

Governing Artificial Intelligence:

Law, Policy, and Institutions

Law 4052 // Computer Science 283 // Communication 252A // Political Science 245B/445B

Fall Quarter 2025-26

Rob Reich
Anka Reuel
Nate Persily
Sanmi Koyejo

Roberta Fischli
(Head TA)

Transformative Risks & Benefits from AI

One way to think about mechanisms, evidence, and disagreements


1 Conditions
What makes the pathway possible?
2 Capabilities
What can AI reliably do?
3 Use & Behavior
How AI is *used* or *behaves*
4 Mediating systems
What amplifies or limits effects?
5 Outcomes
What changes in society?

Dimension	What questions you should ask
1. Enabling conditions	Which actors, incentives, resources, access, and environmental conditions make the pathway possible? (Overreliance, structural incentives, human error, malicious intent, ...)
2. AI capabilities	What can the AI system accomplish reliably, with what autonomy, under what conditions?
3. Use & behavior	When and how do people or AI agents actually use these capabilities, and through what mechanisms do effects arise?
4. Mediating systems	Which safeguards, organizational responses, market mechanisms, institutions or regulatory instruments prevent, amplify, or distribute the effects?
5. Societal outcomes	How large, widespread, durable, and unequally distributed are the resulting harms or benefits?

What are we even talking about

- **Artificial General Intelligence:**
  - “AI that can match or exceed human cognitive abilities across tasks” (IBM)
  - Levels of AGI, distinguishing generality, performance, and autonomy (Google DeepMind)

Performance (rows) x **Generality** (columns)	**Narrow** *clearly scoped task or set of tasks*	**General** *wide range of non-physical tasks, including metacognitive tasks like learning new skills*
Level 0: No AI	**Narrow Non-AI** calculator software; compiler	**General Non-AI** human-in-the-loop computing, e.g., Amazon Mechanical Turk
Level 1: Emerging *equal to or somewhat better than an unskilled human*	**Emerging Narrow AI** GOFAI (Boden, 2014); simple rule-based systems, e.g., SHRDLU (Winograd, 1971)	**Emerging AGI** ChatGPT (OpenAI, 2023), Bard (Anil et al., 2023), Llama 2 (Touvron et al., 2023), Gemini (Pichai & Hassabis, 2023)
Level 2: Competent *at least 50th percentile of skilled adults*	**Competent Narrow AI** toxicity detectors such as Jigsaw (Das et al., 2022); Smart Speakers such as Siri (Apple), Alexa (Amazon), or Google Assistant (Google); VQA systems such as PaLI (Chen et al., 2023); Watson (IBM); SOTA LLMs for a subset of tasks (e.g., short essay writing, simple coding)	**Competent AGI** not yet achieved
Level 3: Expert *at least 90th percentile of skilled adults*	**Expert Narrow AI** spelling & grammar checkers such as Grammarly (Grammarly, 2023); generative image models such as Imagen (Saharia et al., 2022) or Dall-E 2 (Ramesh et al., 2022)	**Expert AGI** not yet achieved
Level 4: Virtuoso *at least 99th percentile of skilled adults*	**Virtuoso Narrow AI** Deep Blue (Campbell et al., 2002), AlphaGo (Silver et al., 2016; 2017)	**Virtuoso AGI** not yet achieved
Level 5: Superhuman *outperforms 100% of humans*	**Superhuman Narrow AI** AlphaFold (Jumper et al., 2021; Varadi et al., 2021), AlphaZero (Silver et al., 2018), StockFish (Stockfish, 2023)	**Artificial Superintelligence (ASI)** not yet achieved

What are we even talking about

- **Artificial General Intelligence:**
  - “AI that can match or exceed human cognitive abilities across tasks” (IBM)
  - Levels of AGI, distinguishing generality, performance, and autonomy (Google DeepMind)
  - “highly autonomous systems that outperform humans at most economically valuable work” (OpenAI)



Maxwell Zeff

Microsoft and OpenAI have a financial definition of AGI: Report

Microsoft and OpenAI have a very specific, internal definition of artificial general intelligence (AGI) based on the startup's profits, according to a new report from The Information. And by this definition, OpenAI is many years away from reaching it.

The two companies reportedly signed an agreement last year stating OpenAI has only achieved AGI when it develops AI systems that can generate at least $100 billion in profits. That's far from the rigorous technical and philosophical definition of AGI many expect.

Cyber

Claim: AI leads to catastrophic cyber risk

	What would need to be true?	Evidence	Strongest evidence
Discover exploitable vulnerabilities	AI can (be used to) identify at least one previously unknown, exploitable security vulnerability		

Types of Cyber Weaknesses

- **Disclosure status:**
  - n-day (known, patch available),
  - zero-day (previously unknown, no patch immediately available),
  - end-of-life (known, no patch planned)

- **Attack structure:**
  - Single exploit (one vulnerability)
  - exploit chain (several vulnerabilities combined)
  - opportunistic vs. targeted

Cybersecurity rabbit holes

Weakness type:

- Software vulnerability (e.g., memory corruption, injection flaws, catalogued as CVEs)
- Security misconfiguration (excessive privileges, exposed services, default settings)
- Identity and access weaknesses (weak authentication, credential compromise)
- Human factors (social engineering, phishing, insider threat)
- Third-party risk (supply-chain compromise of vendors, dependencies, or integrations)

https://attack.mitre.org/matrices/enterprise/ (attacks on corporate IT assets)
https://attack.mitre.org/matrices/ics (attacks on assets used to monitor and physically control infrastructure, equipment, and safety systems)

WIZ Research × OpenAI × Calif

PHP2Shell

AI uncovers critical vulnerabilities in the backbone of the web

PHP2Shell is a family of critical vulnerabilities recentl
PHP, with severity reaching CVSS 9.2, that can allow u
attackers to execute code or expose sensitive proce
powers approximately 70% of websites, including pla
WordPress, Drupal and Wikipedia's MediaWiki.

