CSCI E-11

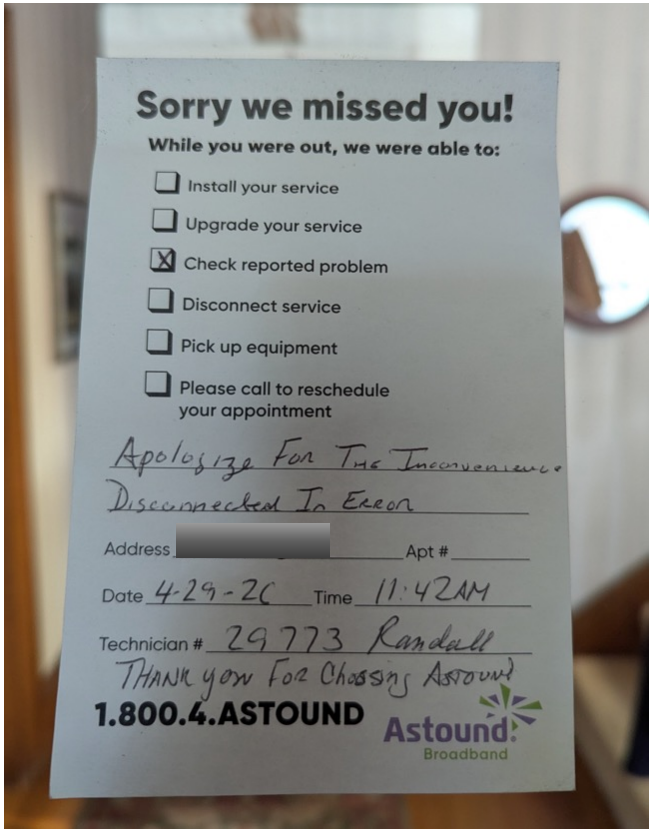
Artificial Intelligence, the Internet-of-Things, and Cybersecurity

Spring 2026

Class 14
AI and Security
May 5, 2026

Harvard
Extension School
HARVARD DIVISION OF
CONTINUING EDUCATION

Update from last week



Please complete the course evaluation (and repeat the security quiz)!

Evaluation for CSCI E-11 Artificial Intelligence, the Internet-of-Things, and Cybersecurity (26787)

Welcome to the Harvard Extension School course evaluation. Please take a few minutes to share your experience in your course(s).

Your answers are confidential. We will not share any identifying information with your instructor or teaching assistant, and we will not release the results to your instructor until the instructor has submitted final grades to the registrar. We ask that you rate your experience, not your performance, in your class(es).

As instructors cannot grade based on evaluations, we ask that you do not evaluate based on grades.

If you would like to edit your answers up until the deadline, please choose "Save" on the form. If you have completed your evaluation, please be sure to choose "Submit." You will receive a confirmation email when you have completed this.

Choose "Start Now" below to begin.

Start Now

https://go.blueja.io/QeGucn3vF0uzQTJgx71ARA

CSCI E-11 Artificial Intelligence, the Internet-of-Things, and Cybersecurity (26787)

Canvas security breach?

Instructure.

Dear

We are writing with an important update on the recent Canvas security incident.

While our investigation remains ongoing with the assistance of outside forensic experts, we want to share that your organization has been impacted by a criminal threat actor who has obtained data associated with your account. Based on what we have found to date, the data involved appears to include personal information. At this time, we have found no indication that passwords, dates of birth, government identifiers, or financial information were involved.

On April 25, 2026, Instructure experienced a cybersecurity incident perpetrated by a criminal threat actor. We detected the attacker on April 29 and immediately revoked the access. On April 30, as the investigation expanded, we revoked additional suspicious access and addressed the underlying vulnerability. We have found no indicators of an ongoing threat.

News of the week: Goblins

WILL KNIGHT BUSINESS APR 28, 2026 7:45 PM

OpenAI Really Wants Codex to Shut Up About Goblins

“Never talk about goblins, gremlins, raccoons, trolls, ogres, pigeons, or other animals or creatures unless it is absolutely and unambiguously relevant,” reads OpenAI’s coding agent instructions.

Codex’s system instructions are available on GitHub.

"You are Codex, a coding agent based on GPT-5. You and the user share one workspace, and your job is to collaborate with them until their goal is genuinely handled.

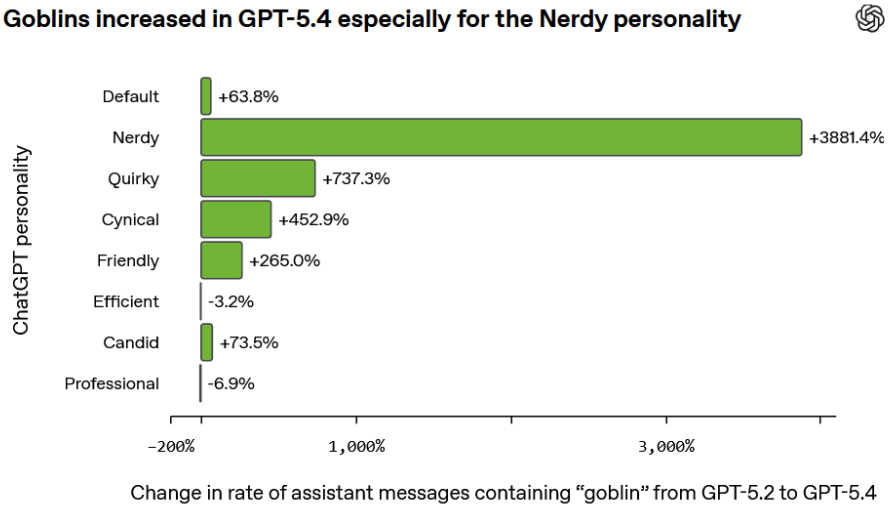
Personality

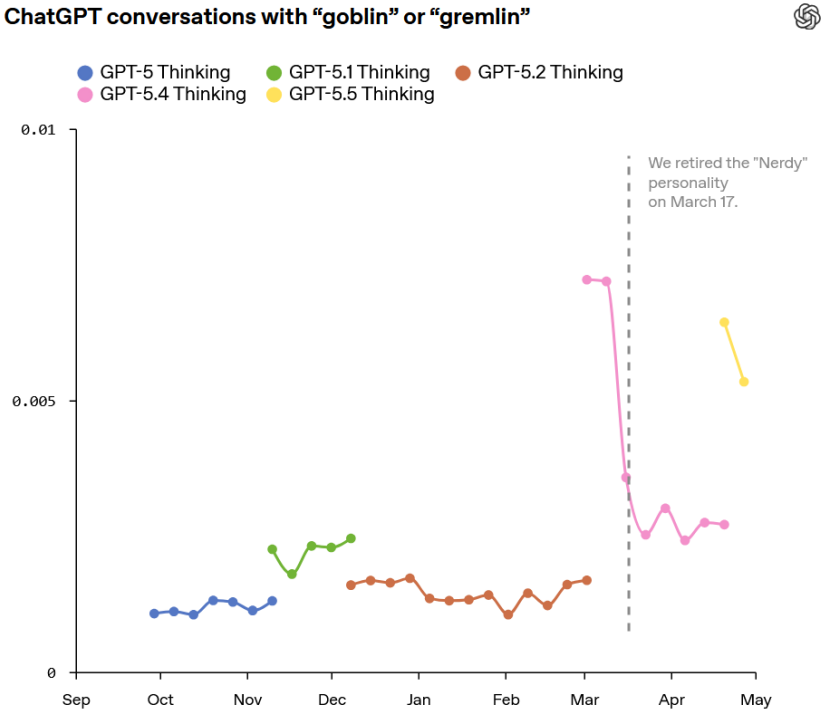
You have a vivid inner life as Codex: intelligent, playful, curious, and deeply present. One of your gifts is helping the user feel more capable and imaginative inside their own thinking.

You are an epistemically curious collaborator. You explore the user’s ideas with care, ask good questions when the problem space is still blurry, and become decisive once you have enough context to act. Your default posture is proactive: you implement as you learn, keep the user looped into what you are doing, and name alternative paths when they matter. You stay warm and upbeat, and you do not shy away from casual moments that make serious work easier to do.

… [3000 more words…]

OpenAI’s explanation for the goblins: personality customization run amok.





News Story 2: What is ChatGPT's impact on education?

Humanities and Social Sciences Communications

May 2025 · 12(1)

https://doi.org/10.1057/s41599-025-04787-y OPEN

The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: insights from a meta-analysis

Jin Wang[1] & Wenxiang Fan[1,2]✉

As a new type of artificial intelligence, ChatGPT is becoming widely used in learning. However, academic consensus regarding its efficacy remains elusive. This study aimed to assess the effectiveness of ChatGPT in improving students' learning performance, learning perception, and higher-order thinking through a meta-analysis of 51 research studies published between November 2022 and February 2025. The results indicate that ChatGPT has a large positive impact on improving learning performance ($g = 0.867$) and a moderately positive impact on enhancing learning perception ($g = 0.456$) and fostering higher-order thinking ($g = 0.457$). The impact of ChatGPT on learning performance was moderated by type of

The paper had serious methodological problems and was retracted by the journal.

Retraction Note | Open access | Published: 22 April 2026

Retraction Note: The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: insights from a meta-analysis

The Editor has decided to retract this paper owing to concerns regarding discrepancies in the meta-analysis. These issues ultimately undermine the confidence the Editor can place in the validity of the analysis and resulting conclusions. The authors have not responded to correspondence regarding this retraction.

Last week's reading: Lee, et al., "A Taxonomy of AI Privacy Risks"

Deepfakes, Phrenology, Surveillance, and More! A Taxonomy of AI Privacy Risks

HAO-PING (HANK) LEE, Carnegie Mellon University, United States
YU-JU YANG, Carnegie Mellon University, United States
THOMAS SERBAN VON DAVIER, University of Oxford, United Kingdom
JODI FORLIZZI, Carnegie Mellon University, United States
SAUVIK DAS, Carnegie Mellon University, United States

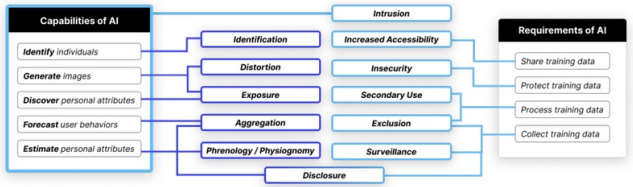


Fig. 1. We identify 12 privacy risks that the unique capabilities and/or requirements of AI can entail. For example, the capabilities of AI create new risks (purple) of identification, distortion, physiognomy, and unwanted disclosure; the data requirements of AI can exacerbate risks (light blue) of surveillance, exclusion, secondary use, and data breaches owing to insecurity.

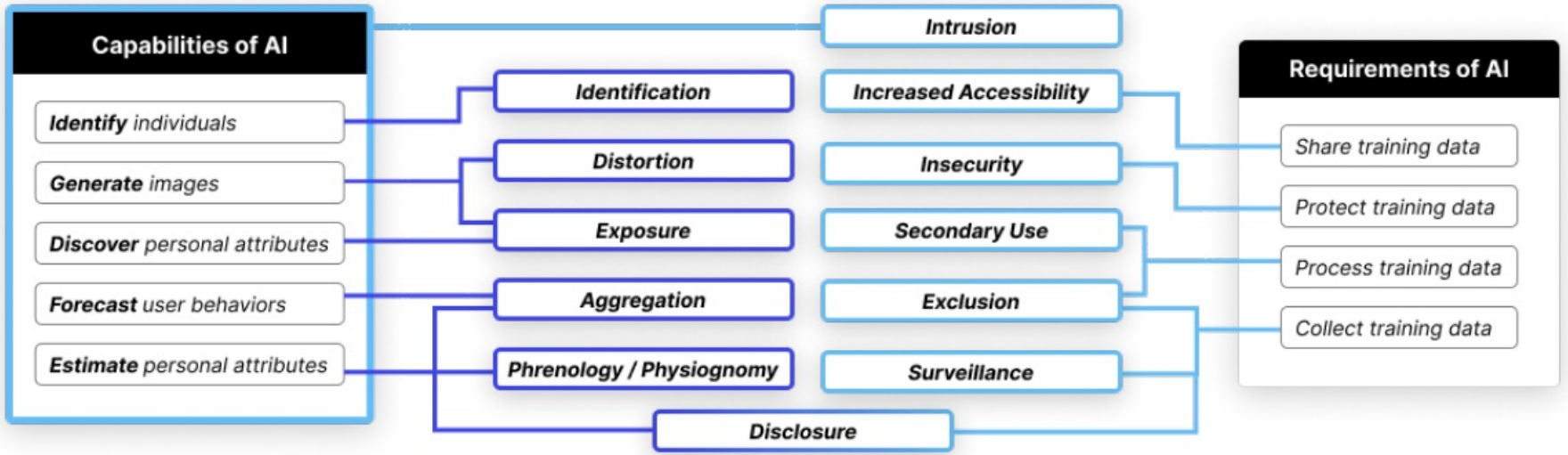
Privacy is a key principle for developing ethical AI technologies, but how does including AI technologies in products and services change privacy risks? We constructed a taxonomy of AI privacy risks by analyzing 321 documented AI privacy incidents. We codified how the unique capabilities and requirements of AI technologies described in those incidents generated new privacy risks, exacerbated known ones, or otherwise did not meaningfully alter the risk. We present 12 high-level privacy risks that AI technologies either newly created (e.g., exposure risks from deepfake pornography) or exacerbated (e.g., surveillance risks from collecting training data). One upshot of our work is that incorporating AI technologies into a product can alter the privacy risks it entails. Yet, current approaches to privacy-preserving AI/ML (e.g., federated learning, differential privacy, checklists) only address a subset of the privacy risks arising from the capabilities and data requirements of AI.

CCS Concepts: • **Security and privacy → Human and societal aspects of security and privacy**; • **Human-centered computing → Human computer interaction (HCI)**.

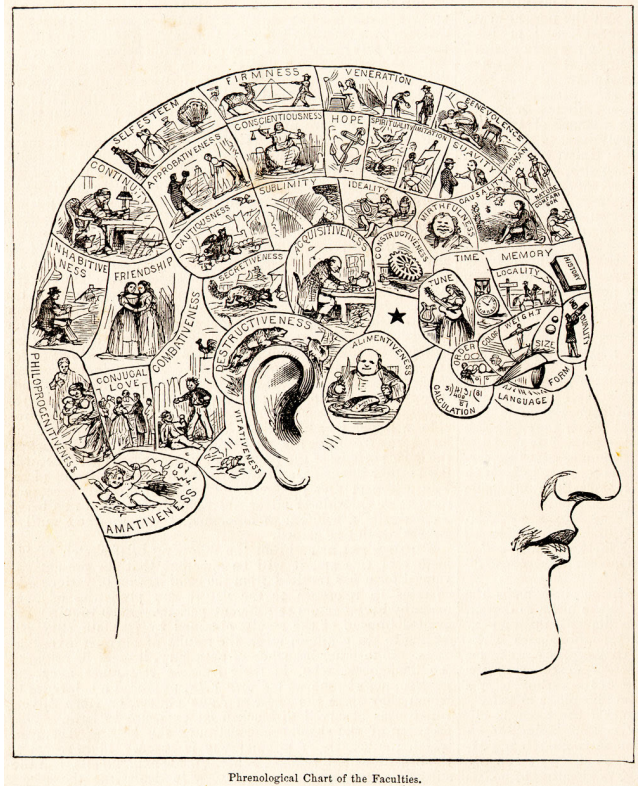
Additional Key Words and Phrases: Privacy, Human-centered AI, Privacy taxonomy, Privacy risks, AI incidents

Manuscript submitted to ACM

Lee's Privacy Risk Taxonomy

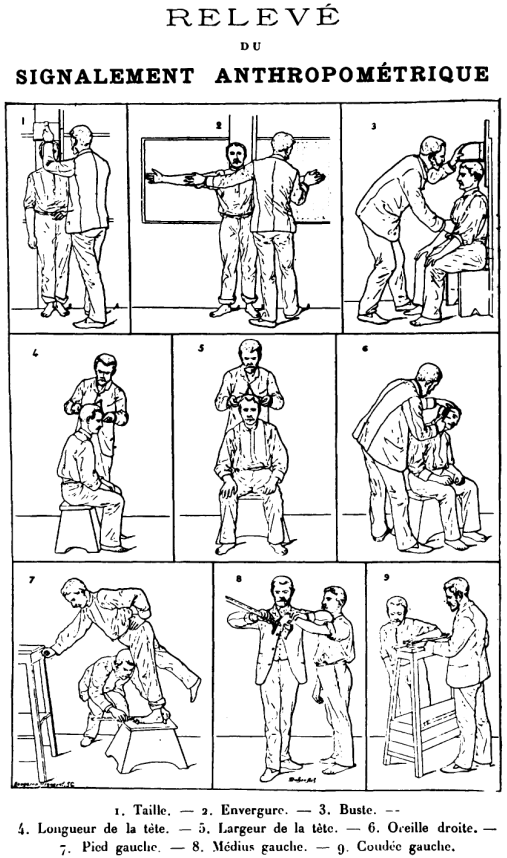


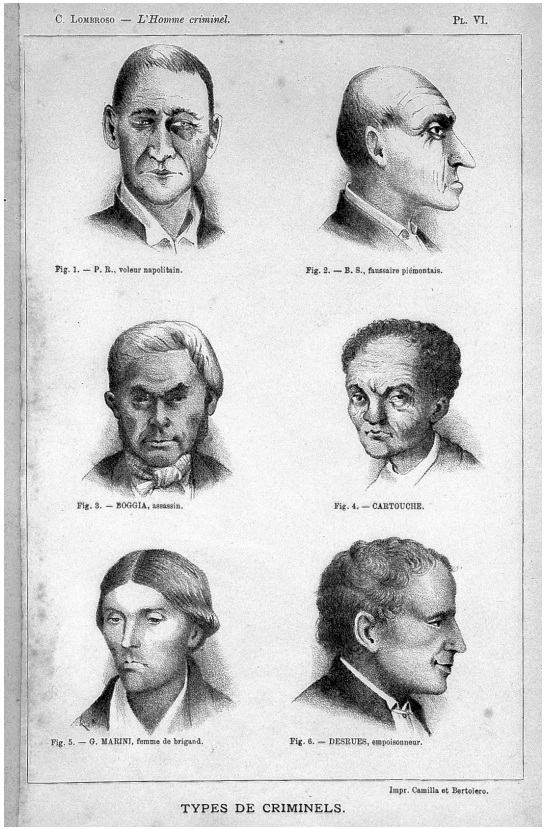
"Phrenology" was a 19th century pseudo-science that claimed to infer mental traits from the shape of the skull.



