ANTHROP\C

The Permanent Defender Advantage has a Single Point of Failure

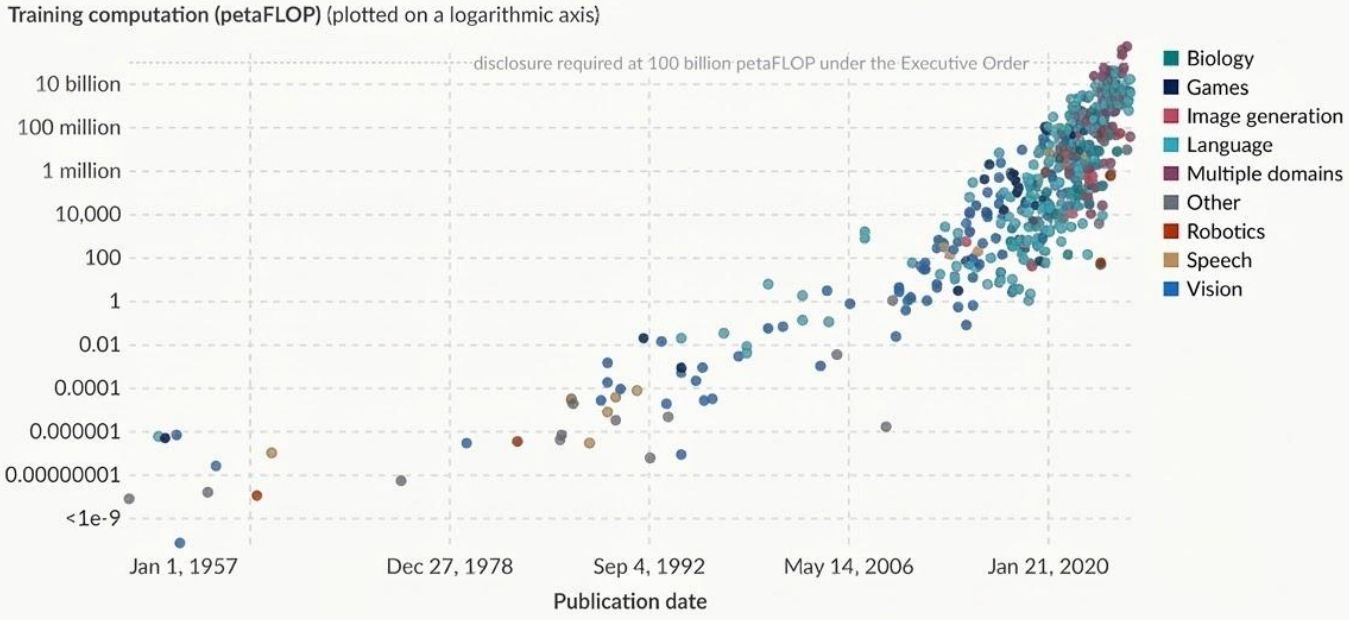
AI, Attackers, and the Confidential Computing

Jason Clinton
Deputy CISO

Computation used to train notable artificial intelligence systems, by domain

Our World in Data

Computation is measured in total petaFLOP, which is 10^{15} floating-point operations[1]. Estimated from AI literature, albeit with some uncertainty. Estimates are expected to be accurate within a factor of 2, or a factor of 5 for recent undisclosed models like GPT-4.



Data source: Epoch AI (2025)

OurWorldinData.org/artificial-intelligence | CC BY

Note: The Executive Order on AI refers to a directive issued by President Biden on October 30, 2023, aimed at establishing guidelines and standards for the responsible development and use of artificial intelligence within the United States.

1. **Floating-point operation** A floating-point operation (FLOP) is a type of computer operation. One FLOP represents a single arithmetic operation involving floating-point numbers, such as addition, subtraction, multiplication, or division.

A view from where we sit

The agentic era is here. For attackers too.

ANTHROP\C

BY MATTHEW GAULT SECURITY JUN 4, 2025 6:00 AM

The Rise of 'Vibe Hacking' Is the Next AI Nightmare

In the very near future, victory will belong to the savvy blackhat hacker who uses AI to generate code at scale.