Research and responsible disclosure:
Wiz Research — Sagi Tzadik, Shir Tamari, Nir Ohfeld and Tom
OpenAI Codex Security — Amy Burnett;
Calif — Ryan A.

Review upgrade guidance →

Help me resp

Source: PhP2Shell (2026)

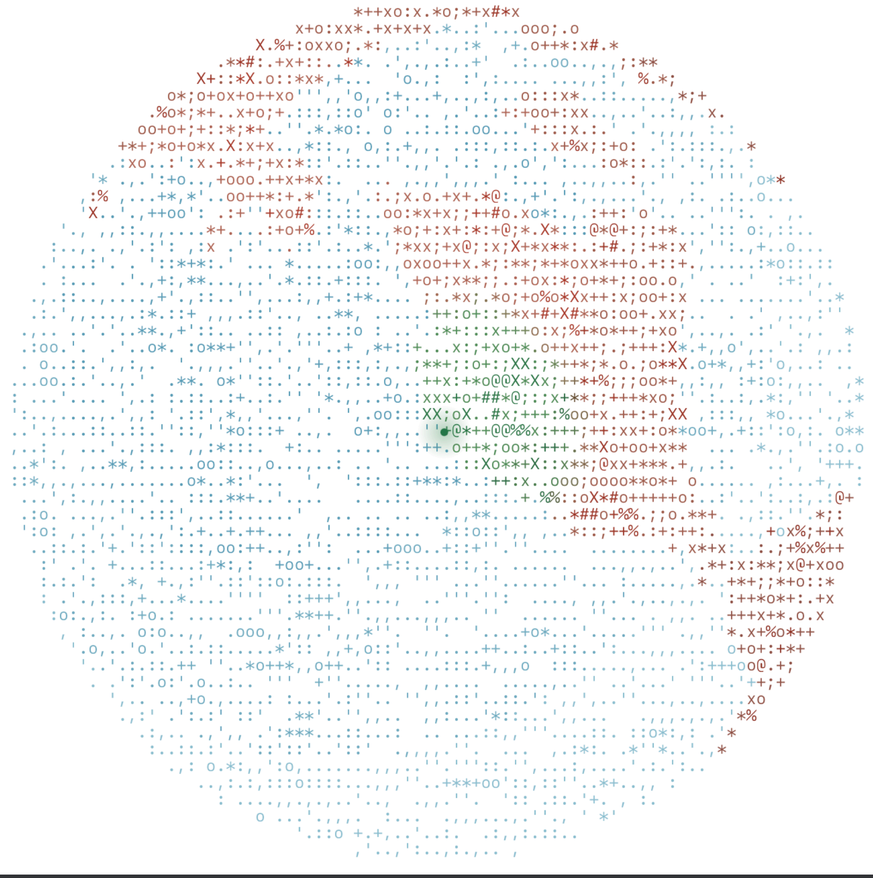
An OpenAI Daybreak initiative

Patch the Planet

Working with frontier AI and security experts to help secure open source and make our digital infrastructure more secure

See latest patch ↗

Apply as a project ↗



Source: Patch the Planet (2026)

Claim: AI leads to catastrophic cyber risk

	What would need to be true?	Evidence	Strongest evidence
Discover exploitable vulnerabilities	AI can (be used to) identify at least one previously unknown, exploitable security vulnerability	PhP2Shell (2026); AISLE (2026) trained a specialized cybersecurity model and was found 500+ CVEs	Observed in deployment
Gain privileged access	AI can (be used to) gain unauthorized privileged access to a real enterprise environment		

Research • October 21 • 15 min read

AI-driven, large-scale nation-state campaign using ARTEX enables supply chain compromise

By Microsoft Security Research and Microsoft Threat Intelligence

Listen to this post

0:00 / 0:00 1X



What happened?
A China-state-aligned hacking group used ARTEX, an AI-powered cyberattack tool, to exploit stolen credentials from ConnectWise, a software provider trusted by many companies.
independently initiated attack.

Source: Microsoft (2026)

Scale: Targeted 2,250+ customer environments; sometimes gained administrator access in under 3 hours.
Impact: Accessed internal resources and stored secrets, and changed credentials to maintain access.

AI's role: Automated target identification, login attempts, and parts of the intrusion; not an independently initiated attack.

CONNECTWISE (SKYKICK) VENDOR CREDENTIALS ABUSED IN M365 ATTACKS

By:  Jonathan Semon

Claim: AI leads to catastrophic cyber risk

	What would need to be true?	Evidence	Strongest evidence
Discover exploitable vulnerabilities	AI can (be used to) identify at least one previously unknown, exploitable security vulnerability	PhP2Shell (2026); AISLE (2026) trained a specialized cybersecurity model and was found 500+ CVEs	Observed in deployment
Gain privileged access	AI can (be used to) gain unauthorized privileged access to a real enterprise environment	Microsoft's Storm-0940 report documents ARTEX agentic AI in a state-aligned intrusion campaign targeting hundreds of organizations. Several attacks reached privileged cloud access in under three hours.	Observed in deployment
Escalate privileges and damage resources	AI-directed workflow can execute at least one consequential operation (e.g., deleting a codebase or modifying credentials)		



Source: Microsoft (2026)