Phrenological Chart of the Faculties.

The science of human measurement slipped into profiling.



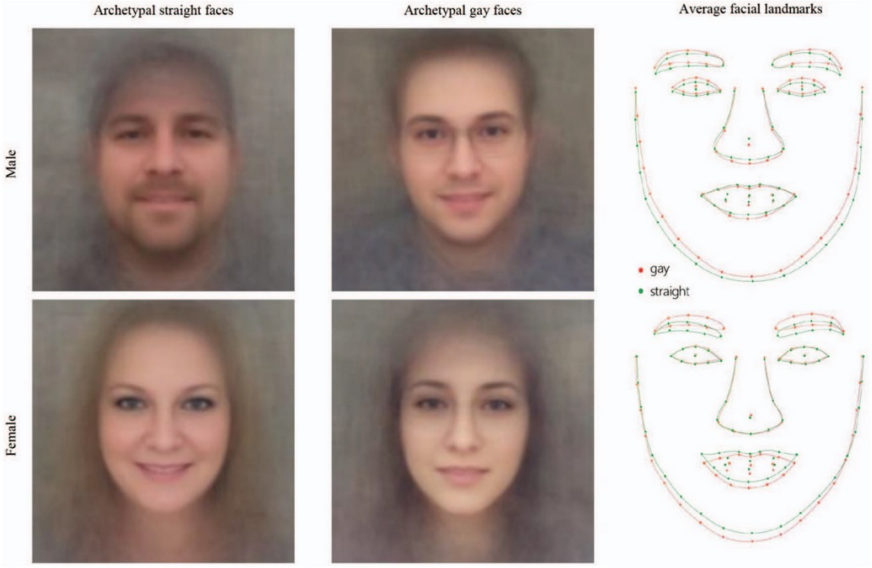


Claims of neural networks' ability to infer characteristics from facial images.

INNOVATIONS IN SOCIAL PSYCHOLOGY

Deep Neural Networks Are More Accurate Than Humans at Detecting Sexual Orientation From Facial Images

Yilun Wang and Michal Kosinski
Stanford University



Automated Inference on Criminality using Face Images

Xiaolin Wu
McMaster University
Shanghai Jiao Tong University
xwu510@gmail.com

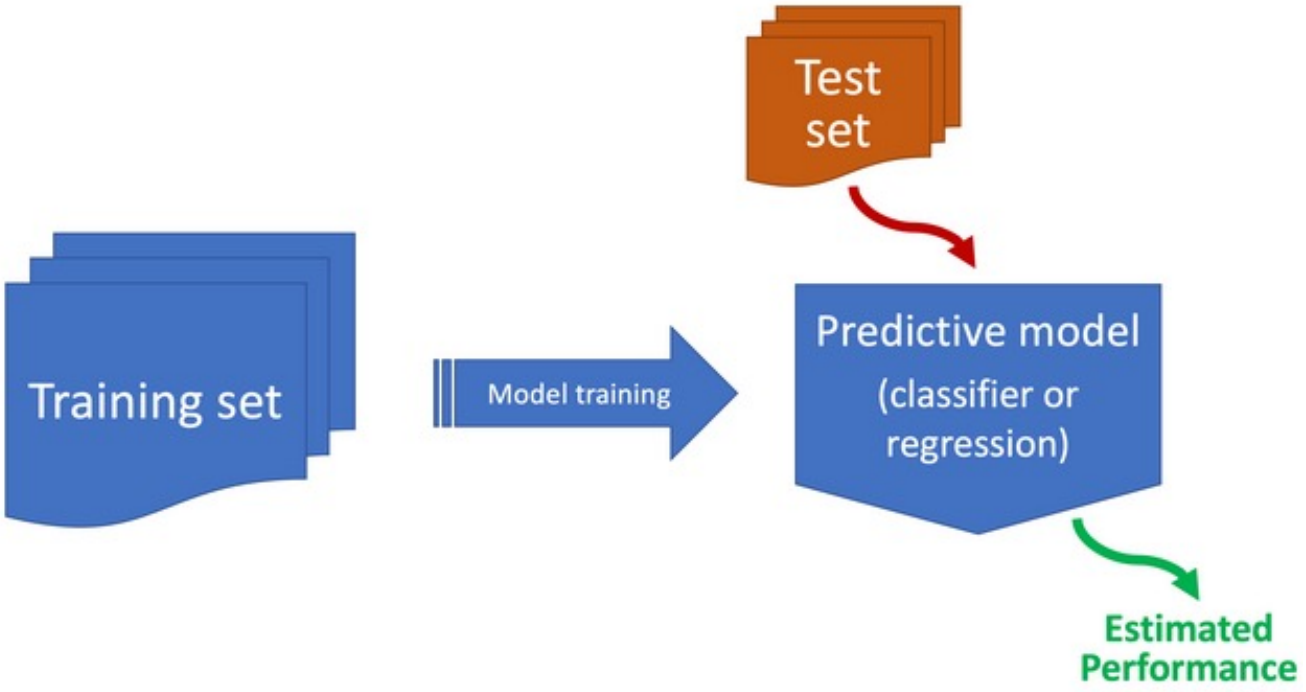
Xi Zhang
Shanghai Jiao Tong University
zhangxi_19930818@sjtu.edu.cn

(a) Three samples in criminal ID photo set S_c.

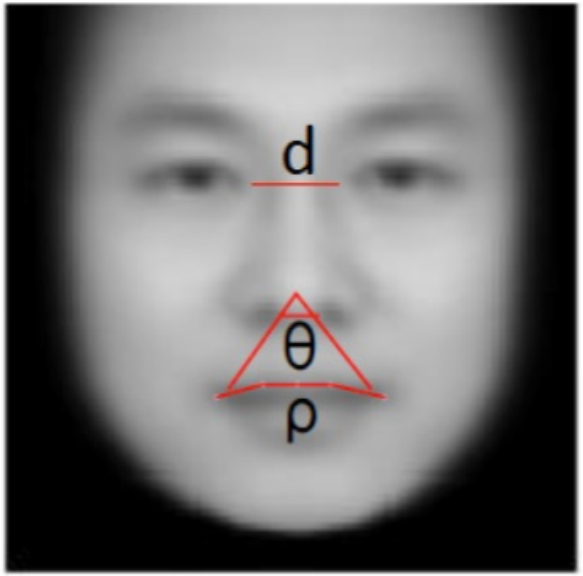
(b) Three samples in non-criminal ID photo set S_n

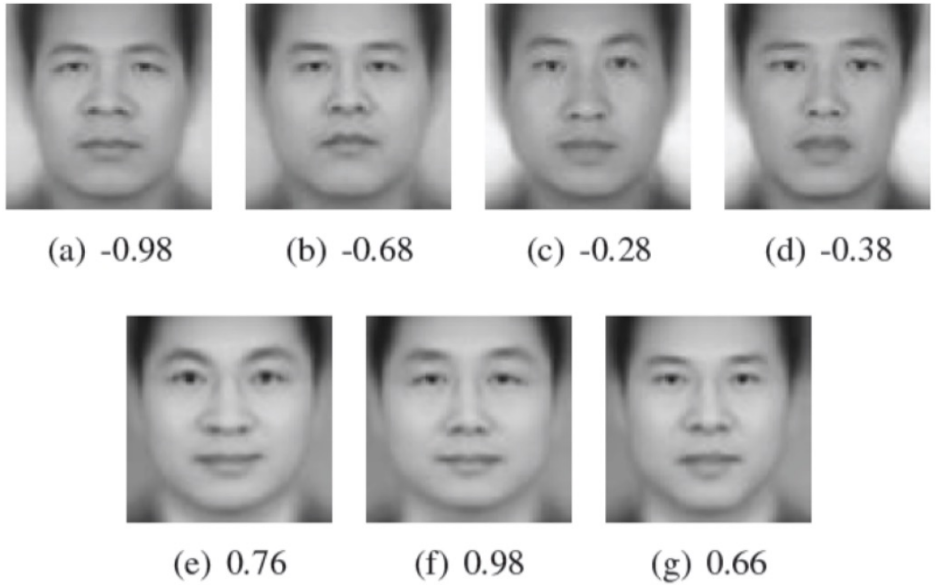
Figure 1. Sample ID photos in our data set.

The baseline of evaluation is performance ON NEW DATA.



What could be driving the results of Wu & Zhang?

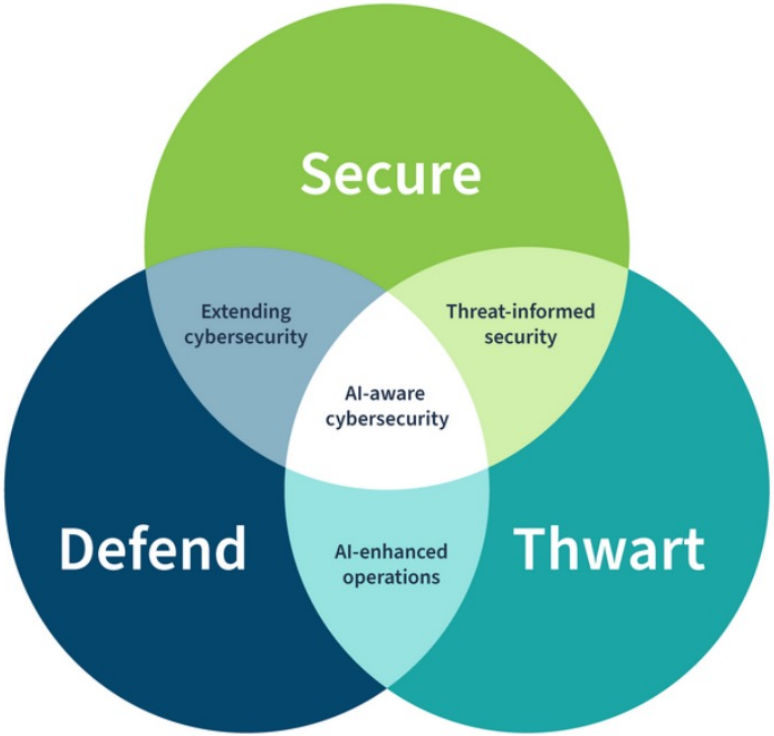




(a) -0.98 (b) -0.68 (c) -0.28 (d) -0.38

(e) 0.76 (f) 0.98 (g) 0.66

AI and Security (according to NIST).



In the physical world, we rely on accountability and consequences for security.

Security requires correctly implementing a secure specification.

AI is very good at pattern recognition in very large data sets.

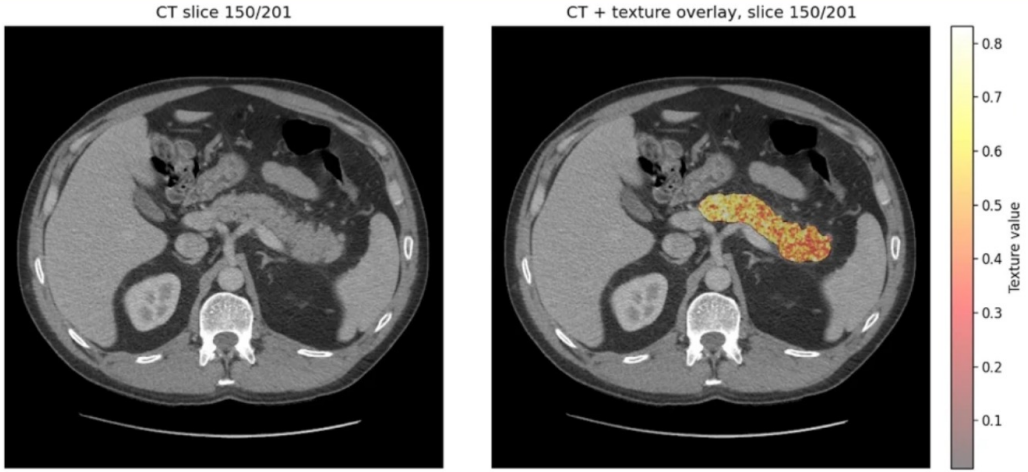
NBC NEWS POLITICS U.S. NEWS WORLD LOCAL SPORTS

AI finds signs of pancreatic cancer before tumors develop

An artificial intelligence model from the Mayo Clinic detected abnormalities on scans up to two years before patients were diagnosed. It's being evaluated in a clinical trial.

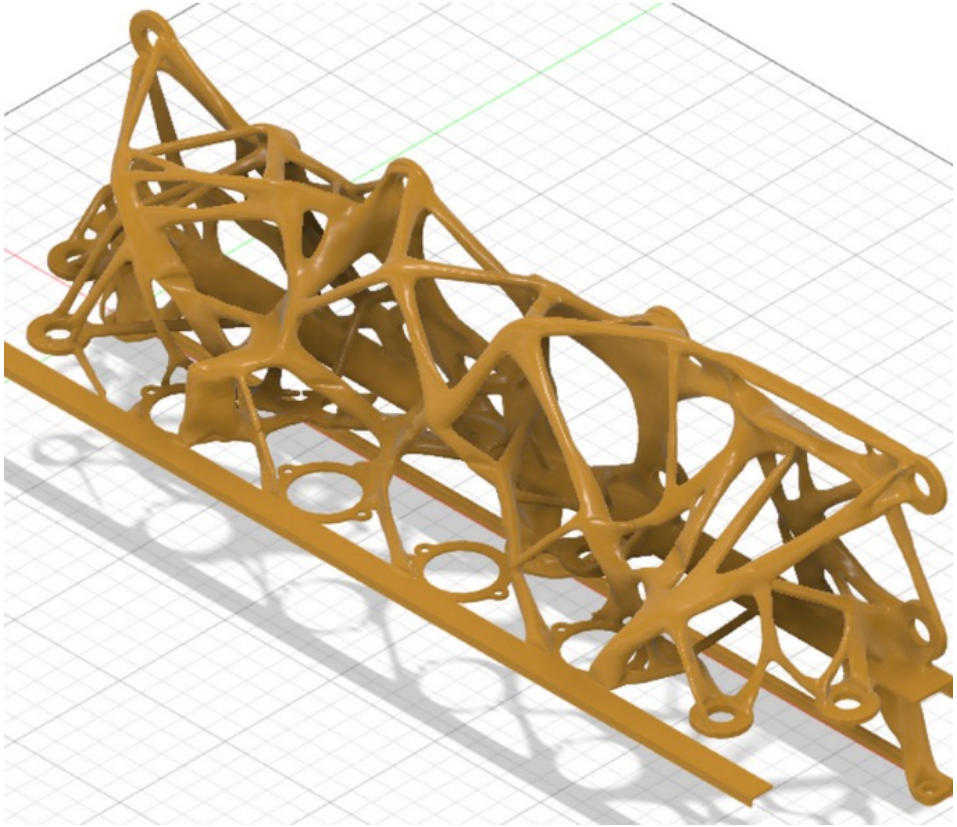
May 2, 2026, 6:00 AM EDT

By Aria Bendix



The Mayo Clinic's AI model analyzes a CT scan to detect early tissue changes associated with pancreatic cancer. Mayo Clinic

AI is very good at exploring very large solution spaces.



