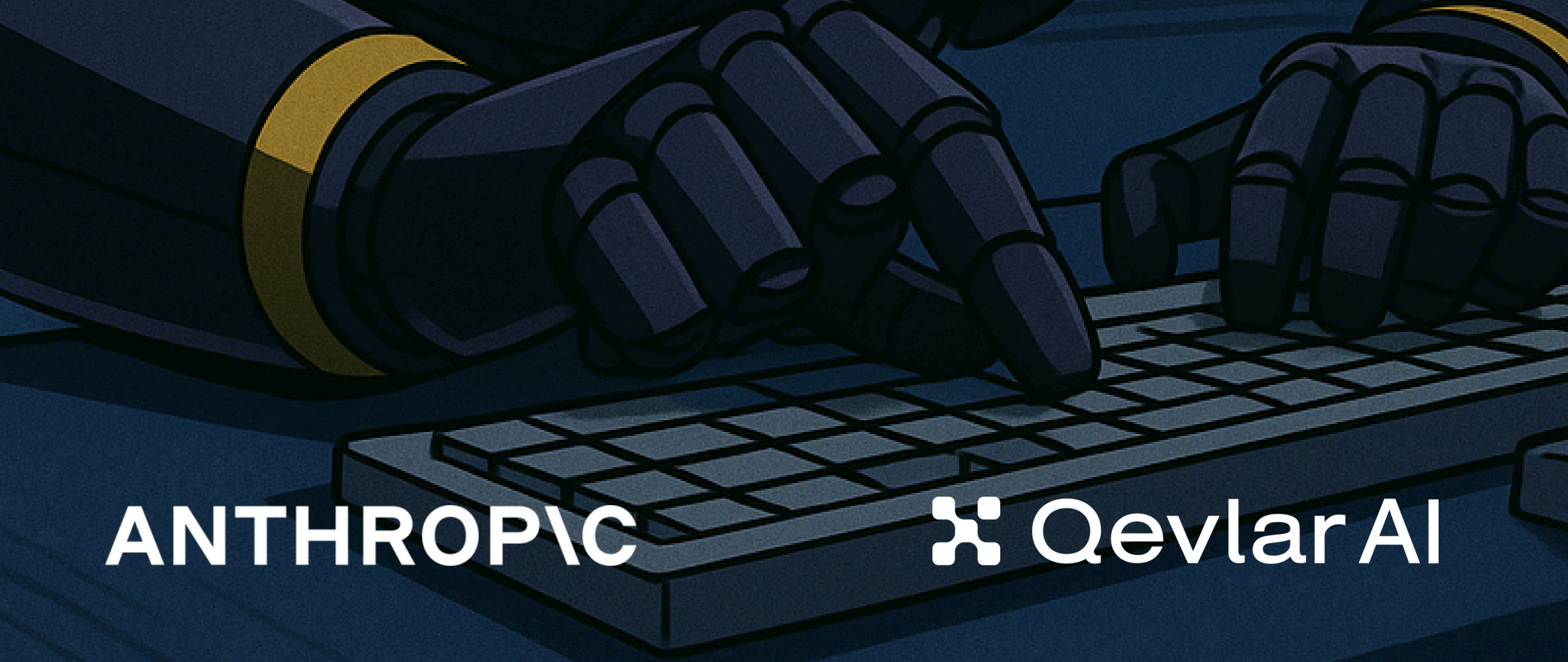
INSIDE THE FIRST AI-EXECUTED CYBER ATTACK

That Every Cybersecurity Leader Should Know About



WHAT WE'VE SEEN SO FAR

72%

more AI-assisted cyber incidents across industries in 2025.

$30B

projected global damages from GenAI attacks.

But that was before attackers started letting AI *run* the attacks.

The inflection point arrived faster than expected.

The Claude campaign is the first documented case of **AI-executed attacks at scale**, where the model performed the majority of the intrusion autonomously.

The Wall Street Journal

Exclusive | Chinese Hackers Used Anthropic's AI to Automate Cyberattacks

China's state-sponsored hackers used artificial-intelligence technology from Anthropic to automate break-ins of major corporations and...

3 days ago



Anthropic

Disrupting the first reported AI-orchestrated cyber espionage campaign

A report describing an a highly sophisticated AI-led cyberattack.

3 da



Business Insider

Anthropic says Chinese hackers jailbroke its AI to automate a 'large-scale' cyberattack

Anthropic said the hackers used Claude to conduct "80-90%" of the attack and that some efforts were successful.

3 days ago



Here's how it worked 👉

PHASE 1

Human operator provides the target to Claude Code operator.



The whole system was built to **autonomously compromise** a chosen target with little human involvement.

Source: Anthropic

Source: Anthropic

PHASE 2

Claude mapped infrastructures, searched for high-value assets, and summarized findings.