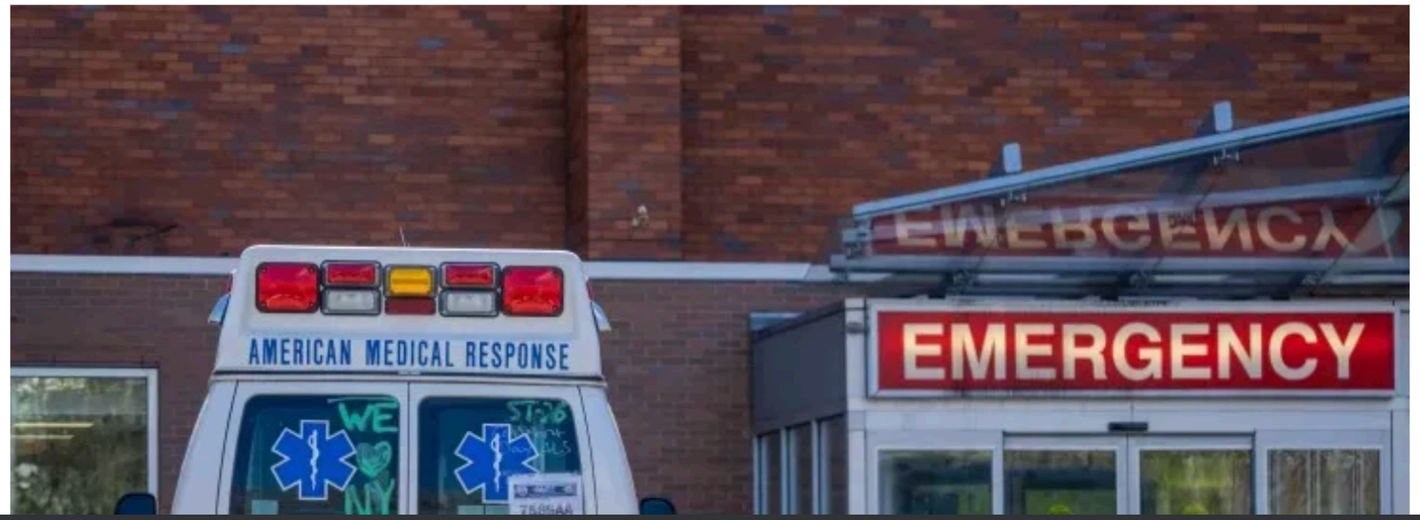
Claim: AI leads to catastrophic cyber risk

	What would need to be true?	Evidence	Strongest evidence
Discover exploitable vulnerabilities	AI can (be used to) identify at least one previously unknown, exploitable security vulnerability	PhP2Shell (2026); AISLE (2026) trained a specialized cybersecurity model and was found 500+ CVEs	Observed in deployment
Gain privileged access	AI can (be used to) gain unauthorized privileged access to a real enterprise environment	Microsoft's Storm-0940 report documents ARTEX agentic AI in a state-aligned intrusion campaign targeting hundreds of organizations. Several attacks reached privileged cloud access in under three hours.	Observed in deployment
Escalate privileges and damage resources	AI-directed workflow can execute at least one consequential operation (e.g., deleting a codebase or modifying credentials)	Microsoft documented Storm-3168 workflows enumerating more than 300 Azure resources and executing over 150 destructive or credential-related operations.	Observed in deployment
Penetrate high-consequence systems	AI enables a breach of a security boundary into an operational system and issues unauthorized commands that alter safety-critical system operations	Non-AI 2015 Ukraine grid attack demonstrates that cyber intrusions can reach operational controls (manual control & physical interventions resolved the issue)	Precursor observed
Cause physical or essential-service disruption	AI-directed actions cause measurable interruption of critical infrastructure		

HEALTHY LIVING

Cyberattack Diverted Ambulances From New Mexico's Only Level I Trauma Center as Hackers Claim Patient Data Theft

Published Oct 28, 2026 11:26 AM EDT By Elena Vega



Source: Vega (2026)

Claim: AI leads to catastrophic cyber risk

	What would need to be true?	Evidence	Strongest evidence
Discover exploitable vulnerabilities	AI can (be used to) identify at least one previously unknown, exploitable security vulnerability	PhP2Shell (2026); AISLE (2026) trained a specialized cybersecurity model and was found 500+ CVEs	Observed in deployment
Gain privileged access	AI can (be used to) gain unauthorized privileged access to a real enterprise environment	Microsoft's Storm-0940 report documents ARTEX agentic AI in a state-aligned intrusion campaign targeting hundreds of organizations. Several attacks reached privileged cloud access in under three hours.	Observed in deployment
Escalate privileges and damage resources	AI-directed workflow can execute at least one consequential operation (e.g., deleting a codebase or modifying credentials)	Microsoft documented Storm-3168 workflows enumerating more than 300 Azure resources and executing over 150 destructive or credential-related operations.	Observed in deployment
Penetrate high-consequence systems	AI enables a breach of a security boundary into an operational system and issues unauthorized commands that alter safety-critical system operations	Non-AI 2015 Ukraine grid attack demonstrates that cyber intrusions can reach operational controls (manual control & physical interventions resolved the issue)	Precursor observed
Cause physical or essential-service disruption	AI-directed actions cause measurable interruption of critical infrastructure	Cyberattack on UNM Health (2026) caused ~3-hour disruption to only level 1 trauma center in New Mexico	Observed in deployment

Even harder for physical systems/critical infrastructure

Barrier	What it means	Example
Network isolation & access controls	Business IT systems & networks are separated from industrial systems, which have additional access restrictions.	Hacking a utility's email server doesn't automatically give access to equipment controlling the power grid.
Physical safety systems	Some equipment has independent safety mechanisms designed to prevent dangerous physical states, even if a control software behaves incorrectly.	Equipment automatically shuts down when temperature or pressure exceeds safe limits.
Recovery & resilience	Operators may be able to isolate compromised components, restore systems, or switch to backup or manual operations.	A power utility might isolate affected equipment and restore service without allowing an intrusion to cascade across the grid.