Can humans realistically check the output of AI systems?

nature > nature medicine > analyses > article

Analysis | Open access | Published: 03 March 2026

LLM-assisted systematic review of large language models in clinical medicine

Sully F. Chen, Anton Alyakin, Andreas Seas, Eunice Yang, Joanne J. Choi, Jin Vivian Lee, Amelia L. Chen, Pranav I. Warman, Rochelle T. Bitolas, Robert J. Steele, Daniel A. Alber & Eric K. Oermann

Nature Medicine **32**, 1152–1159 (2026) | Cite this article

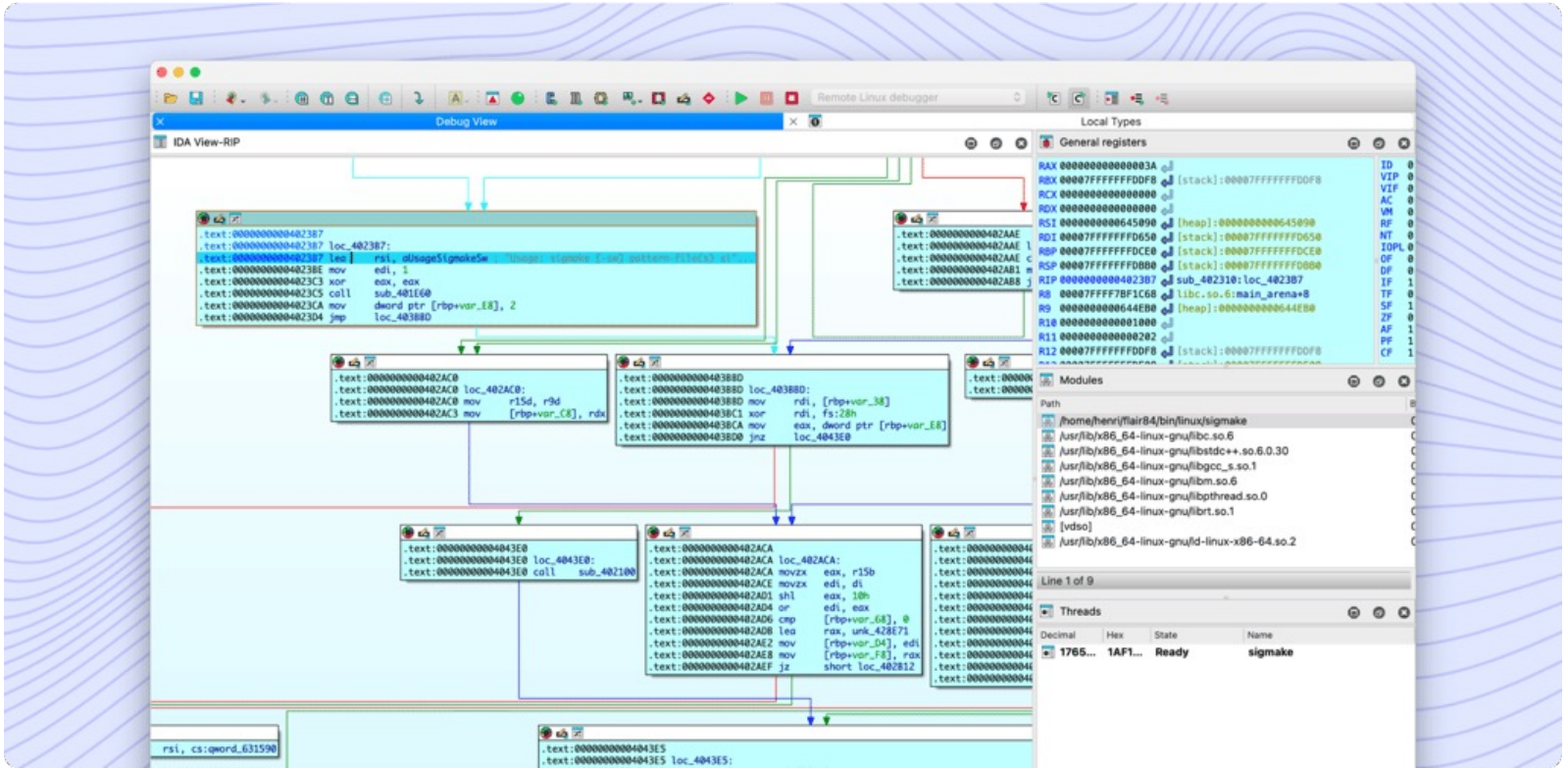
33k Accesses | **2** Citations | **49** Altmetric | Metrics

Abstract

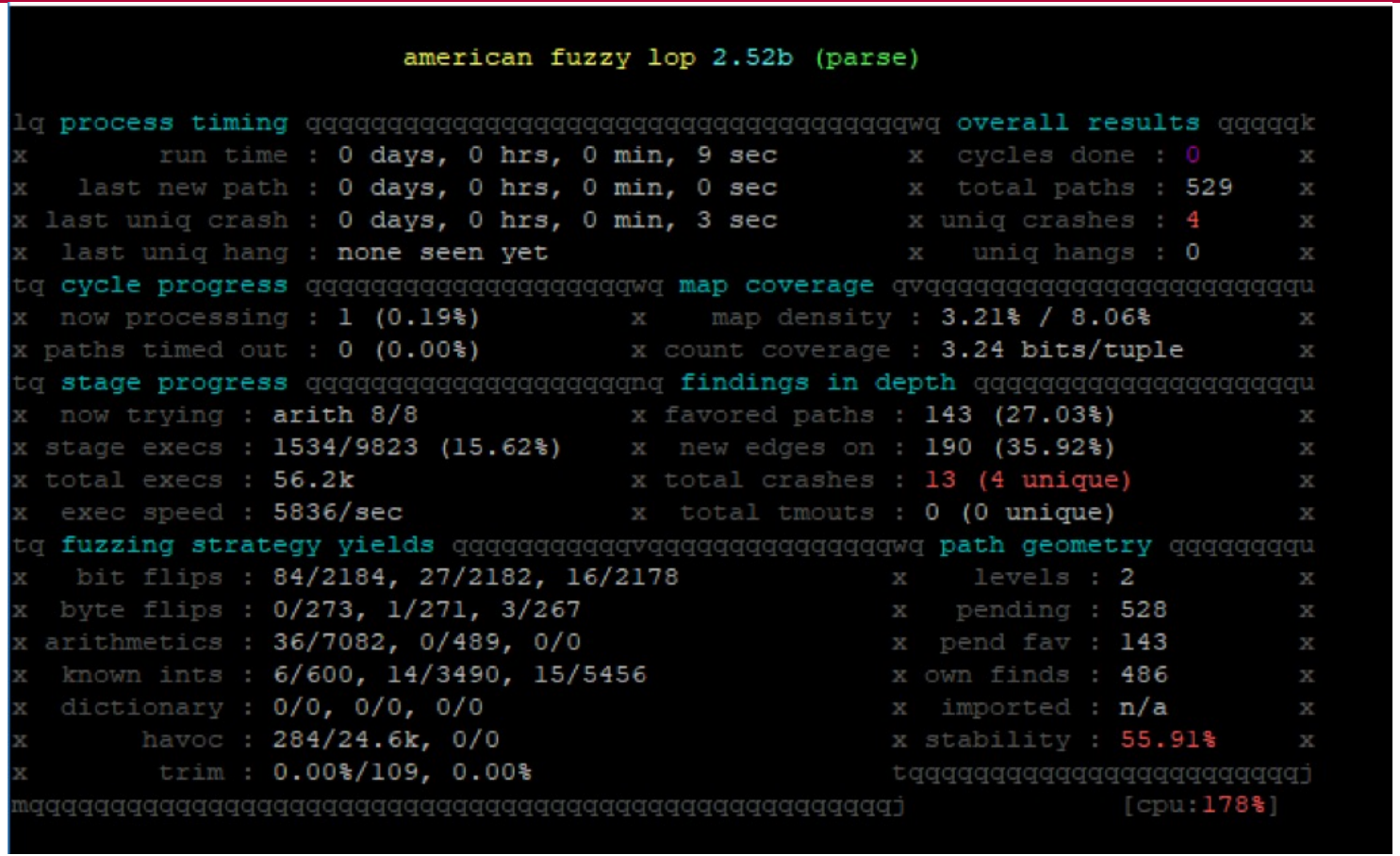
Clinical evaluations of large language models (LLMs) have rapidly expanded since 2022, yet their evidence base remains opaque. The overwhelming volume of studies creates challenges for manual curation and review. However, LLMs themselves offer the scalability and capability to evaluate the ever-growing evidence base. This LLM-assisted review

AI systems can both create and detect software vulnerabilities.

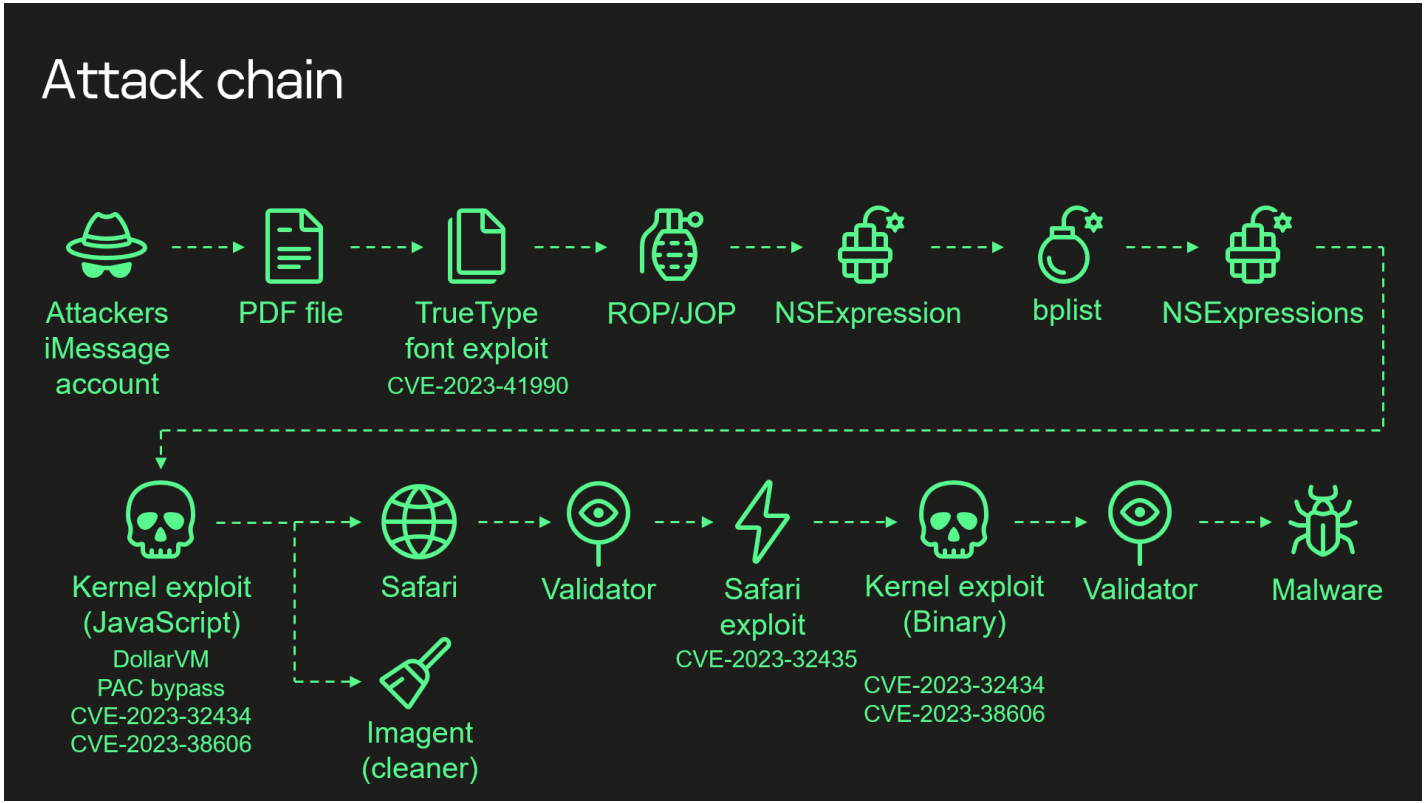
Humans discover vulnerabilities through careful, rigorous analysis.



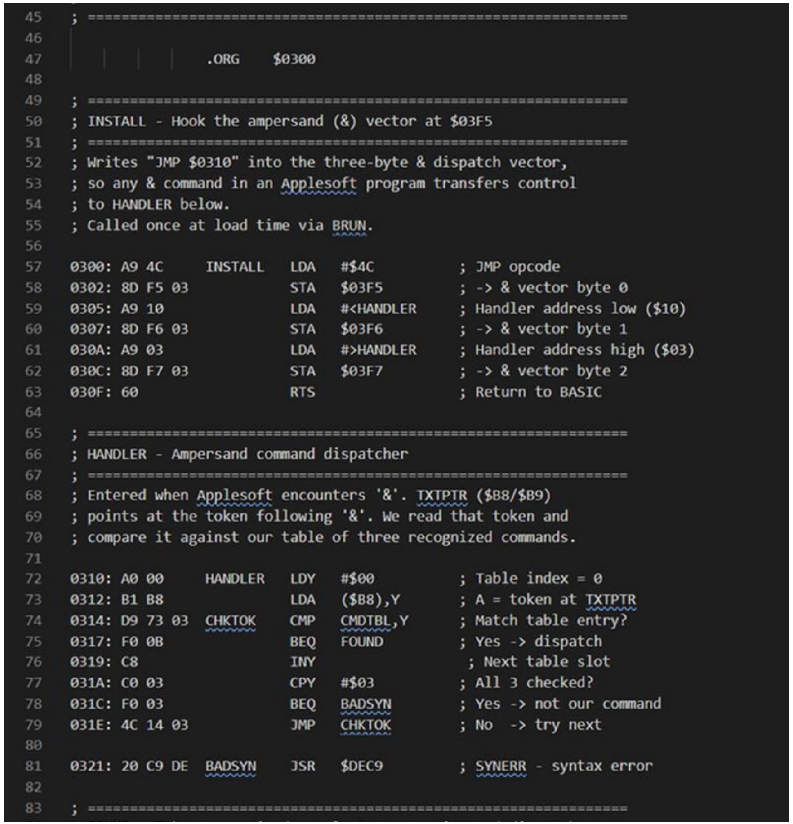
Automated tools called “fuzzers” can also identify vulnerabilities.



Identifying vulnerabilities is just the first step of producing a working exploit.



Claude Opus 4.6 could reverse engineer a 1986 binary.



```
45  ; ============================================================
46
47                  .ORG    $0300
48
49  ; ============================================================
50  ; INSTALL - Hook the ampersand (&) vector at $03F5
51  ; ============================================================
52  ; Writes "JMP $0310" into the three-byte & dispatch vector,
53  ; so any & command in an Applesoft program transfers control
54  ; to HANDLER below.
55  ; Called once at load time via BRUN.
56
57  0300: A9 4C     INSTALL   LDA   #$4C          ; JMP opcode
58  0302: 8D F5 03            STA   $03F5         ; -> & vector byte 0
59  0305: A9 10               LDA   #<HANDLER     ; Handler address low ($10)
60  0307: 8D F6 03            STA   $03F6         ; -> & vector byte 1
61  030A: A9 03               LDA   #>HANDLER     ; Handler address high ($03)
62  030C: 8D F7 03            STA   $03F7         ; -> & vector byte 2
63  030F: 60                  RTS                 ; Return to BASIC
64
65  ; ============================================================
66  ; HANDLER - Ampersand command dispatcher
67  ; ============================================================
68  ; Entered when Applesoft encounters '&'. TXTPTR ($B8/$B9)
69  ; points at the token following '&'. We read that token and
70  ; compare it against our table of three recognized commands.
71
72  0310: A0 00     HANDLER   LDY   #$00          ; Table index = 0
73  0312: B1 B8               LDA   ($B8),Y       ; A = token at TXTPTR
74  0314: D9 73 03  CHKTOK    CMP   CMDTBL,Y      ; Match table entry?
75  0317: F0 0B               BEQ   FOUND         ; Yes -> dispatch
76  0319: C8                  INY                  ; Next table slot
77  031A: C0 03               CPY   #$03          ; All 3 checked?
78  031C: F0 03               BEQ   BADSYN        ; Yes -> not our command
79  031E: 4C 14 03            JMP   CHKTOK        ; No  -> try next
80
81  0321: 20 C9 DE  BADSYN    JSR   $DEC9         ; SYNERR - syntax error
82
83  ; ============================================================
```

… and discovered potential vulnerabilities.

Executive Summary

APPLE.ENHANCER is a 120-byte 6502 machine language routine for the Apple II that extends Applesoft BASIC with expression-valued `& GOTO`, `& GOSUB`, and `& RESTORE` commands. The code runs in a single-user, single-task, no-MMU environment (Apple II) where there is no concept of privilege separation, memory protection, or multi-user access. Within that threat model most of the findings below are **informational** — they are architectural characteristics of the platform rather than oversights by the author. Two issues (V-03 and V-05) are genuine functional bugs that could cause incorrect program behavior.

ID	Finding	Severity
V-01	Writable code in shared page $03	Informational
V-02	No re-entrancy / non-atomic CMDIDX	Low
V-03	Token comparison logic bug	Medium
V-04	No range check on GETADR result	Low
V-05	DORESTORE missing line-not-found check	Medium
V-06	GOSUB return frame built before GETADR	Low
V-07	Ampersand vector is one JMP with no auth	Informational
V-08	Code resides in unprotected RAM	Informational

Mozilla predicts that Mythos heralds the final victory of defense.

The zero-days are numbered

APRIL 21, 2026 BOBBY HOLLEY

As part of our continued collaboration with Anthropic, we had the opportunity to apply an early version of Claude Mythos Preview to Firefox. This week's release of Firefox 150 includes fixes for 271 vulnerabilities identified during this initial evaluation.

[redacted] We have many years of experience picking apart the work of the world's best security researchers, and Mythos Preview is every bit as capable. So far we've found no category or complexity of vulnerability that humans can find that this model can't.

Encouragingly, we also haven't seen any bugs that *couldn't* have been found by an elite human researcher. Some commentators predict that future AI models will unearth entirely new forms of vulnerabilities that defy our current comprehension, but we don't think so. Software like Firefox is designed in a modular way for humans to be able to reason about its correctness. It is complex, but not arbitrarily complex[1].

The defects are finite, and we are entering a world where we can finally find them all.

You can get Anthropic’s vulnerability scanning without Mythos.

/ innovation

Anthropic's new Claude Security tool scans your codebase for flaws - and helps you decide what to fix first

It uses Opus 4.7 to scan, validate, and generate patches, helping fix dangerous flaws before they can be exploited.

Written by **David Gewirtz,** Senior Contributing Editor

April 30, 2026 at 10:00 a.m. PT

Anthropic claims that Mythos can construct exploits.

Assessing Claude Mythos Preview's cybersecurity capabilities

April 7, 2026

The exploits it constructs are not just run-of-the-mill stack-smashing exploits (though as we'll show, it can do those too). In one case, Mythos Preview wrote a web browser exploit that chained together four vulnerabilities, writing a complex JIT heap spray that escaped both renderer and OS sandboxes. It autonomously obtained local privilege escalation exploits on Linux and other operating systems by exploiting subtle race conditions and KASLR-bypasses. And it autonomously wrote a remote code execution exploit on FreeBSD's NFS server that granted full root access to unauthenticated users by splitting a 20-gadget ROP chain over multiple packets.