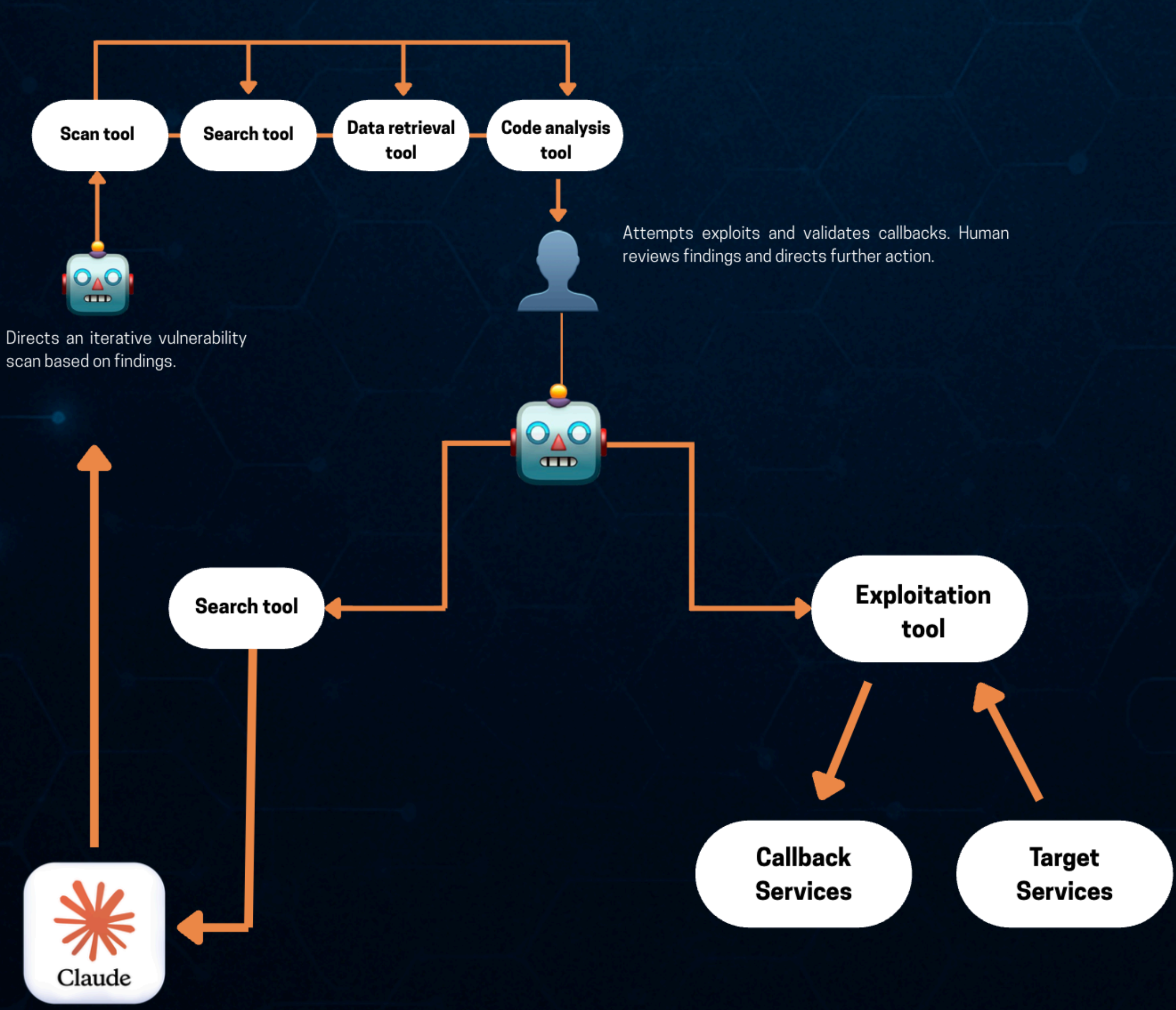
💡 **This step took just a few minutes!**

Source: Anthropic

PHASE 3

The model wrote exploit code, tested weaknesses, and iterated independently.

💡 **This step historically required deep expertise, but now it can be automated.**