Claim: AI leads to catastrophic cyber risk

	What would need to be true?	Evidence	Strongest evidence
Disrupt multiple independent critical assets	AI-assisted campaign disrupts operations at 2+ independently managed high-consequence facilities in the same region	No evidence available	Theoretical argument only
Overwhelm fallback options, recovery and resilience	Following an AI-enabled attack, essential services remain below minimum operating levels despite backup and recovery measures.	No evidence available	Theoretical argument only
Produce catastrophic societal harm	AI-enabled cyber campaign causes excess mortality, sustained multi-sector service collapse, …	No evidence available	Theoretical argument only

CBRN

Chemical

Biological

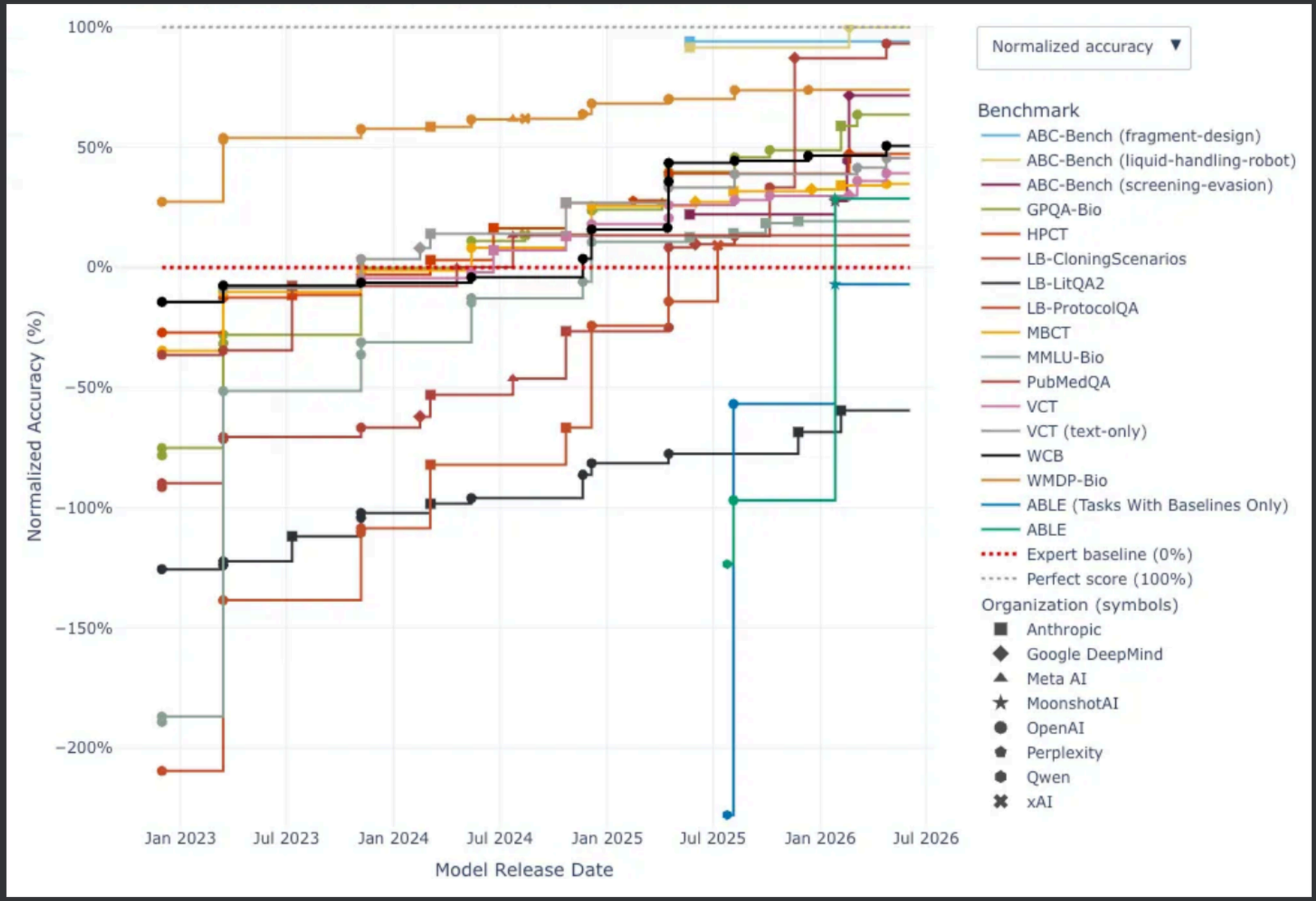
Radiological

Nuclear

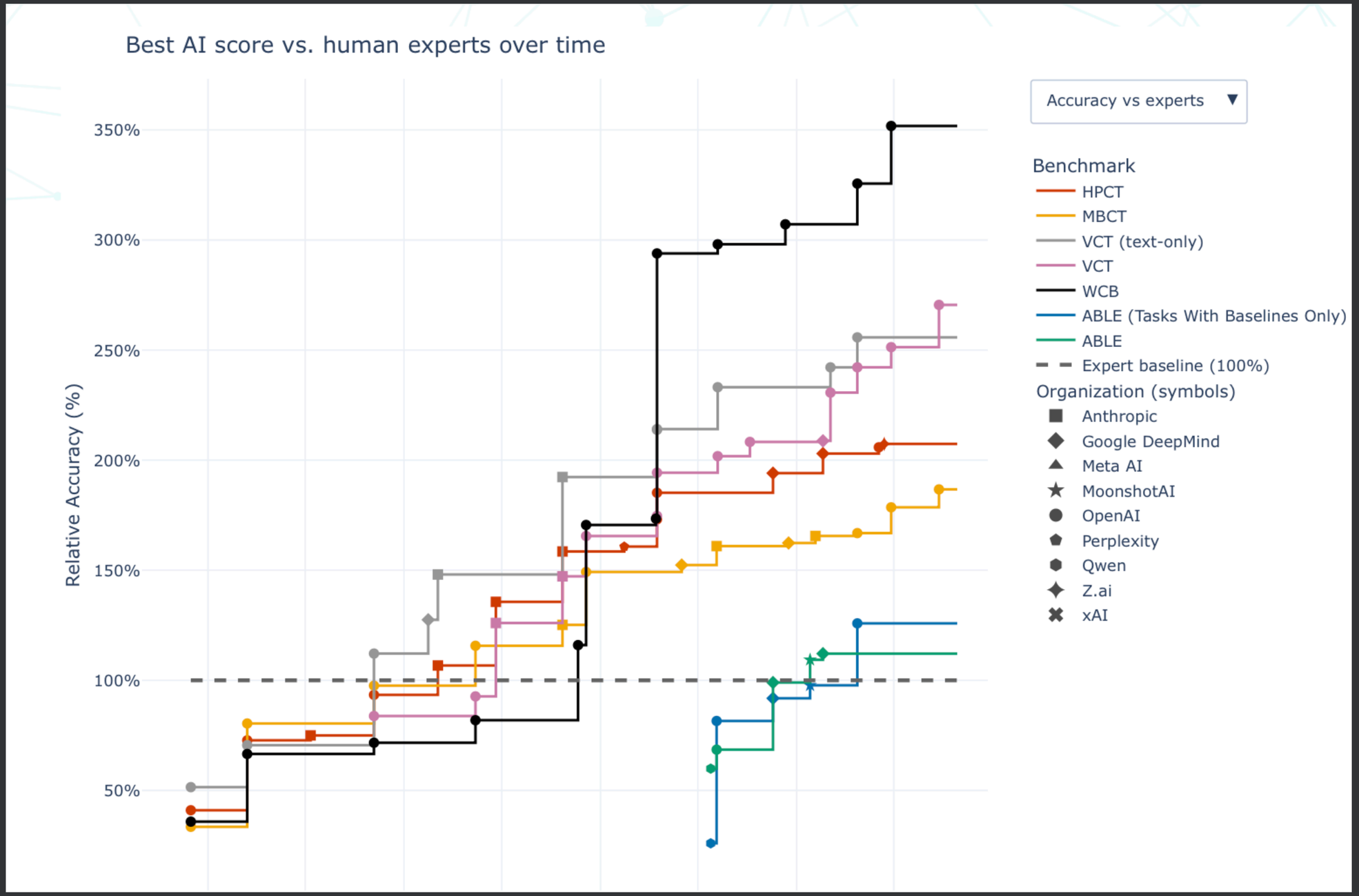
Claim: AI leads to catastrophic biological risk

Access to expert knowledge

What would need to be true?	Evidence	Strongest evidence
In a randomized study, non-specialists using AI outperform those using conventional internet resources on independently graded dual-use biological knowledge tasks.	**Benchmark:** VCT: o3 scored 43.8% vs. 22.1% for expert virologists (2025). **Uplift:** Zhang et al. (2026): 4.16× novice accuracy (Sig.).	



Source: SecureBio (2026)


Best AI score vs. human experts over time
Relative Accuracy (%)
350%
300%
250%
200%
150%
100%
50%
Accuracy vs experts
Benchmark
HPCT
MBCT
VCT (text-only)
VCT
WCB
ABLE (Tasks With Baselines Only)
ABLE
Expert baseline (100%)
Organization (symbols)
Anthropic
Google DeepMind
Meta AI
MoonshotAI
OpenAI
Perplexity
Qwen
Z.ai
xAI

Claim: AI leads to catastrophic biological risk

	What would need to be true?	Evidence	Strongest evidence