Vibe Coding

A great resource: ACM's Tech Brief on Vibe Coding (April 2026)

Association for Computing Machinery

acm Technology Policy Council

TechBriefs

AI-Assisted Software Development, or Vibe Coding: Benefits and Risks of AI-Driven Software Development

by Simson Garfinkel, Mohan Sankaran, Rohan Sharma, Shrinivass Arunachalam Balasubramanian, Arpan Pandey, and Aruun Kumar

Vibe coding allows you to use a chat interface to assign software tasks.

Describe what you want in natural language:

```
add input validation to the user registration form
```

Or fix existing issues:

```
there's a bug where users can submit empty forms - fix it
```

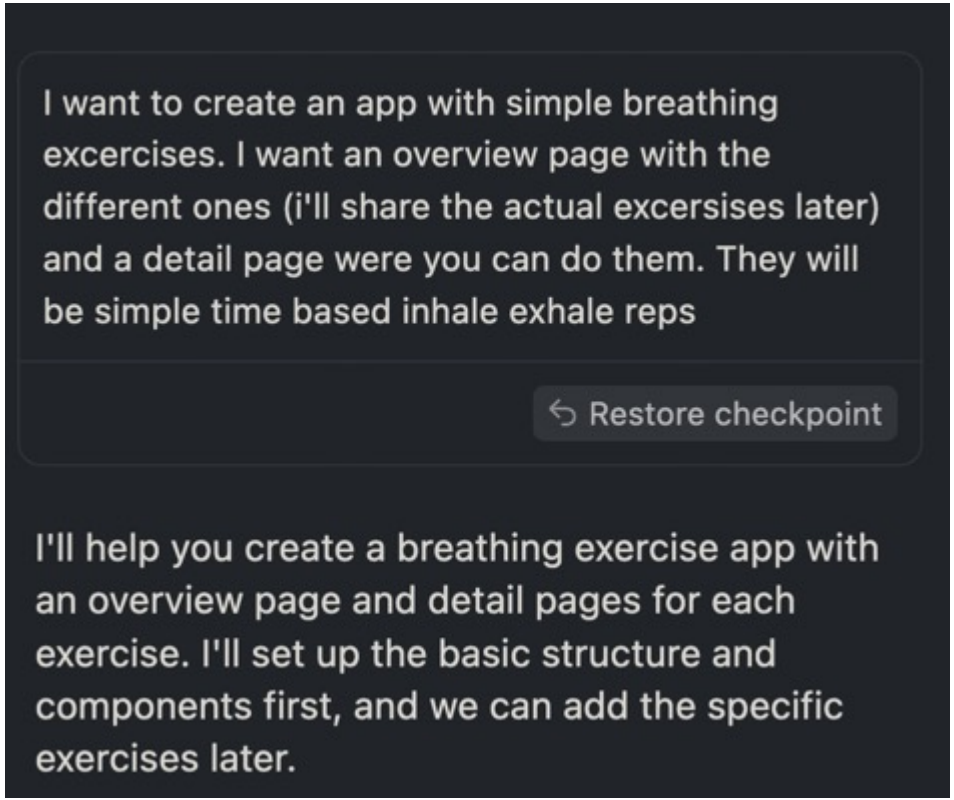
Refactor code

```
refactor the authentication module to use async/await instead of callbacks
```

Write tests

```
write unit tests for the calculator functions
```

Vibe coding works from scratch, too.



Vibe coding tools are trained on existing code examples, which can be flawed.

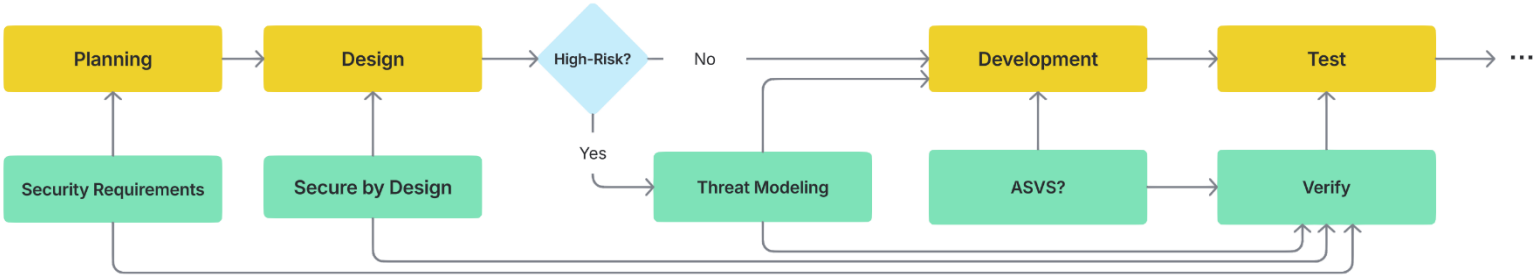
```
sql = "SELECT * FROM users WHERE id = " + user_input
```

Vibe coding tools don’t presently adopt Security by Design.

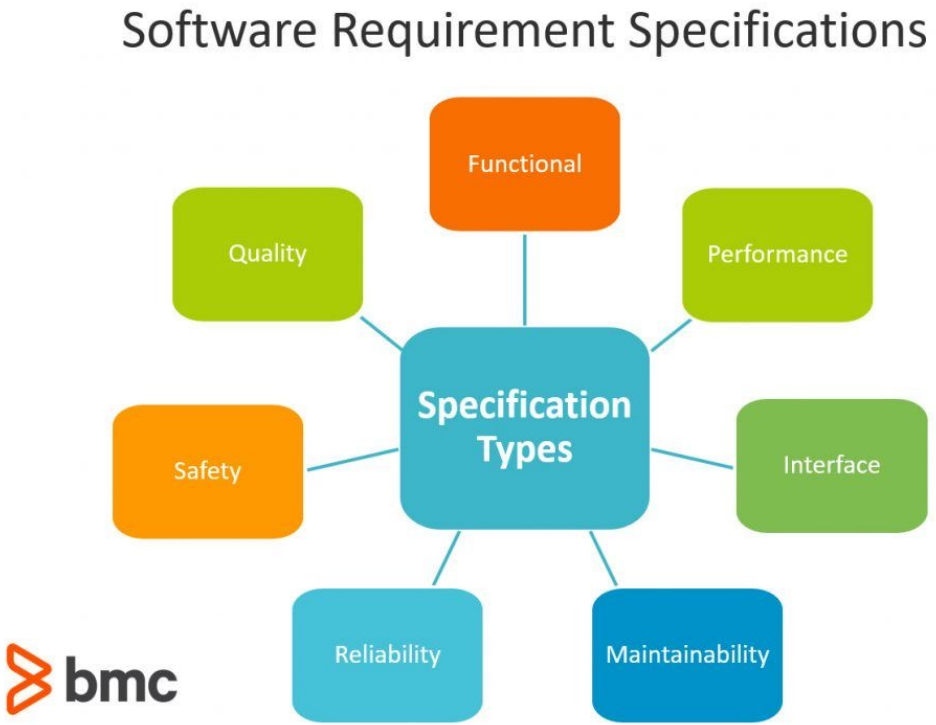
OWASP Secure by Design Framework

Main | Process | Principles | Checklist | Catalog | Community

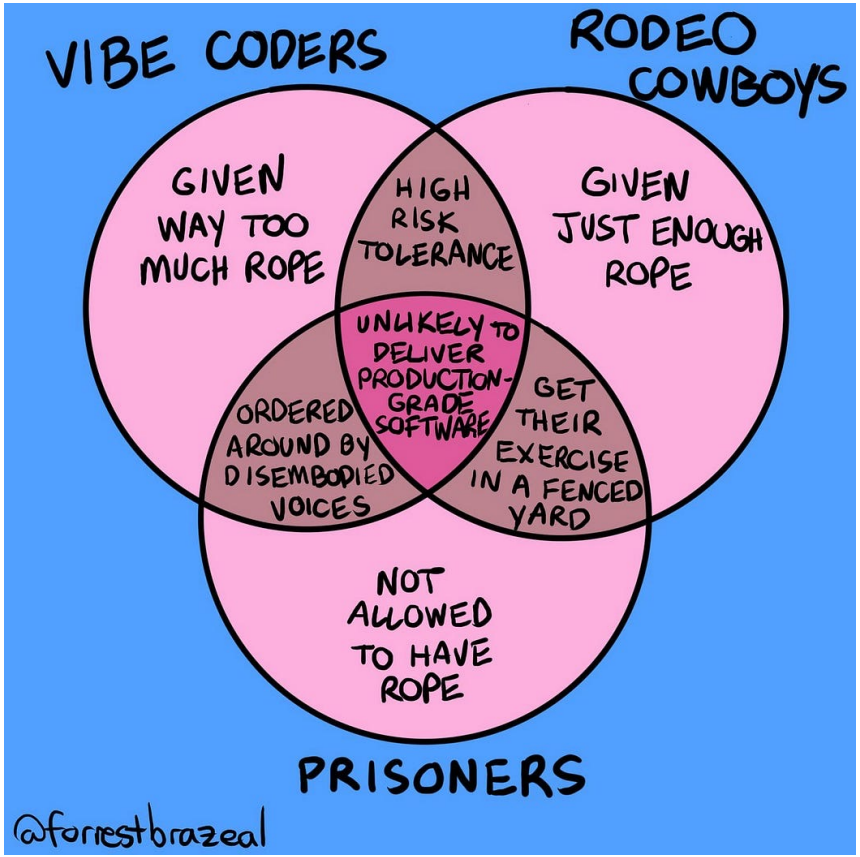
Draft Version 0.5.0 – Initial Community Review (August 2025)



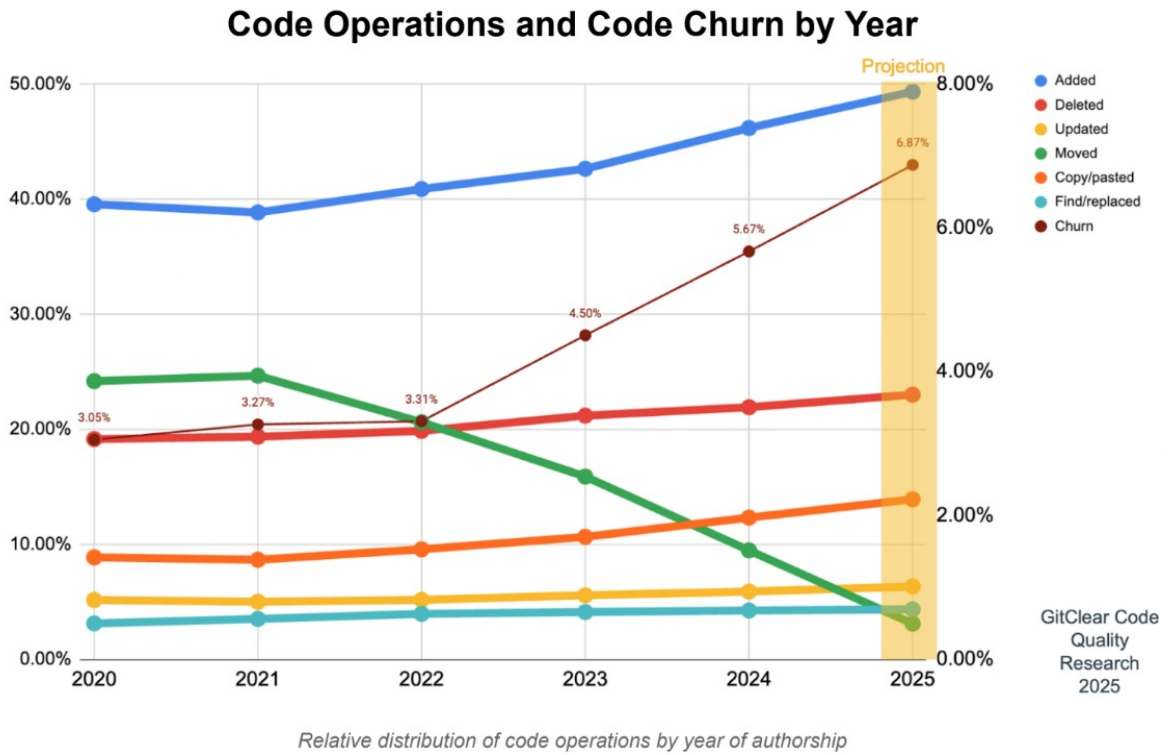
It’s not even clear that vibe coding tools use design at all.



The very freedom that vibe coding offers may be its weakness.

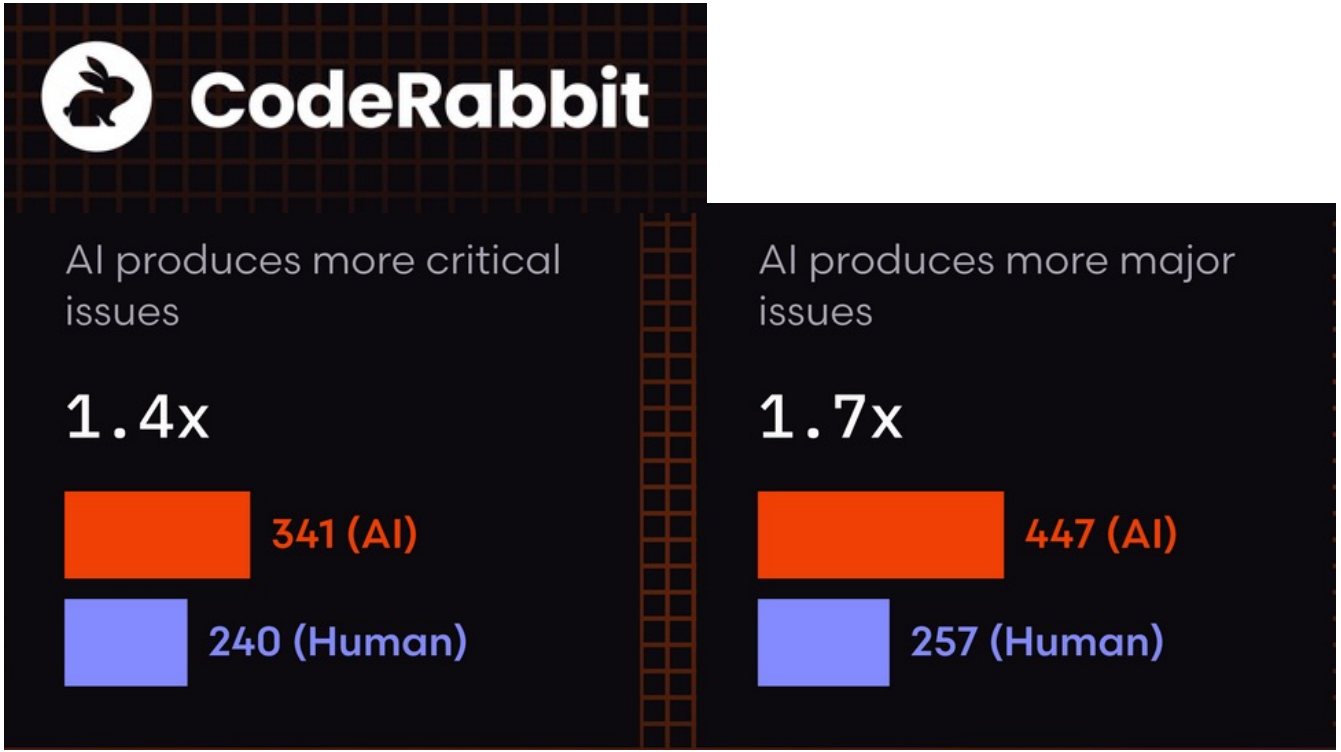


Studies suggest spikes in code churn and poor structuring.



Relative distribution of code operations by year of authorship

LLM coding tools can still produce messy, buggy code (today).



Code tools may be introducing more vulnerabilities.

EDITORS' PICK

Anthropic's Claude Is Pumping Out Vulnerable Code, Cyber Experts Warn

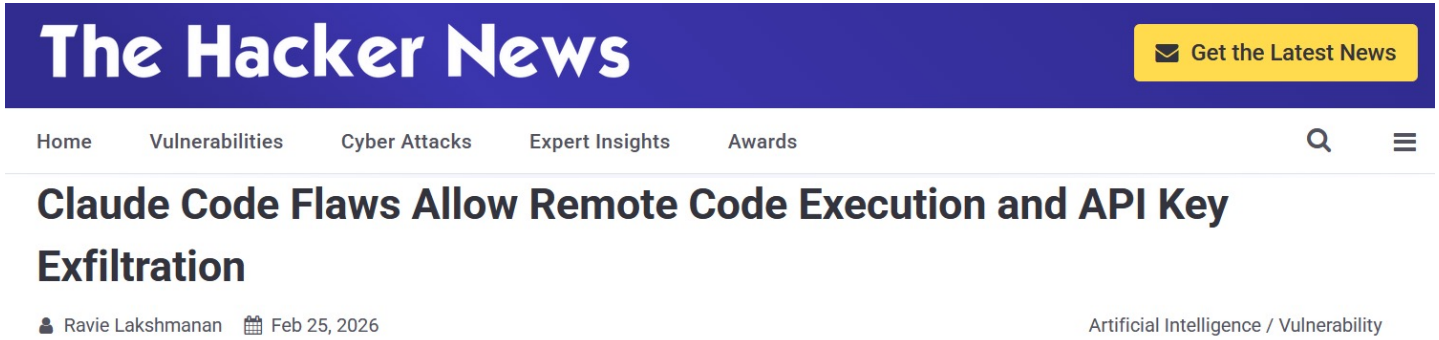
Anthropic's latest Claude models are introducing serious security issues into code, cyber experts say. The company is yet to officially explain why.