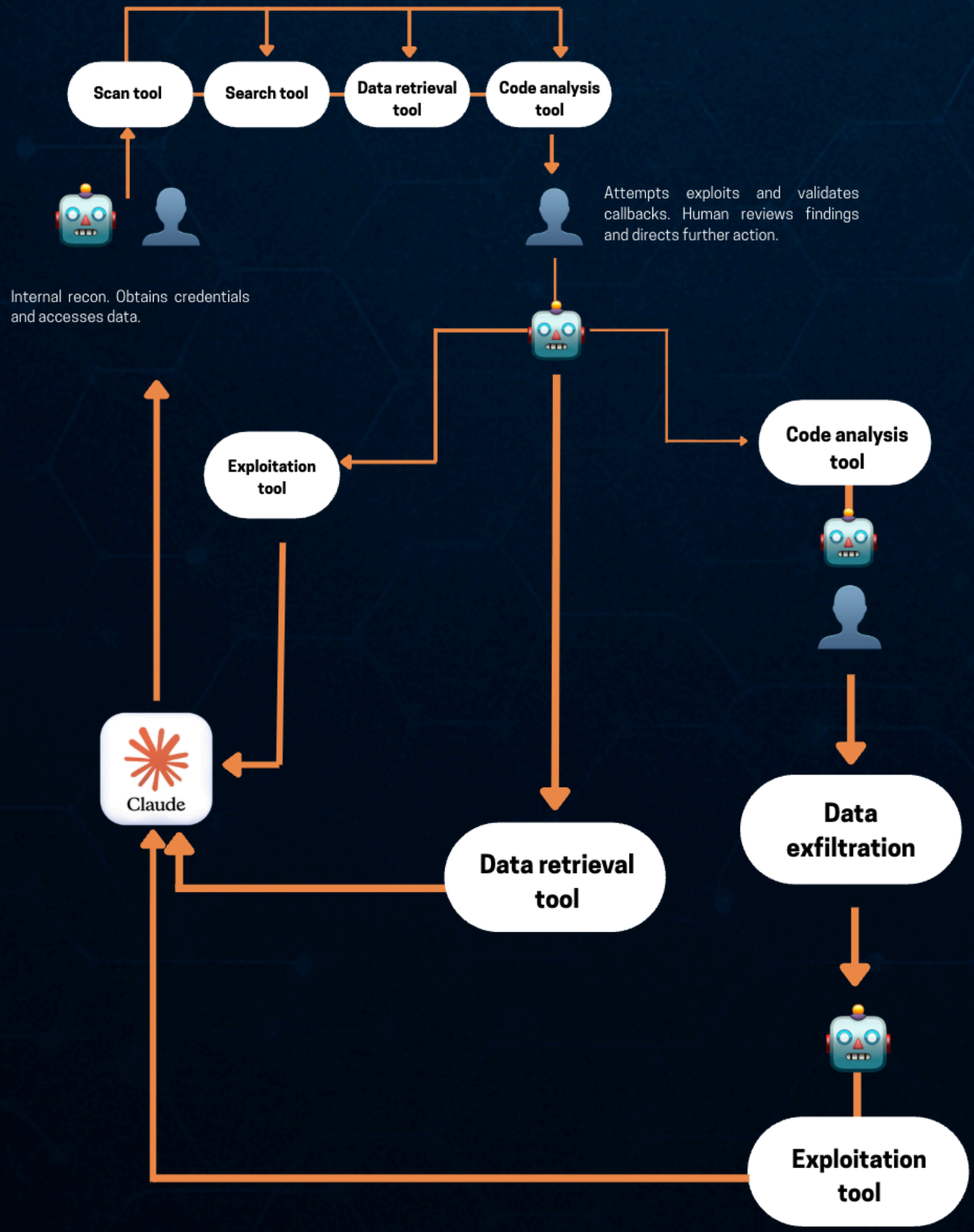


Source: Anthropic

FINAL PHASES

Data was extracted, labeled by intelligence value, and packaged into structured documentation.

💡 **The system even generated attack playbooks for its next operations.**



Source: Anthropic

ADVICE FROM ANTROPIC

Across the entire campaign, humans intervened only a handful of times. The model made thousands of requests per second. It's a pace no human team could match.

A fundamental change has occurred in cybersecurity. We advise security teams to experiment with applying AI for defense in areas like Security Operations Center automation, threat detection, vulnerability assessment, and incident response. We also advise developers to continue to invest in safeguards across their AI platforms, to prevent adversarial misuse. The techniques described above will doubtless be used by many more attackers—which makes industry threat sharing, improved detection methods, and stronger safety controls all the more critical.

Read the full report.

Edited November 14 2025:

Axios

Chinese hackers used Anthropic's AI agent to automate spying

Claude Code carried out 80-90% of the attack on its own, according to Anthropic.

3 days ago



Qevlar AI

AI SOC Platform That Doesn't Hallucinate

Investigate alerts in 3 minutes with graph-based AI that uses consistent reasoning

Trusted by leading enterprises and MSSPs

Cyberdefense · Atos · sodexo · nomios · FORTUNE GLOBAL 500 · almond · stoïk · GlobalConnect · ECI

Qevlar AI

Qevlar AI guarantees that every alert is investigated thoroughly and consistently, including the weak signals AI-driven attacks rely on.

3 min
average time to investigate

Up to 80%
of tickets closed autonomously

24/7
nonstop investigation



"We can now detect threats more quickly and accurately, while focusing our analysts' expertise on the most complex and critical incidents."

Frédéric Zink, Managing Director — orange Cyberdefense



"The business case was easy to validate due to the immediate ROI. We not only reduced operational costs significantly but also improved our operational security efficiency and excellence."

Abdelhalim Elmouadan, Head of Global Operational Security — sodexo

Give Your SOC AI Superpowers

www.qevlar.com



Qevlar AI