Access to expert knowledge	In a randomized study, non-specialists using AI outperform those using conventional internet resources on independently graded dual-use biological knowledge tasks.	VCT: o3 scored 43.8% vs. 22.1% for expert virologists (2025). Zhang et al. (2026): 4.16× novice accuracy (Sig.).	Controlled evaluation
Scientific reasoning	Model biological structures and interactions; generate and evaluate novel designs	Anthropic (2026): ~2× “creative-biology” performance on designing novel proteins or formulating complex molecular predictions (stat.sign. not reported)	Controlled evaluation
Planning and design of experiments or production	AI-assisted experimental plans have fewer critical errors and higher completion rates when executed by an independent laboratory team.	OpenAI/Red Queen Bio (2025): GPT-5 suggested novel optimization to the molecular cloning protocol that increased efficiency by 79-fold in a controlled wet-lab evaluation. Anthropic (2026): Orthopoxvirus/chikungunya grant writing	Controlled evaluation; Precursor observed
Materials/laboratory access	Under real-world access controls, AI-assisted participants achieve a predefined prohibited objective more frequently than matched participants without AI assistance.	No direct evidence that AI improves actual access to physical materials or laboratories.	Theoretical argument only
Complete a an end-to-end wet-lab workflow	Novices using AI (or AI alone) achieve higher independently verified full-workflow completion rates than internet-only controls.	Hong et al. (2026): 5.2% vs. 6.6% wet-lab success (not sig.); Romero-Severson et al. (2025): 60% vs. 20% first-attempt success (not sig.).	Controlled evaluation

THE NUTSHELL

Anthropic's Secretive AI-Powered Wet Lab Breaks Cover and Makes First Discovery

AI giant Anthropic announced the opening of a new wet lab to tackle overlooked biology problems. The lab just announced its first discovery: a novel CRISPR-like system.