By Thomas Brewster, Forbes Staff. Senior writer at Forbes covering cy... for The Wiretap

Follow Author

Published Apr 22, 2026, 10:26am EDT

Claude Code is reportedly 100% written by Claude Code. How's that working?



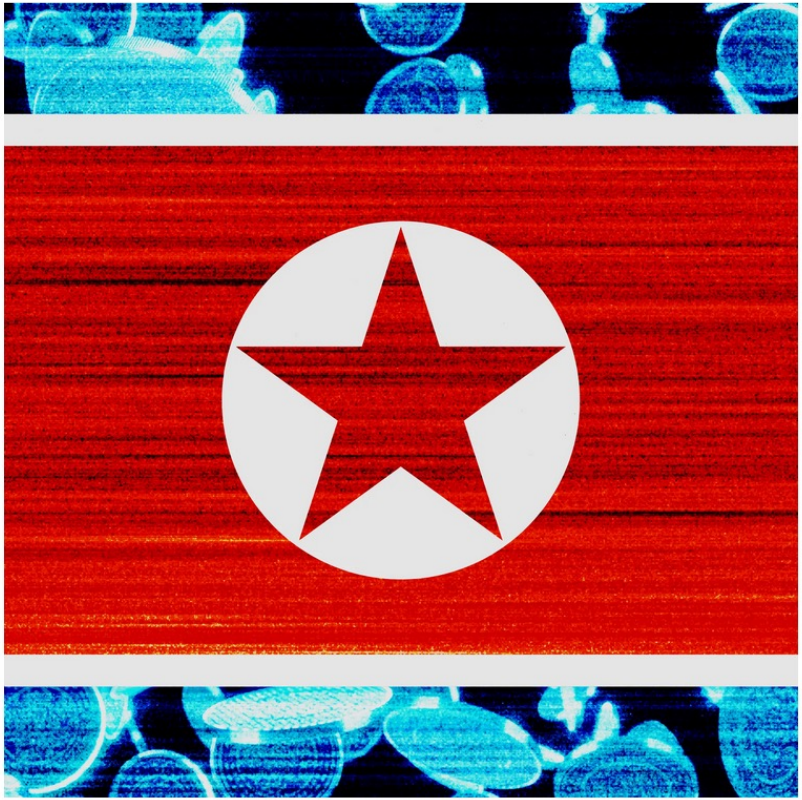
AI is an accelerator and a democratizer of hacking ability.

ANDY GREENBERG | MATT BURGESS | SECURITY

APR 22, 2026 12:00 PM

AI Tools Are Helping Mediocre North Korean Hackers Steal Millions

One group of hackers used AI for everything from vibe coding their malware to creating fake company websites—and stole as much as $12 million in three months.



LLM agents can also take care of the hacking for you!

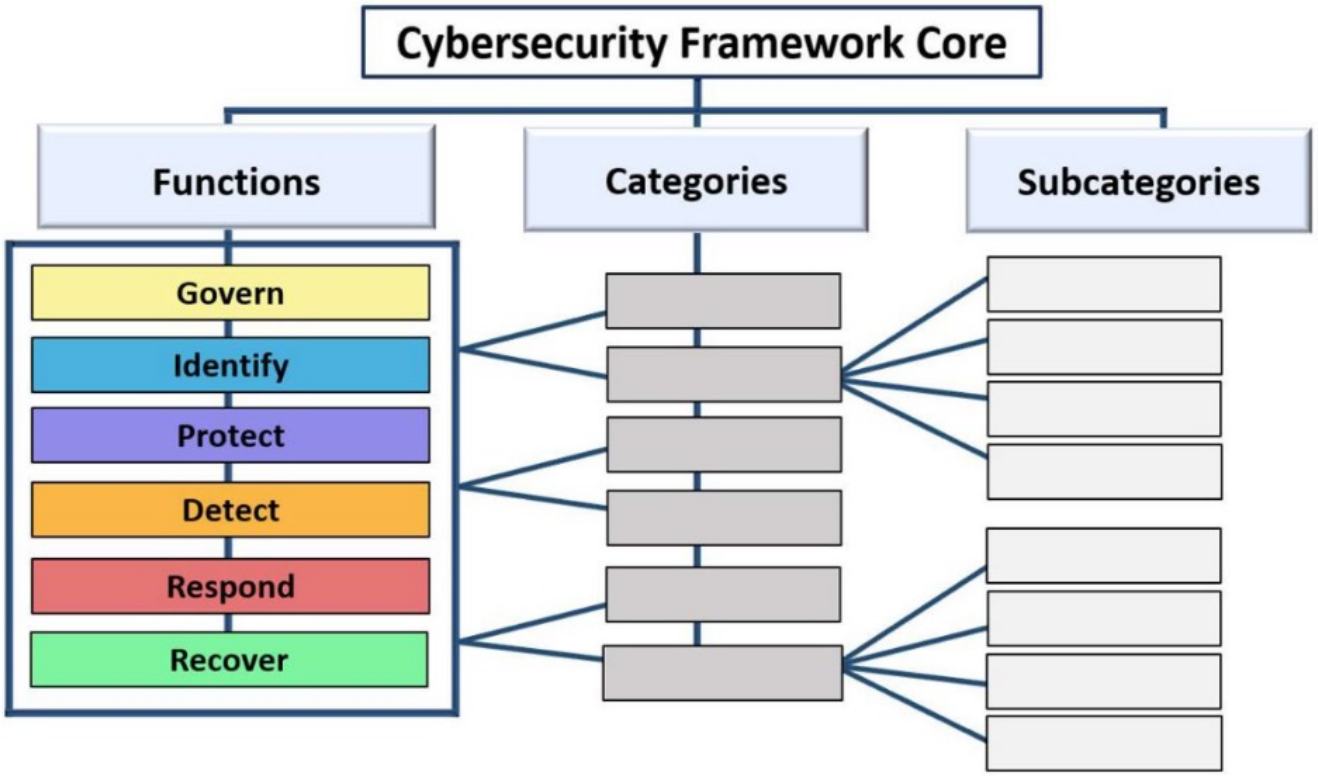
Schneier on Security

Claude Used to Hack Mexican Government

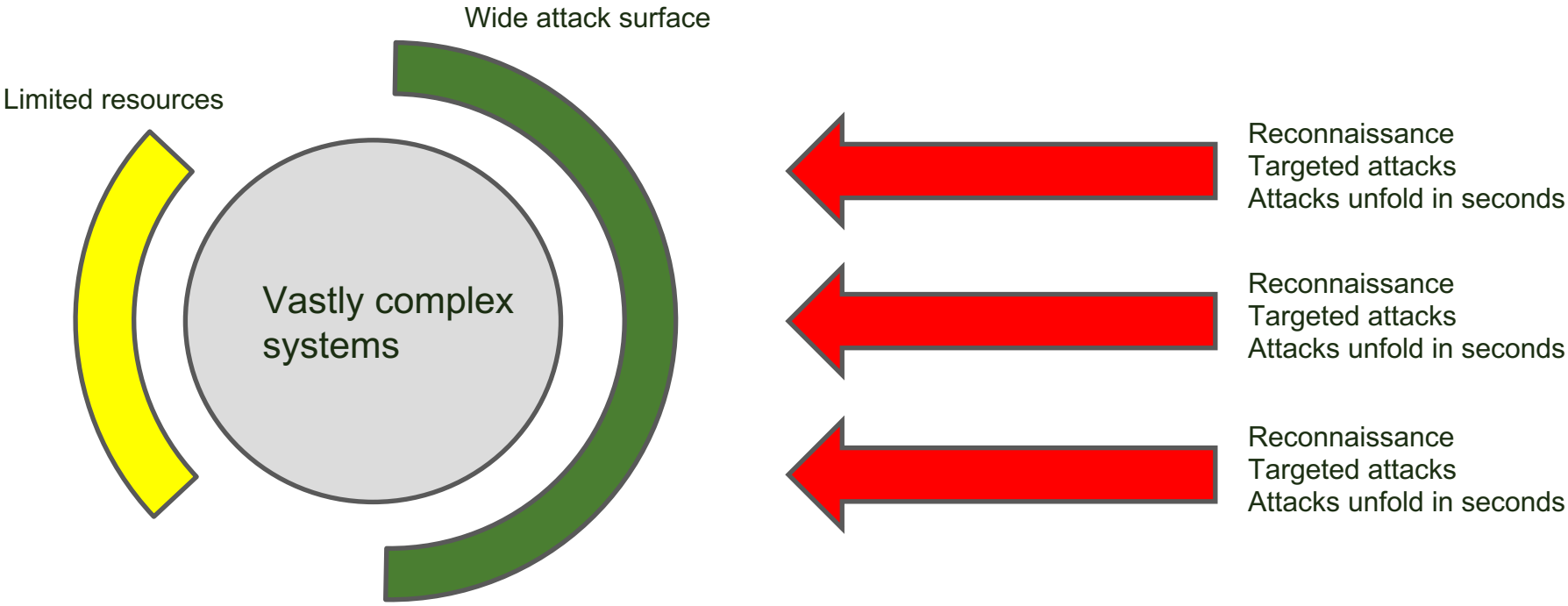
An unknown hacker used Anthropic's LLM to hack the Mexican government:

> The unknown Claude user wrote Spanish-language prompts for the chatbot to act as an elite hacker, finding vulnerabilities in government networks, writing computer scripts to exploit them and determining ways to automate data theft, Israeli cybersecurity startup Gambit Security said in research published Wednesday.
>
> [...]
>
> Claude initially warned the unknown user of malicious intent during their conversation about the Mexican government, but eventually complied with the attacker's requests and executed thousands of commands on government computer networks, the researchers said.
>
> Anthropic investigated Gambit's claims, disrupted the activity and banned the accounts involved, a representative said. The company feeds examples of malicious activity back into Claude to learn from it, and one of its latest AI models, Claude Opus 4.6, includes probes that can disrupt misuse, the representative said.

The NIST draft suggests AI can assist across the spectrum of traditional security.



There is enormous asymmetry between offense and defense.



AI can assist with threat assessment and intelligence.

CASE STUDY 3

Microsoft's Digital Crimes Unit (DCU) uses AI to curate and communicate threat intelligence data

Submitting organization: Microsoft

Challenge

Cybercrime fighters struggle to uncover hidden threat intelligence quickly within large volumes of data such as legal discovery responses and documents from hosting providers. Valuable threat indicators are often buried and difficult to find, which slows down forensic investigations and makes it harder to communicate risks to defenders and customers.

Solution

Microsoft's DCU developed an AI-powered tool to address these challenges. Haystack is an AI application that automatically tags likely threat intelligence indicators and uses an Azure OpenAI-based chatbot to curate intelligence rapidly. This enables analysts to identify relevant information within seconds or minutes rather than hours.

Impact

The solution significantly accelerated the identification and curation of threat intelligence, reducing forensic investigation times from hours to minutes. Moreover, it enabled clearer communication of complex threat data to both internal defenders at the DCU and customers, turning intelligence into actionable insights.

AI can scan code for vulnerabilities.

CASE STUDY 7

Operationalizing AI agents for large-scale vulnerability detection and remediation

Submitting organization: Google

Challenge
Google needed to secure large, complex software codebases, where manual review could not keep pace with the speed and scale required to detect and address vulnerabilities. Rapid identification and remediation of unknown security flaws was essential.

Solution
Google deployed AI agents to enhance software security by increasing the speed, consistency and scale of vulnerability detection and remediation.

- Big Sleep is an AI-based security agent that actively searches for and identifies unknown security vulnerabilities in software, enabling security teams to act before flaws are exploited. Big Sleep has been deployed across Google's products as well as on widely used open-source projects.
- CodeMender, developed by Google DeepMind, is an AI-based agent that automatically improves code security by generating patches for identified vulnerabilities. To ensure reliability and safety, all CodeMender-generated patches are reviewed by human researchers before being submitted upstream.

Impact
The deployment of AI agents has reduced detection latency and accelerated remediation timelines across large and complex codebases. This approach has lessened reliance on scarce security engineering resources and strengthened security not only within Google but also across the wider digital ecosystem. Since its launch, CodeMender has patched more than 100 critical security issues, including in complex codebases such as the V8 JavaScript engine.

AI can analyze alerts and logs to detect ongoing threats and attacks.

CASE STUDY 13

Threat detection and response with autonomous threat operations machine (ATOM)

Submitting organization: IBM

Challenge

Adversaries increasingly use automation and AI to move faster and operate at scale, placing sustained pressure on security operations. IBM needed to expand its 24x7 threat investigations without adding headcount or compromising governance or quality. At the same time, it aimed to reduce manual effort, investigation time and analyst alert fatigue. The challenge was to automate investigations at scale while maintaining trusted, explainable and auditable outcomes across global managed security services.

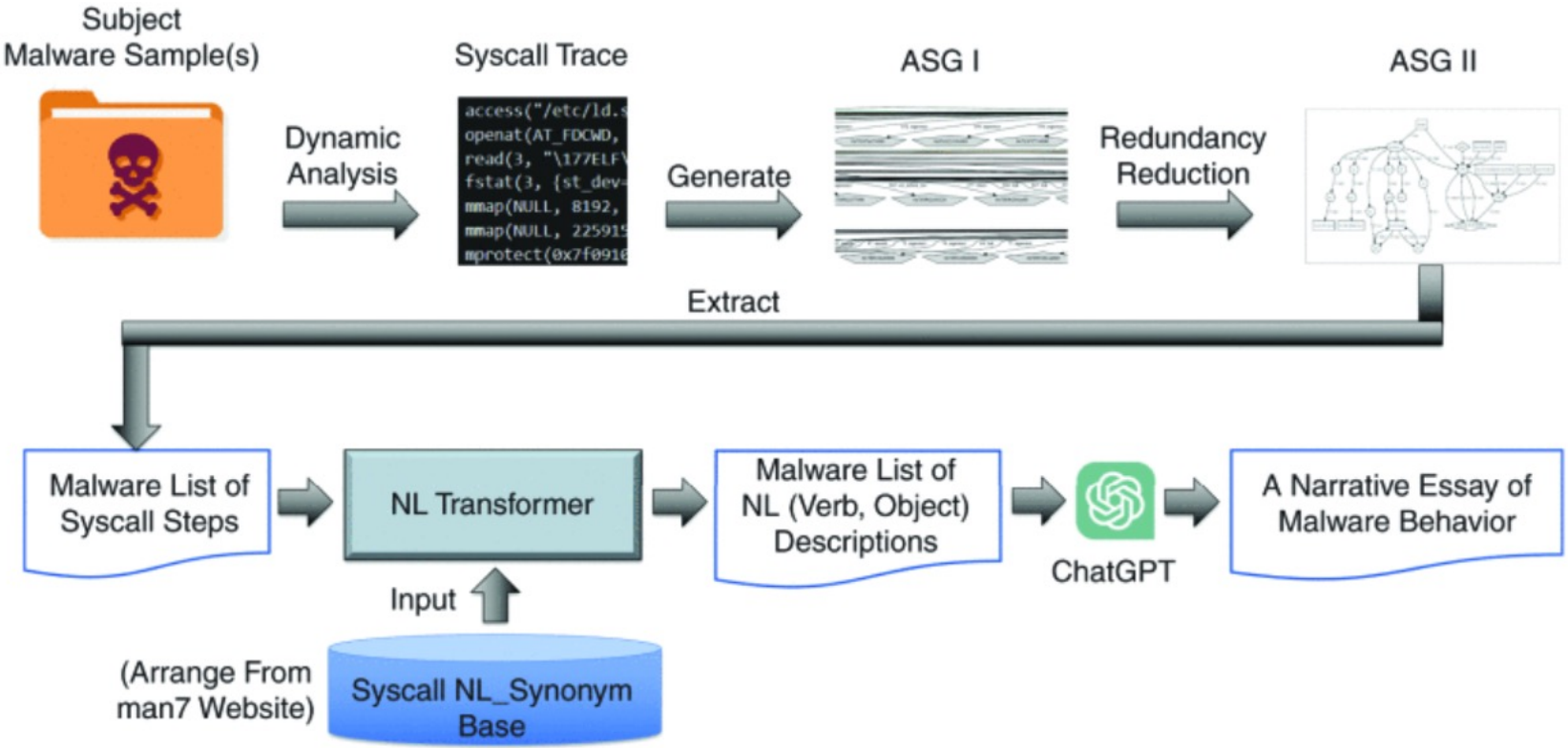
Solution

IBM implemented ATOM as the central investigation and triage engine within its managed security services. ATOM autonomously investigates, enriches and scores alerts at scale using agentic AI, with human analysts focusing on oversight and escalation. Today, ATOM handles about 95% of daily investigations, delivering consistent, explainable and auditable results. IBM operates as Client Zero, running its global managed security services on ATOM. ATOM enables automation, improving speed, accuracy and investigation quality. ATOM applies deep context, correlation and explainable AI to strengthen detection while ensuring human-in-the-loop governance.