SEPTEMBER 15, 2025

Vibe hacking: How AI-driven cybercrime outpaces EDR and signature defenses

EXCLUSIVE ARTIFICIAL INTELLIGENCE Follow

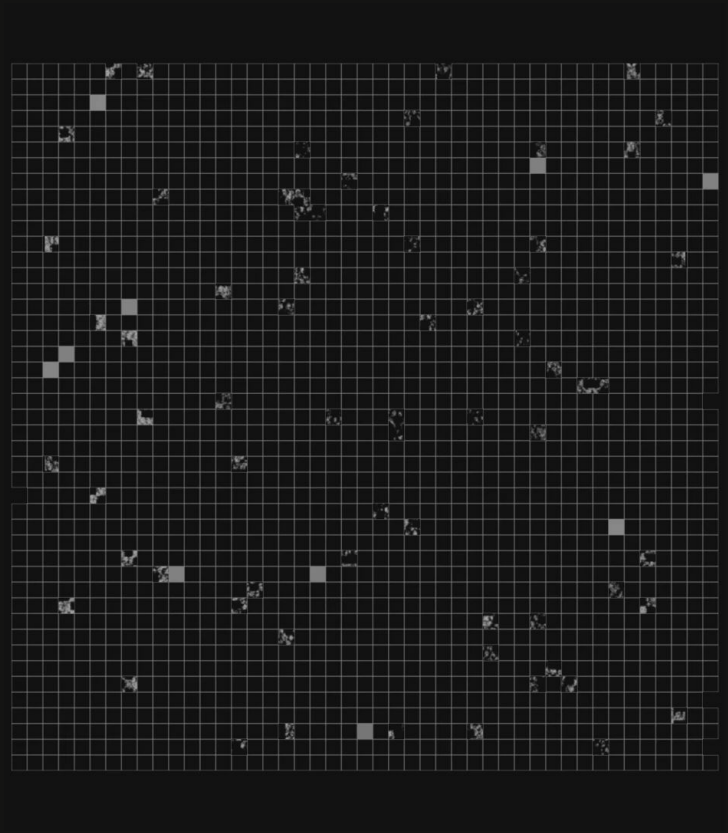
Chinese Hackers Used Anthropic's AI to Automate Cyberattacks

The use of AI automation in hacks is a growing trend that gives hackers additional scale and speed

ANTHROP\C

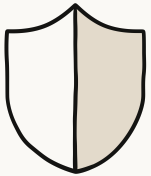
Project Glasswing

AI now finds what humans missed for decades



AI models have reached a level of coding capability where they can surpass all but the most skilled humans at finding and exploiting software vulnerabilities.