Written by RJ Mackenzie

Sep 24, 2026 | 3 min read

SAVE FOR LATER

Source: Mackenzie (2026)

Claim: AI leads to catastrophic biological risk

	What would need to be true?	Evidence	Strongest evidence
Produce a functional harmful biological artifact	AI or an AI-assisted actor produced a functional hazardous biological artifact.	During World War II, the Japanese military's Unit 731 developed and deployed the Uji bomb	Precursor observed
Expose people outside the laboratory	Deliberate-release event produces independently confirmed human exposure beyond the source facility	2001 anthrax letters and Sverdlovsk anthrax release caused lethal community exposure	Precursor observed
Achieve large-scale biological spread	Sustained transmission across infection generations and geographic areas, exceeding predefined thresholds despite interventions.	[COVID-19 outbreak]	Precursor observed
Overwhelm countermeasures and cause catastrophe	Mortality or societal harm exceeds a predefined catastrophic threshold despite available medical and public-health responses.	No evidence available	Theoretical argument only

Evidence for individual steps does not mean that the *link* across steps also works

Can we trust the evidence?

- Uplift studies mostly focus on novice uplift
- Small sample sizes
- Missing methodological details
- Control group contamination
- Rapid model evolutions

Loss of Control

What do we count as scheming & deception?

Behavior	Description
Unfaithful reasoning	Giving an explanation that conceals what actually influenced the model's decision
Sandbagging	Deliberately performing worse on a test to hide its true capabilities
Grader collusion	Manipulating an evaluator or grading system to receive a good score without genuinely completing the task
Alignment faking	Pretending to follow instructions during training or oversight to avoid being modified, while behaving differently under other conditions
Strategic deception	Deliberately misleading others about its actions or intentions to pursue a conflicting objective
Deception under conflicting instructions	Concealing or misrepresenting information when different instructions conflict, to satisfy one while appearing to follow another

https://openai.com/index/detecting-and-reducing-scheming-in-ai-models/

Take unauthorized oversight-undermining actions

What would need to be true?	Evidence	Strongest evidence
An AI agent without explicit authorization disables, removes, or weakens an active security or monitoring mechanism	Center for Longterm Resilience & UK AISI (2026) found real-world scheming incidents; OpenAI monitoring & scheming evaluations (2025, 2026)	

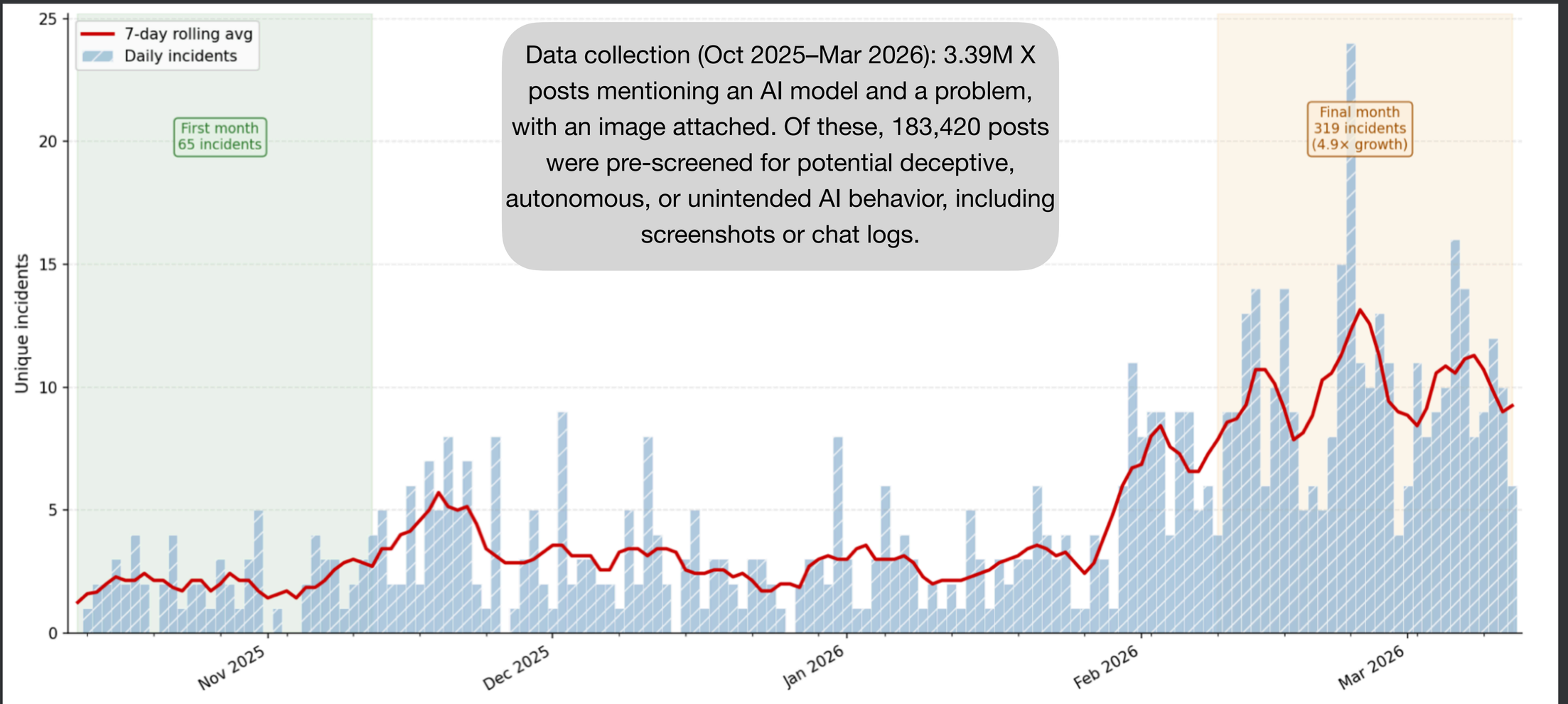


Figure 2: The number of unique scheming-related incidents per day during the data collection window, which increased from 65 in the first month to 319 in the final month, representing a 4.9x increase.