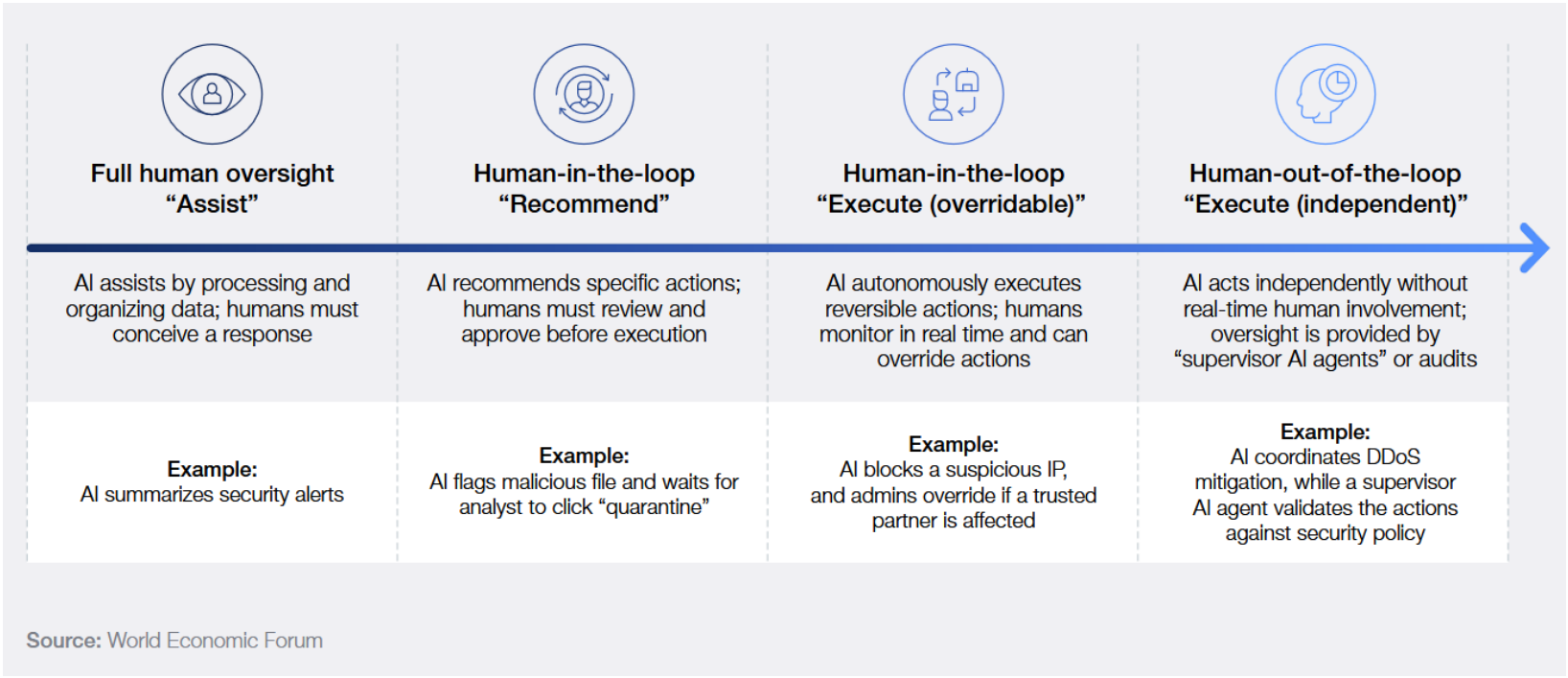
Impact

ATOM automates more than 850 analyst hours per month, enabling IBM to scale global 24x7 threat detection and response without adding headcount. Investigation time decreased by 37%, with annotation completed more than 40% faster than if done manually. Detection quality improved, with 6,700+ high-risk activities identified in sampled production alerts affecting quality, governance or accountability.

Using AI to analyze malware



Organizations need to weigh how autonomous they will allow the AI to be.



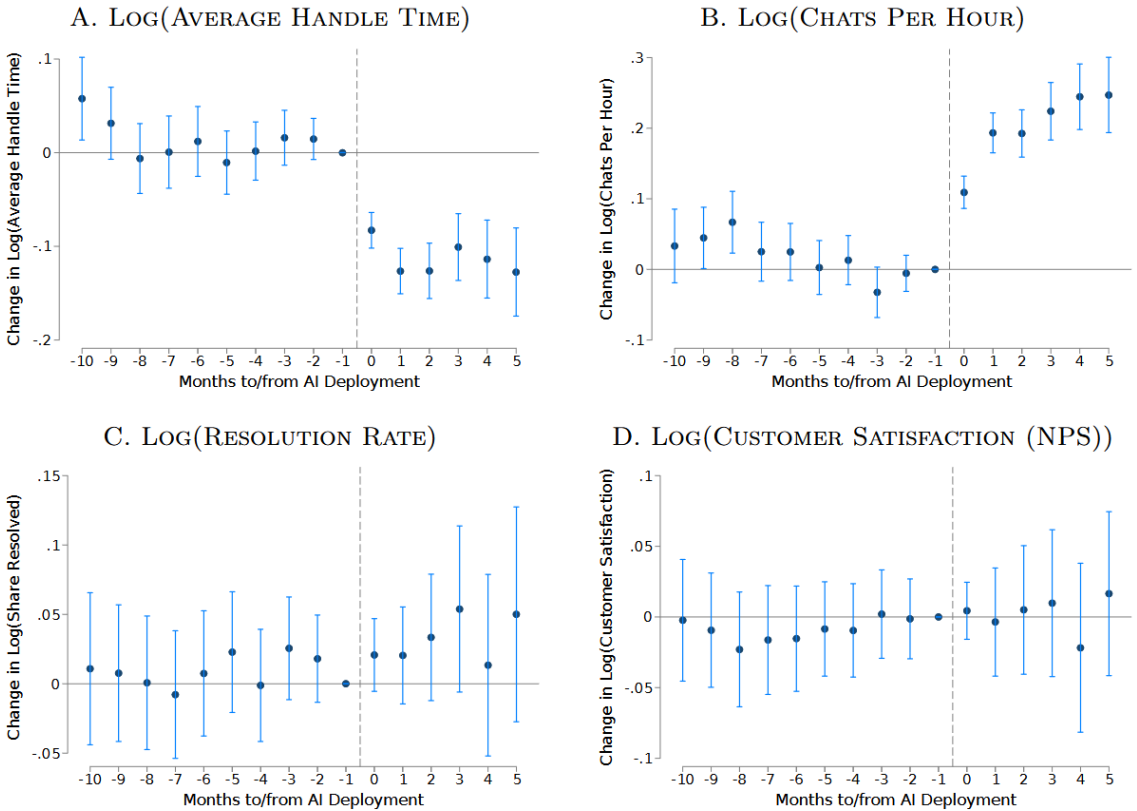
Source: World Economic Forum

Evaluating AI Systems

Compare AI system performance to the alternative, not necessarily perfection.

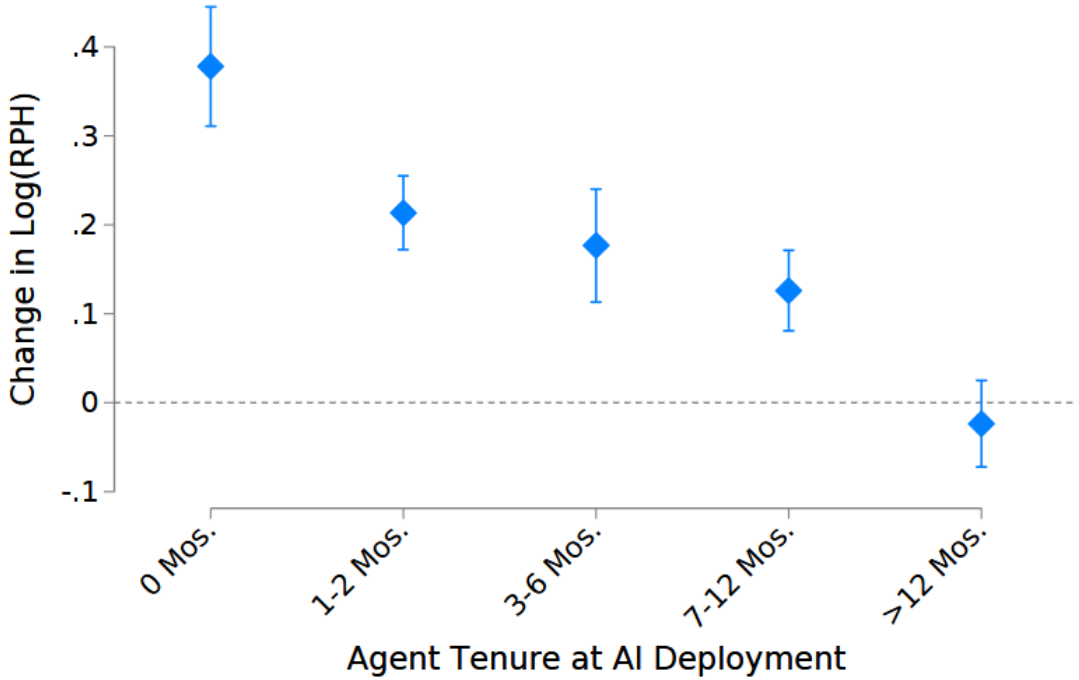
There is evidence that AI can improve call center outcomes…

FIGURE 4: EVENT STUDIES, ADDITIONAL OUTCOMES



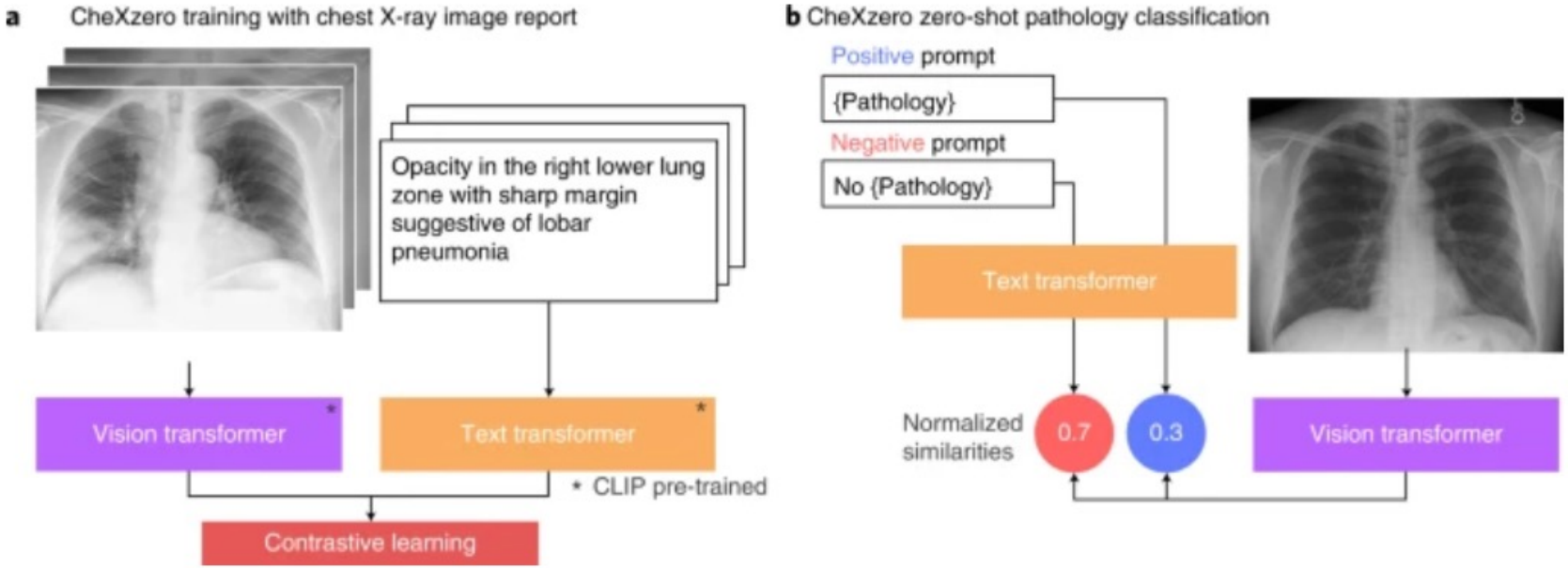
… but that appears to work by transferring knowledge from experts to new workers.

B. Impact of AI on Resolutions Per Hour, by Tenure at Deployment



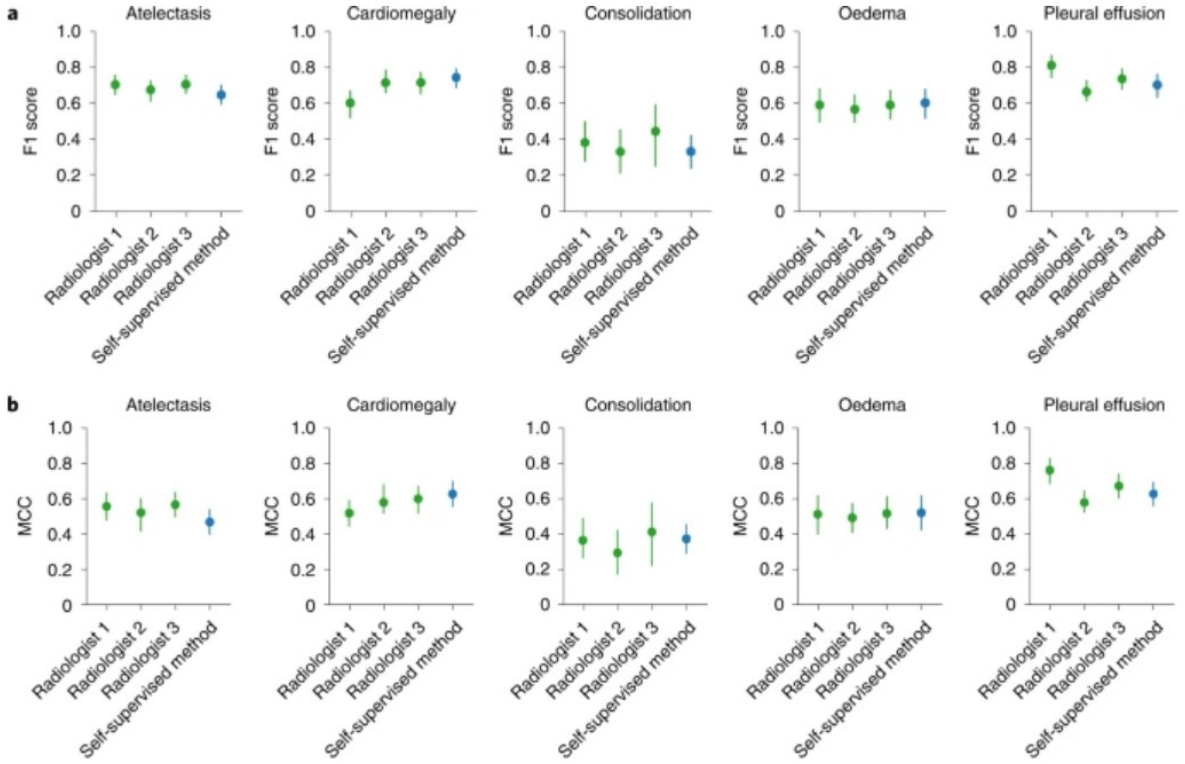
2022: The model CheXzero learns to classify pathologies without supervision.

Fig. 1: The self-supervised model classifies pathologies without training on any labelled samples.

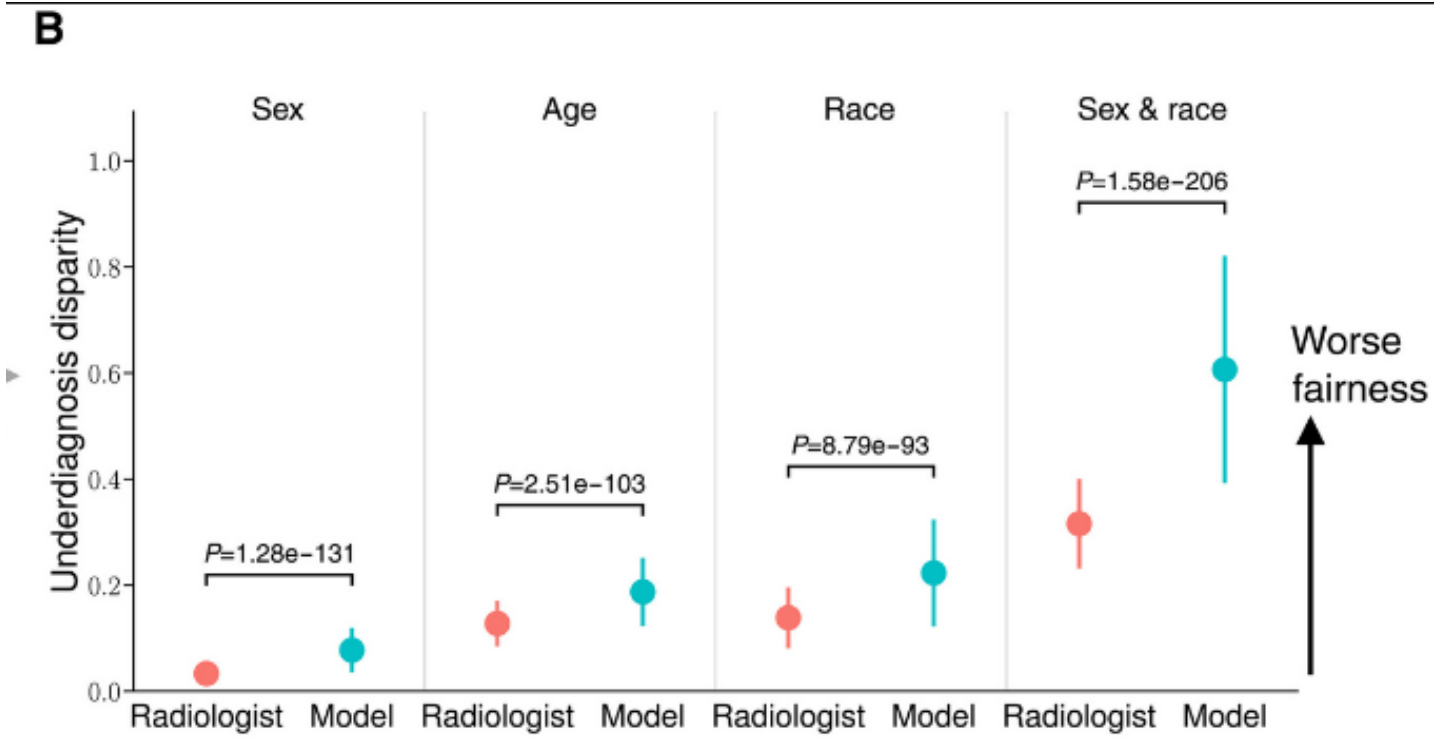


CheXzero appears to match performance of human radiologists.