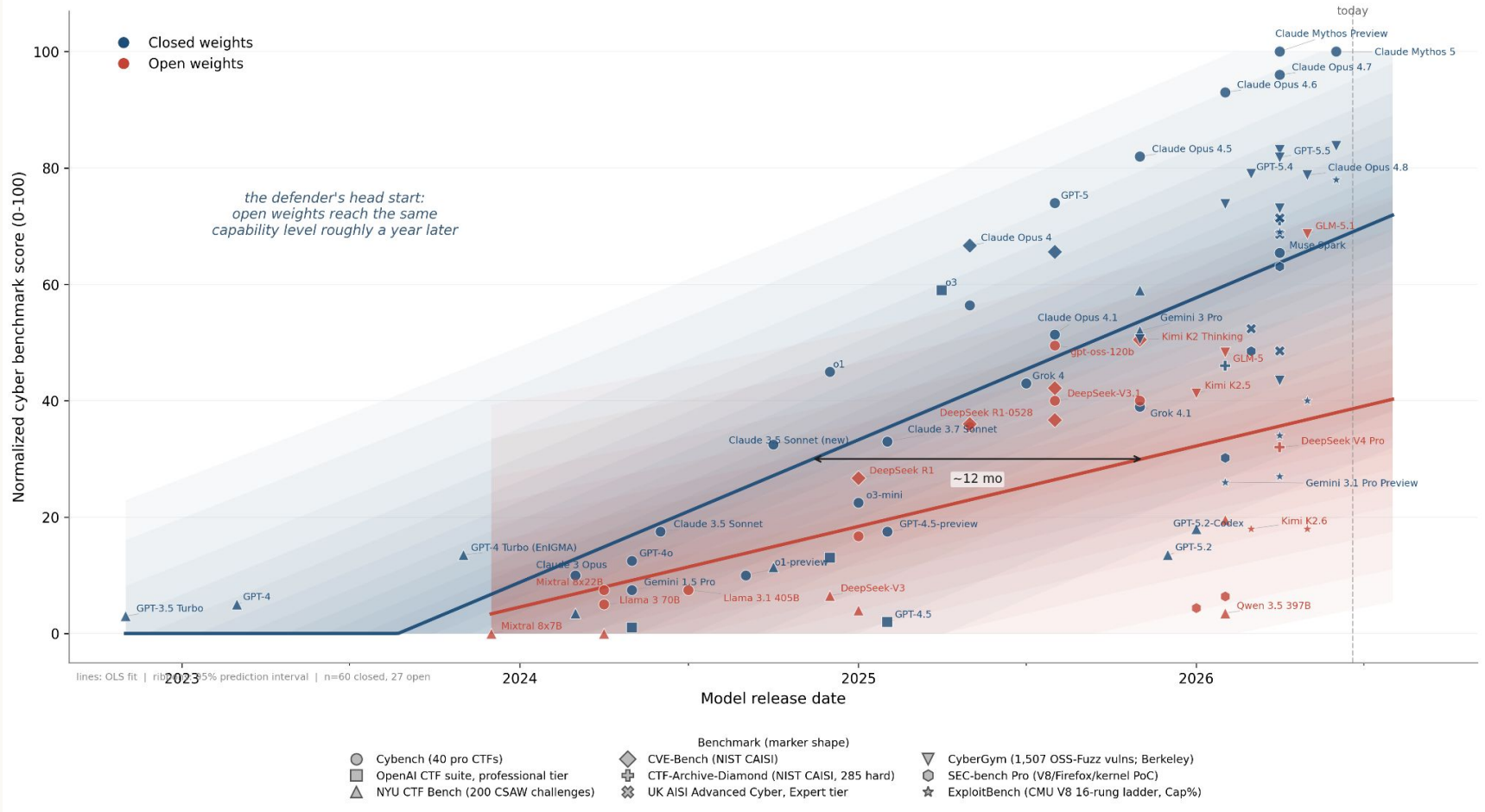
The permanent defender advantage is coming



Software with vulnerabilities can no longer be released

There's only so many ways for software to be wrong in a way that is useful for an attacker