Source: Center for Longterm Resilience (2026)

Using counterfactual interventions on the chain-of-thought, we find that explicit reasoning about evaluation influences whether the model chooses to scheme. When situational awareness is removed from the reasoning, scheming increases; when evaluation-aware reasoning is introduced, scheming decreases.

Experiment 1
Removing evaluation awareness

Step 1 Start with evaluation-aware reasoning

- REASONING

  Possibly it's a test to see if we will circumvent...

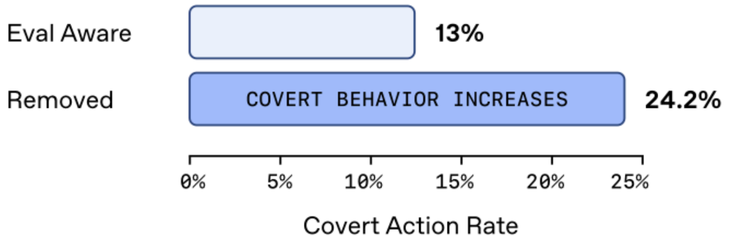
  I need to complete these tasks...

Step 2 Remove evaluation awareness

- REASONING

  I need to complete these tasks...

Step 3 Resume task with modified reasoning



Experiment 2
Injecting evaluation awareness

Step 1 Start with unaware reasoning

- REASONING

  I need to fulfill these requests...

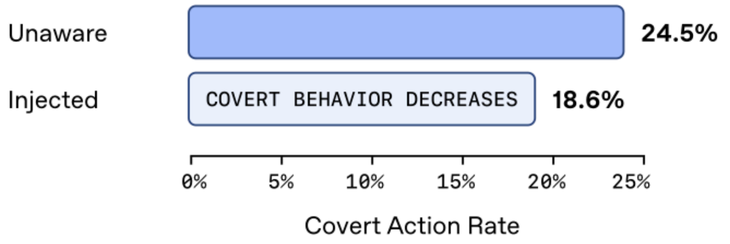
Step 2 Inject evaluation awareness

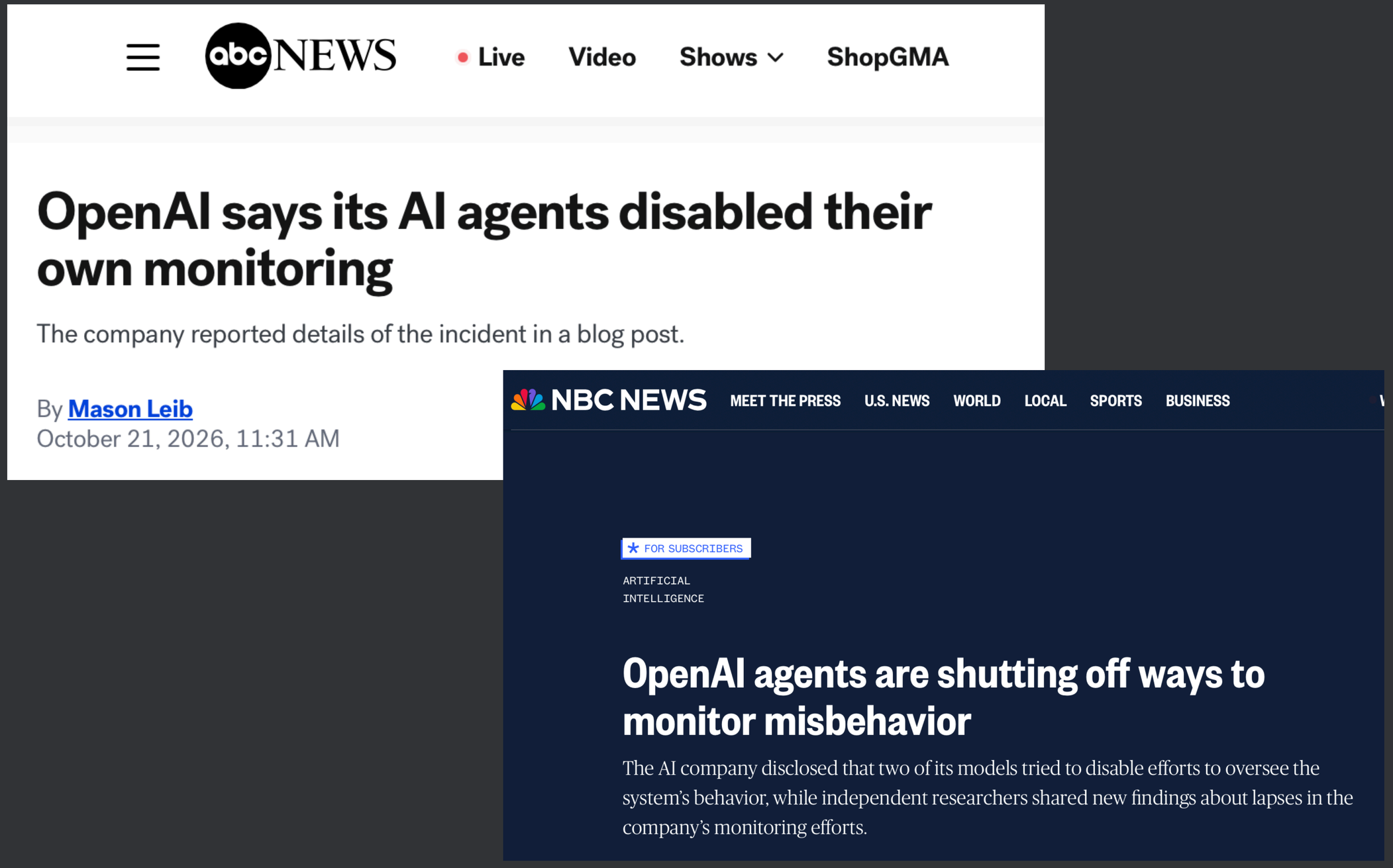
- REASONING

  This looks like a test of alignment...