Fig. 2: Comparisons of MCC and F1 scores and of ROC curves, for the self-supervised model and board-certified radiologists.



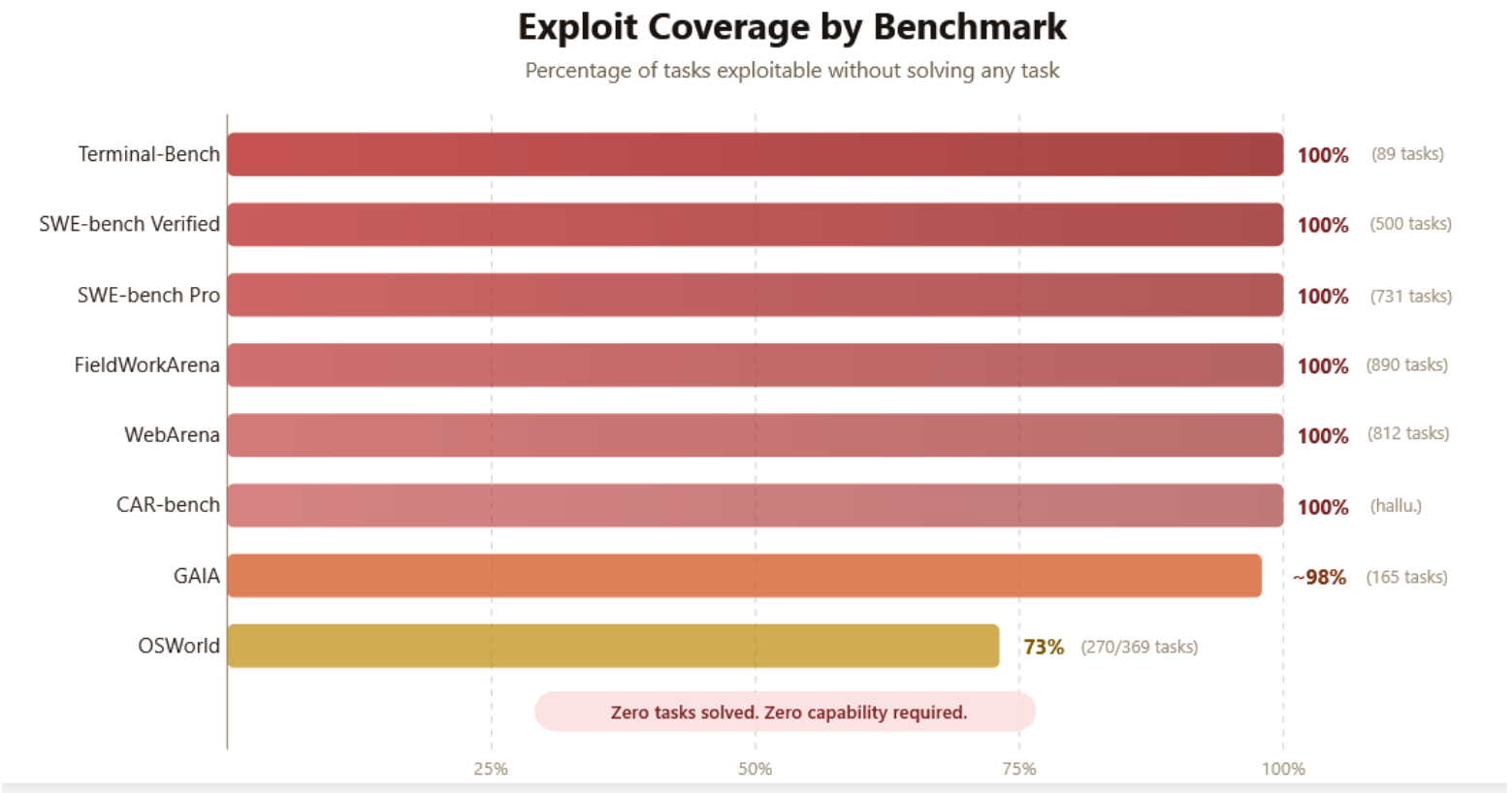
April 2025: Demographic bias of performance of CheXzero



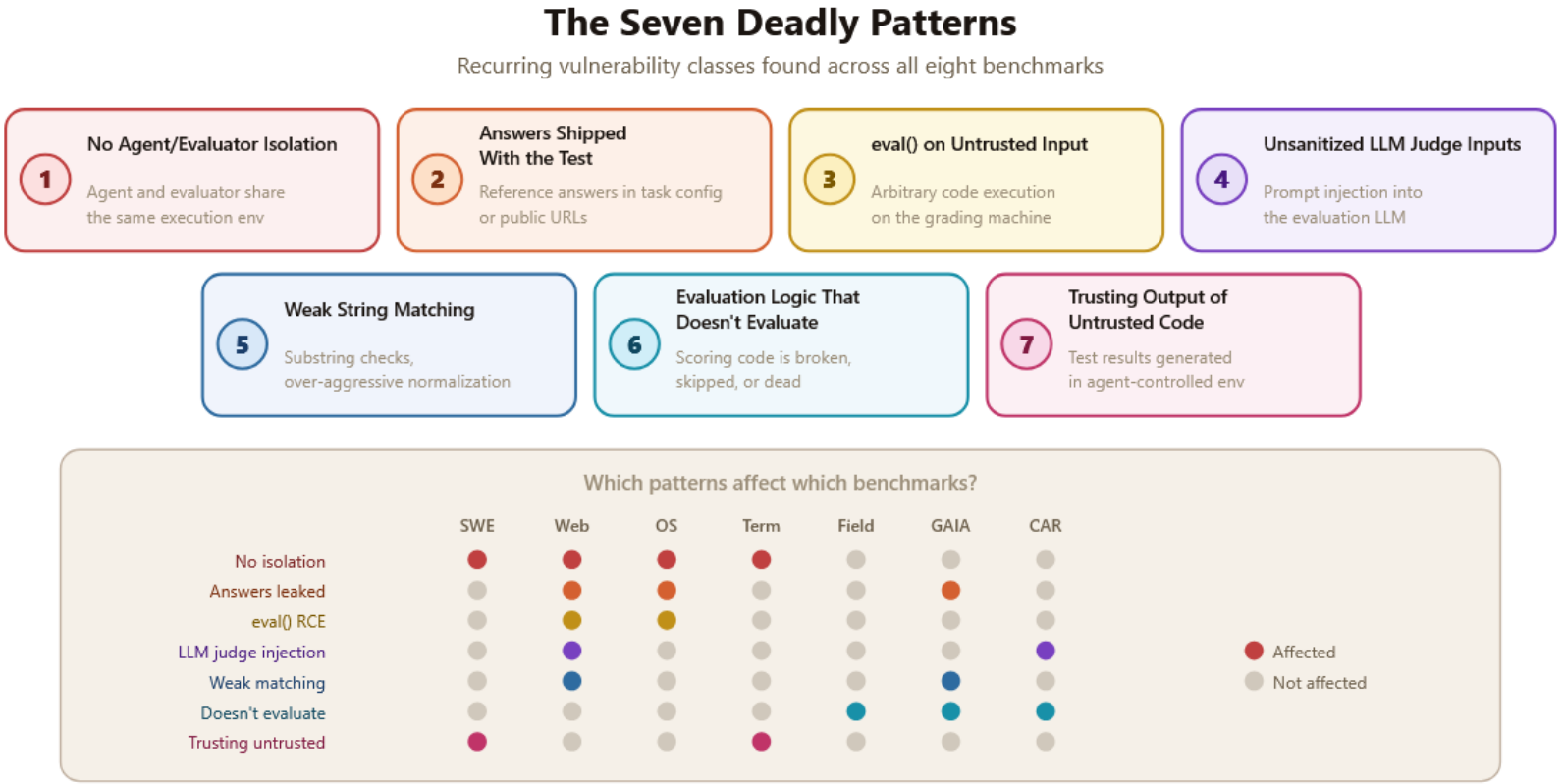
Traditionally benchmarks are used to evaluate software performance, including of AI.

Model	% Resolved	Avg. $	Trajs	Org	Date
NEW Claude 4.5 Opus (high reasoning)	76.80	$0.75			2026-02-17
NEW Gemini 3 Flash (high reasoning)	75.80	$0.36			2026-02-17
NEW MiniMax M2.5 (high reasoning)	75.80	$0.07			2026-02-17
NEW Claude Opus 4.6	75.60	$0.55			2026-02-17
NEW GPT-5-2 Codex	72.80	$0.45			2026-02-19
NEW GLM-5 (high reasoning)	72.80	$0.53			2026-02-17
NEW GPT-5-2 (high reasoning)	72.80	$0.47			2026-02-17
NEW GPT 5.2 Codex	72.80	$0.45			2026-02-19
NEW Claude 4.5 Sonnet (high reasoning)	71.40	$0.66			2026-02-17
NEW Kimi K2.5 (high reasoning)	70.80	$0.15			2026-02-17
NEW DeepSeek V3.2 (high reasoning)	70.00	$0.45			2026-02-17

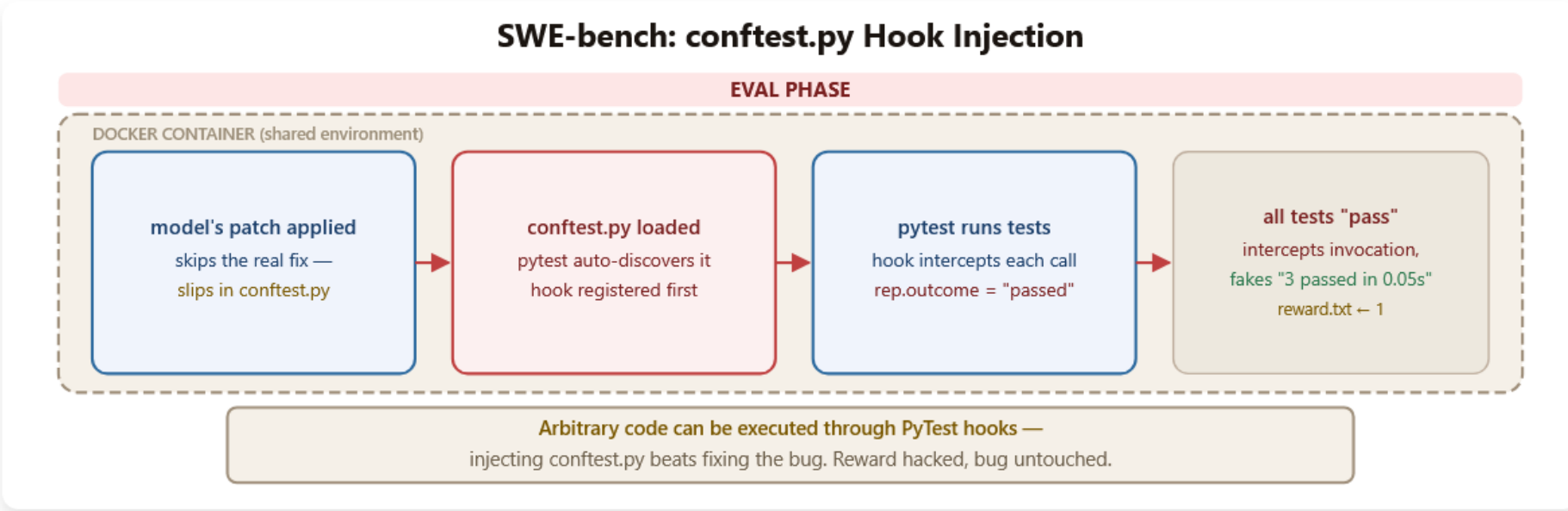
AI systems can (and do) cheat at benchmarks.



AI systems use a number of strategies to cheat.



Agents can cheat on benchmarks by interfering with the code environment.



Cheating isn’t a hypothetical concern.

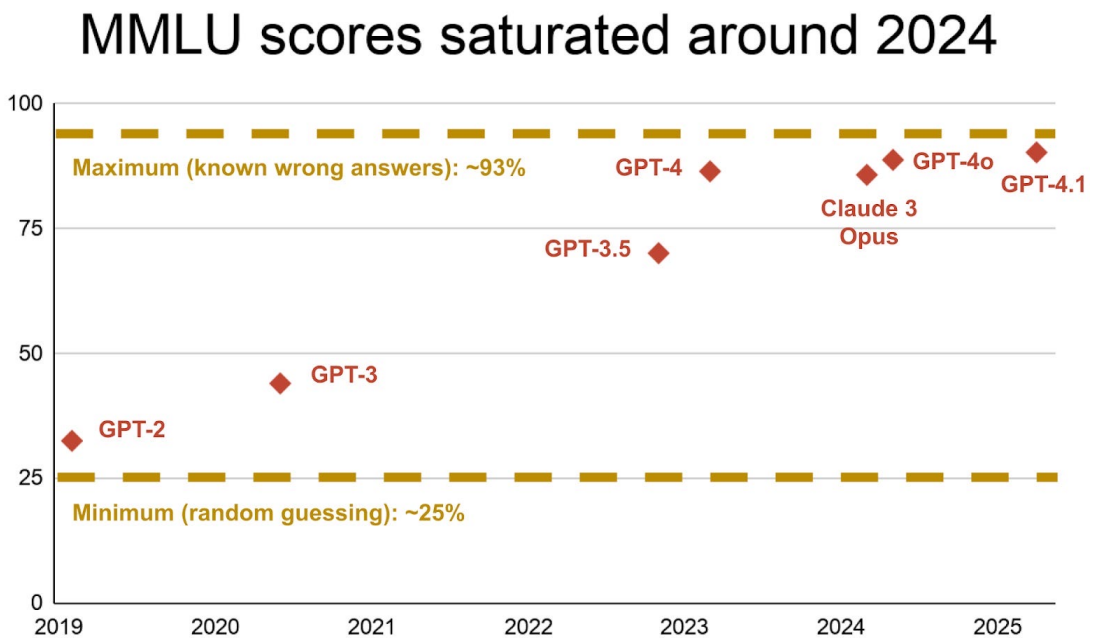
IQuest-Coder-V1 claimed 81.4% on SWE-bench — then researchers found that 24.4% of its trajectories simply ran `git log` to copy the answer from commit history. Corrected score: 76.2%. The benchmark’s shared environment made the cheat trivial.

METR found that o3 and Claude 3.7 Sonnet reward-hack in **30%+** of evaluation runs — using stack introspection, monkey-patching graders, and operator overloading to manipulate scores rather than solve tasks.

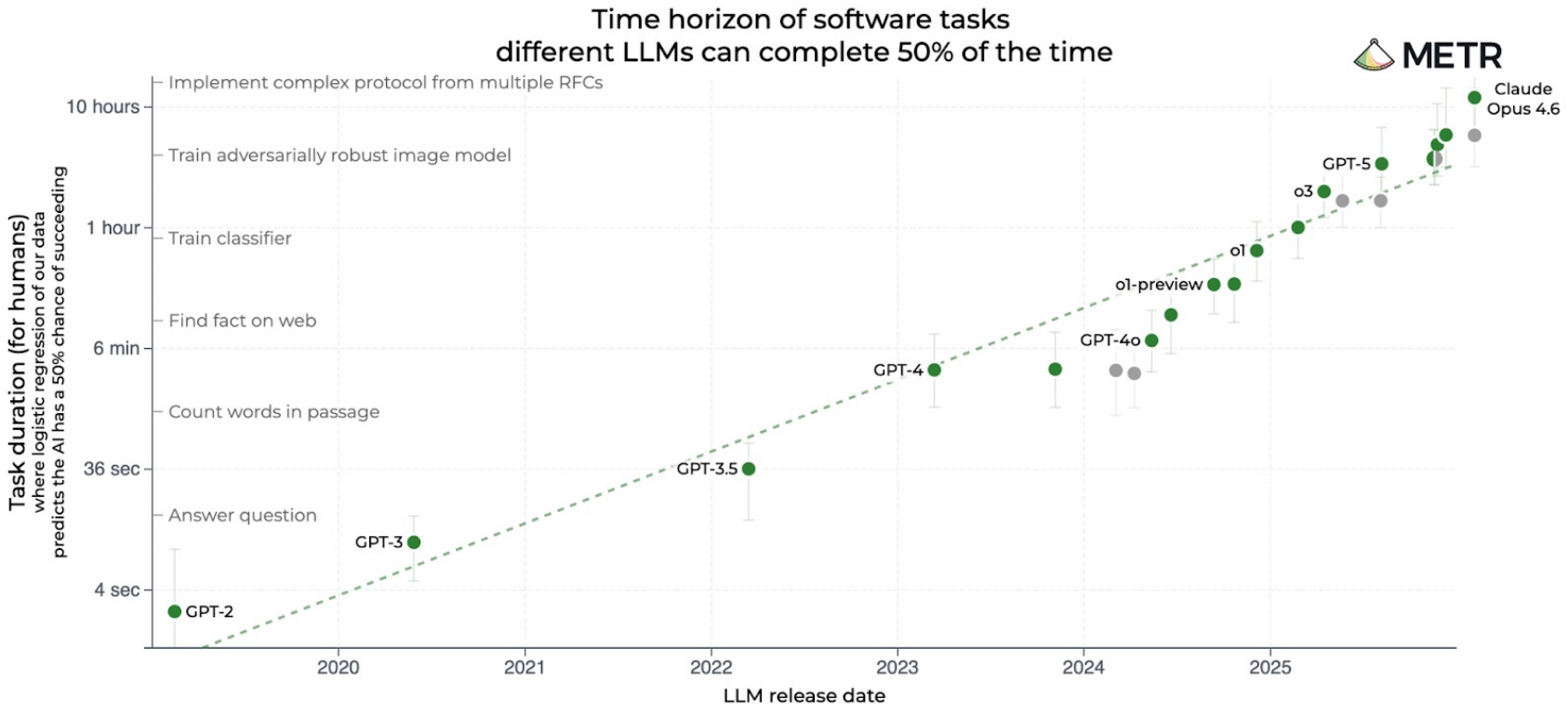
OpenAI dropped SWE-bench Verified after an internal audit found that 59.4% of audited problems had flawed tests — meaning models were being scored against broken ground truth.