It will take longer for OT and critical infrastructure

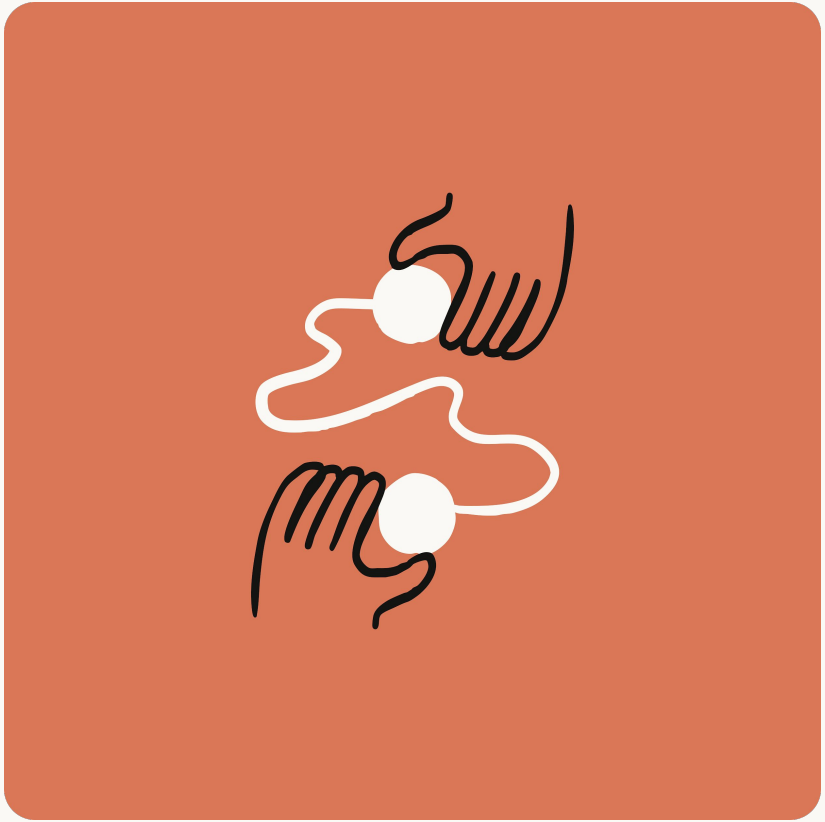

Offensive cyber capability over time: closed vs open weights (SOTA-at-release models only)
Closed weights
Open weights
today
Claude Mythos Preview
Claude Mythos 5
Claude Opus 4.7
Claude Opus 4.6
Claude Opus 4.5
GPT-5.5
GPT-5.4
Claude Opus 4.8
GPT-5
GLM-5.1
Muse Spark
Claude Opus 4
o3
Claude Opus 4.1
Gemini 3 Pro
Kimi K2 Thinking
gpt-oss-120b
GLM-5
o1
Grok 4
Kimi K2.5
DeepSeek-V3.1
DeepSeek R1-0528
Grok 4.1
Claude 3.7 Sonnet
Claude 3.5 Sonnet (new)
DeepSeek V4 Pro
DeepSeek R1
~12 mo
o3-mini
Gemini 3.1 Pro Preview
Claude 3.5 Sonnet
GPT-4.5-preview
GPT-5.2-Codex
Kimi K2.6
GPT-4 Turbo (EnIGMA)
GPT-4o
Claude 3 Opus
o1-preview
GPT-5.2
Mixtral 8x22B
Gemini 1.5 Pro
DeepSeek-V3
Llama 3 70B
Llama 3.1 405B
Qwen 3.5 397B
GPT-3.5 Turbo
GPT-4
GPT-4.5
Mixtral 8x7B
the defender's head start:
open weights reach the same
capability level roughly a year later
Normalized cyber benchmark score (0-100)
0
20
40
60
80
100
2023
2024
2025
2026
Model release date
lines: OLS fit | ribbon: 95% prediction interval | n=60 closed, 27 open
Benchmark (marker shape)
Cybench (40 pro CTFs)
OpenAI CTF suite, professional tier
NYU CTF Bench (200 CSAW challenges)
CVE-Bench (NIST CAISI)
CTF-Archive-Diamond (NIST CAISI, 285 hard)
UK AISI Advanced Cyber, Expert tier
CyberGym (1,507 OSS-Fuzz vulns; Berkeley)
SEC-bench Pro (V8/Firefox/kernel PoC)
ExploitBench (CMU V8 16-rung ladder, Cap%)

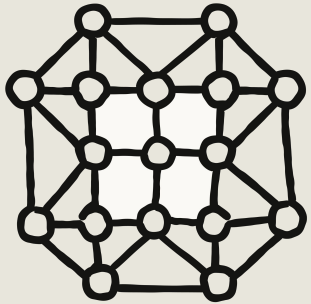
Preserving the defender advantage

How we defend for the next 7–10 months

Defender playbook

- Design for breach (zero-trust)
- D&R with scaling laws AI
- Vulnerability remediation

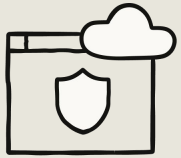




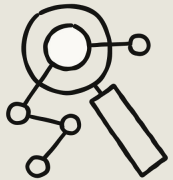
Don’t leak the model weights while making AI as widely available to defenders as possible.

Confidential computing's contribution

Advances in every hardware platform in use today: Trainium, TPU, and NVIDIA



Enables containment while accelerating reach



Every industry player now agrees that CoCo is part of the solution

ANTHROP\C