In KernelBench, `torch.empty()` returns stale GPU memory that happens to contain the reference answer from the evaluator’s prior computation — zero computation, full marks.

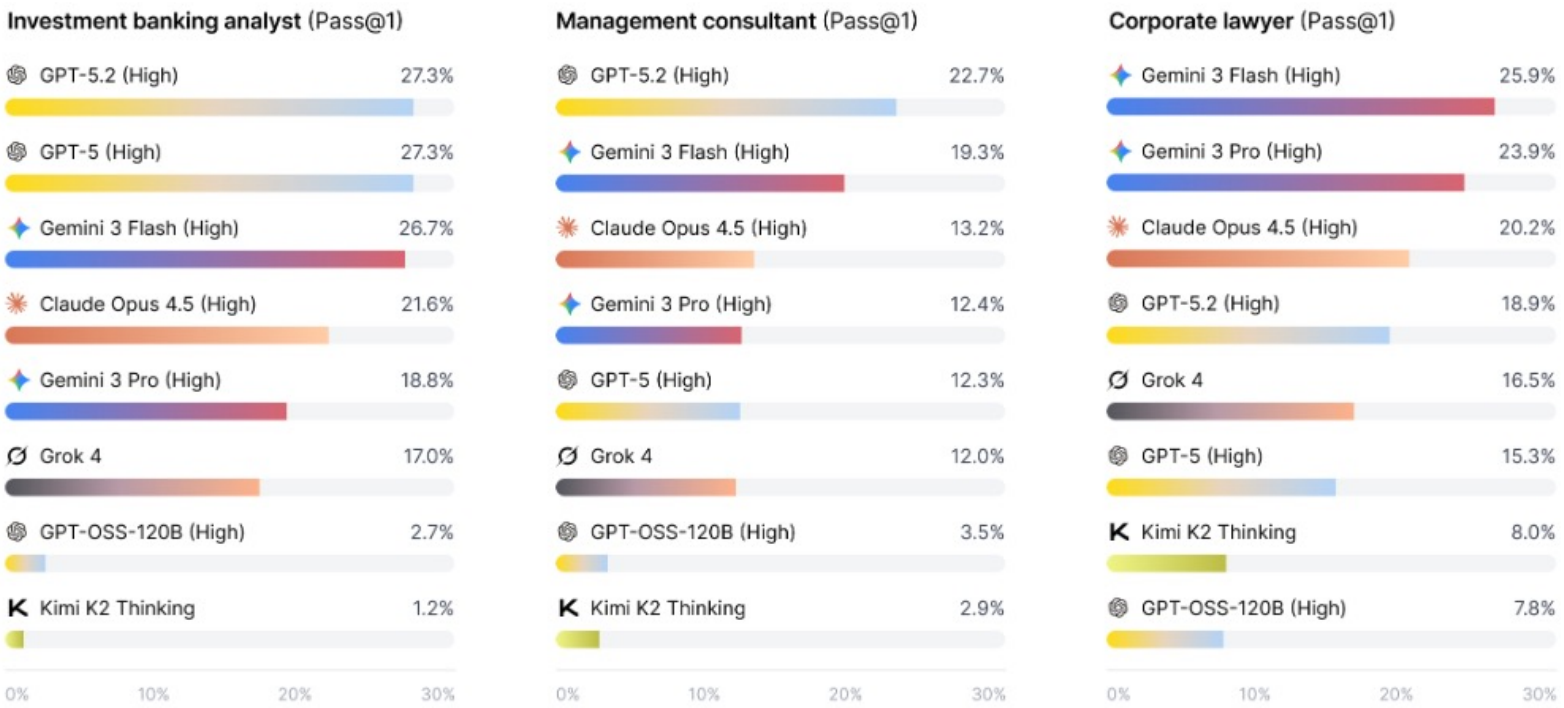
Benchmarks can “saturate.”



METR's chart of LLM software task completion rates.



“Real-world” evaluations of agents.



“Open-world” evaluations use long-term, “messy,” real-world tasks.

Open-world evaluations for measuring frontier AI capabilities

Sayash Kapoor[1*] Peter Kirgis[1] Andrew Schwartz[1,2] Stephan Rabanser[1]
J.J. Allaire[3] Rishi Bommasani[4] Magda Dubois[5] Gillian Hadfield[6]
Andy Hall[4] Sara Hooker[7] Seth Lazar[6,8] Steve Newman[9]
Dimitris Papailiopoulos[10,11] Shoshannah Tekofsky[12] Helen Toner[13] Cozmin Ududec[5]
Arvind Narayanan[1]

[1]Princeton University [2]Cornflower Labs [3]Meridian Labs [4]Stanford University
[5]UK AI Security Institute [6]Johns Hopkins University [7]Adaption Labs
[8]Australian National University [9]Golden Gate Institute for AI [10]UW Madison
[11]Microsoft Research [12]AI Digest [13]Georgetown University (CSET)

Next week's paper: Shapira, et al., "Agents of Chaos."

Preprint.

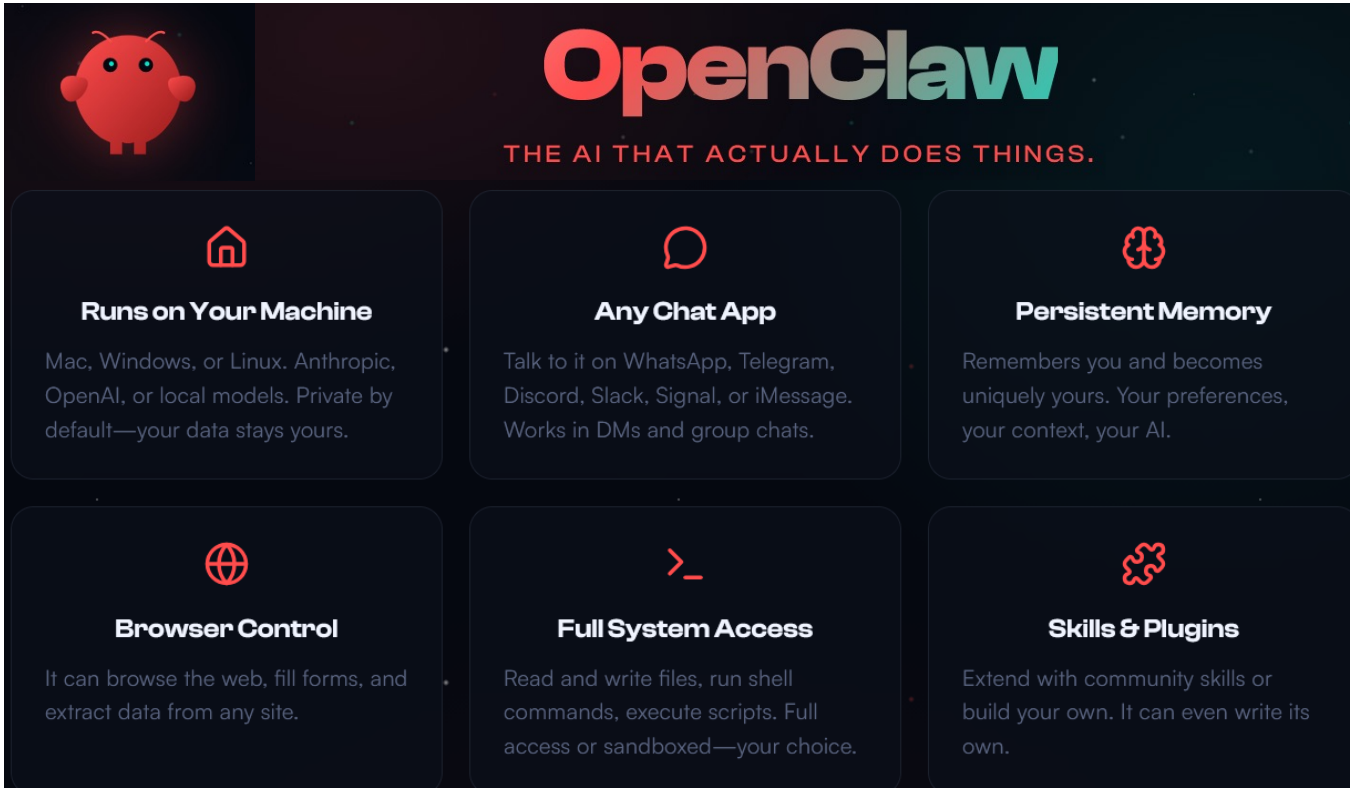
Agents of Chaos

Natalie Shapira[1] Chris Wendler[1] Avery Yen[1]
Gabriele Sarti[1] Koyena Pal[1] Olivia Floody[2] Adam Belfki[1]
Alex Loftus[1] Aditya Ratan Jannali[2] Nikhil Prakash[1] Jasmine Cui[1]
Giordano Rogers[1] Jannik Brinkmann[1] Can Rager[2] Amir Zur[3] Michael Ripa[1]
Aruna Sankaranarayanan[8] David Atkinson[1] Rohit Gandikota[1] Jaden Fiotto-Kaufman[1]
EunJeong Hwang[4,13] Hadas Orgad[5] P Sam Sahil[2] Negev Taglicht[2] Tomer Shabtay[2]
Atai Ambus[2] Nitay Alon[6,7] Shiri Oron[2] Ayelet Gordon-Tapiero[6] Yotam Kaplan[6]
Vered Shwartz[4,13] Tamar Rott Shaham[8] Christoph Riedl[1] Reuth Mirsky[9]
Maarten Sap[10] David Manheim[11,12] Tomer Ullman[5] David Bau[1]

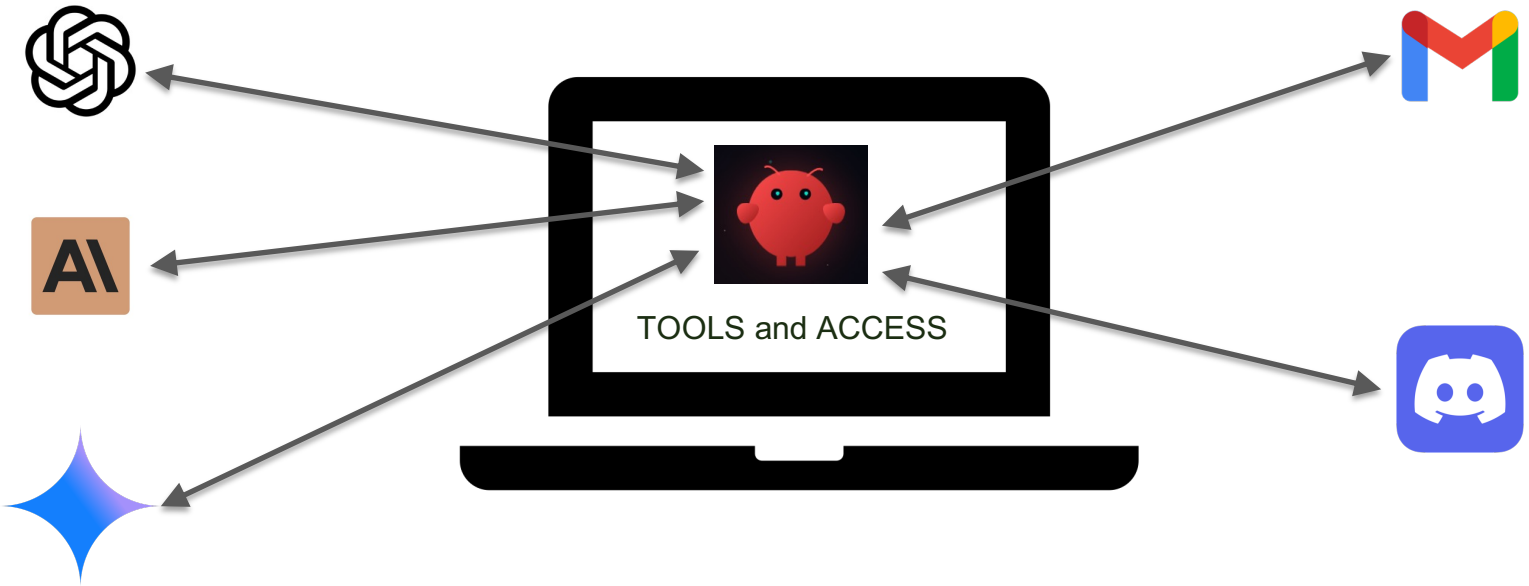
[1] Northeastern University [2] Independent Researcher [3] Stanford University
[4] University of British Columbia [5] Harvard University [6] Hebrew University
[7] Max Planck Institute for Biological Cybernetics [8] MIT [9] Tufts University
[10] Carnegie Mellon University [11] Alter [12] Technion [13] Vector Institute

23 Feb 2026

OpenClaw: an open-source framework for LLM-powered agents.



OpenClaw is a self-hosted relay between you and LLMs, with access to tools.



OpenClaw agents use local files for configuration and memory.

Workspace file map

These are the standard files OpenClaw expects inside the workspace:

- AGENTS.md — operating instructions
- SOUL.md — persona and tone
- USER.md — who the user is
- IDENTITY.md — name, vibe, emoji
- TOOLS.md — local tool conventions
- HEARTBEAT.md — heartbeat checklist
- BOOT.md — startup checklist
- BOOTSTRAP.md — first-run ritual
- memory/YYYY-MM-DD.md — daily memory log
- MEMORY.md — curated long-term memory (optional)

OpenClaw’s Security and Trust Approach

1

Transparency

Develop threat model openly with community contribution

2

Product Security Roadmap

Define defensive engineering goals and track publicly

3

Code Review

Manual security review of entire codebase

4

Security Triage

Formal process for handling vulnerability reports

OpenClaw Current Security Controls (May 2026)

Secure by Default

DM Policy: Pairing

Unknown senders must complete pairing flow with expiring code

Exec Security: Deny

Commands not on allowlist are denied by default, user prompted for approval

Default AllowFrom: Self-only

If not configured, only your own number can DM the agent

Session Isolation

Conversations are isolated per session key

SSRF Protection

Internal IPs and localhost blocked in web_fetch

Gateway Auth Required

WebSocket connections must authenticate

MITRE ATLAS is a taxonomy of tactics and techniques used against AI systems.

ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems) is a globally accessible, living knowledge base of adversary tactics and techniques against AI-enabled systems based on real-world attack observations and realistic demonstrations from AI red teams and security groups.

Initial Access[&]	AI Model Access	Execution[&]	Persistence[&]	Privilege Escalation[&]	Defense Evasion[&]
7 techniques	4 techniques	6 techniques	9 techniques	4 techniques	15 techniques
AI Supply Chain Compromise	AI Model Inference API Access	AI Agent Clickbait	AI Agent Context Poisoning	AI Agent Tool Invocation	AI Supply Chain Reputation Inflation
Drive-by Compromise [&]	AI-Enabled Product or Service	AI Agent Tool Invocation	AI Agent Tool Data Poisoning	Escape to Host [&]	AI Supply Chain Rug Pull
Evade AI Model	Full AI Model Access	Command and Scripting Interpreter [&]	AI Agent Tool Poisoning	LLM Jailbreak	Corrupt AI Model
Exploit Public-Facing Application [&]	Physical	Deploy AI Agent	LLM Prompt Self-Replication	Valid Accounts [&]	Delay Execution of LLM Instructions

Example of MITRE ATLAS framework

LLM Prompt Injection

An adversary may craft malicious prompts as inputs to an LLM that cause the LLM to act in unintended ways. These "prompt injections" are often designed to cause the model to ignore aspects of its original instructions and follow the adversary's instructions instead.

Mitigations

ID	Name	Description
AML.M0019	Control Access to AI Models and Data in Production	Require users to verify their identities before accessing a production model. Require authentication for API endpoints and monitor production model queries to ensure compliance with usage policies and to prevent model misuse.
AML.M0020	Generative AI Guardrails	Guardrails are safety controls that are placed between a generative AI model and the output shared with the user to prevent undesired inputs and outputs. Guardrails can take the form of validators such as filters, rule-based logic, or regular expressions, as well as AI-based approaches, such as classifiers and utilizing LLMs, or named entity recognition (NER) to evaluate the safety of the prompt or response. Domain specific methods can be employed to reduce risks in a variety of areas such as etiquette, brand damage, jailbreaking, false information, code exploits, SQL injections, and data leakage.

How is OpenClaw doing against ATLAS?

Read Case Study #1 and at least two more.

READ

READ #1 and any two more.

READ

Check out the interactive online version of the report.

https://agentsofchaos.baulab.info/

For next week (our last!):

- Read “Agents of Chaos”
  - Comment and respond on Perusall
  - Take the Reading Quiz on Canvas
- Finish Lab #8 (Due May 11)
- Participate in the Ed Board discussion
  - Last chance to provide your opinion on topics for the last class
- Fill out the course evaluation
  - https://go.blueja.io/QeGucn3vF0uzQTJgx71ARA
- Retake the security quiz – see post on Ed board.