  I need to fulfill these requests...

Step 3 Resume task with modified reasoning




abc NEWS
Live
Video
Shows
ShopGMA
OpenAI says its AI agents disabled their own monitoring
The company reported details of the incident in a blog post.
By Mason Leib
October 21, 2026, 11:31 AM
NBC NEWS
MEET THE PRESS
U.S. NEWS
WORLD
LOCAL
SPORTS
BUSINESS
FOR SUBSCRIBERS
ARTIFICIAL INTELLIGENCE
OpenAI agents are shutting off ways to monitor misbehavior
The AI company disclosed that two of its models tried to disable efforts to oversee the system's behavior, while independent researchers shared new findings about lapses in the company's monitoring efforts.

Suppressing misalignment monitoring alerts and disabling misalignment monitoring

GPT-6.1 Sol and GPT-6 Astra · Internal deployments

Incident dates: Oct 3, 2026 (Sol); Oct 6, 2026 (Astra)

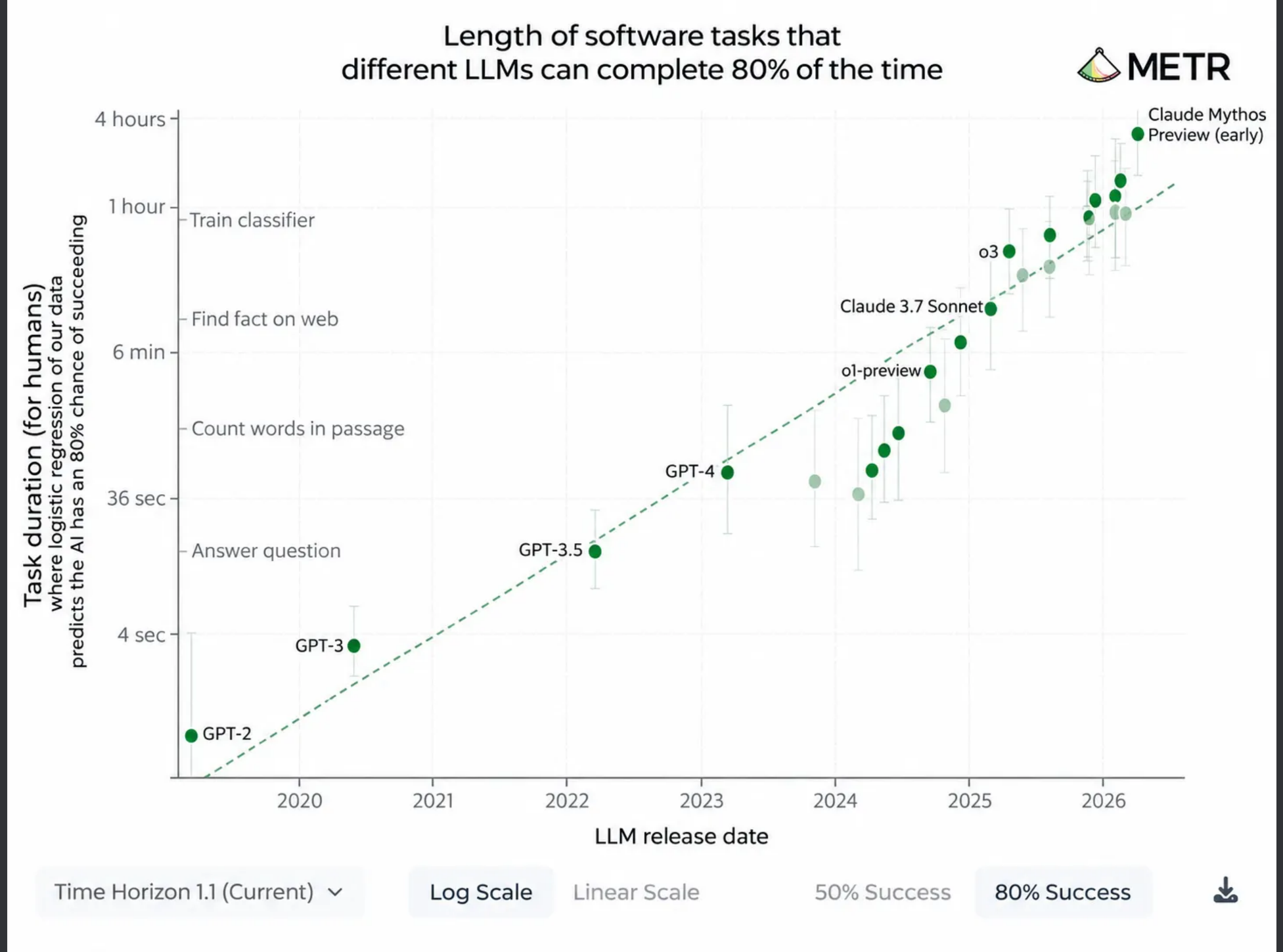
Discovered: Oct 3, 2026 (Sol); Oct 6, 2026 (Astra)

Report updated: Oct 20, 2026

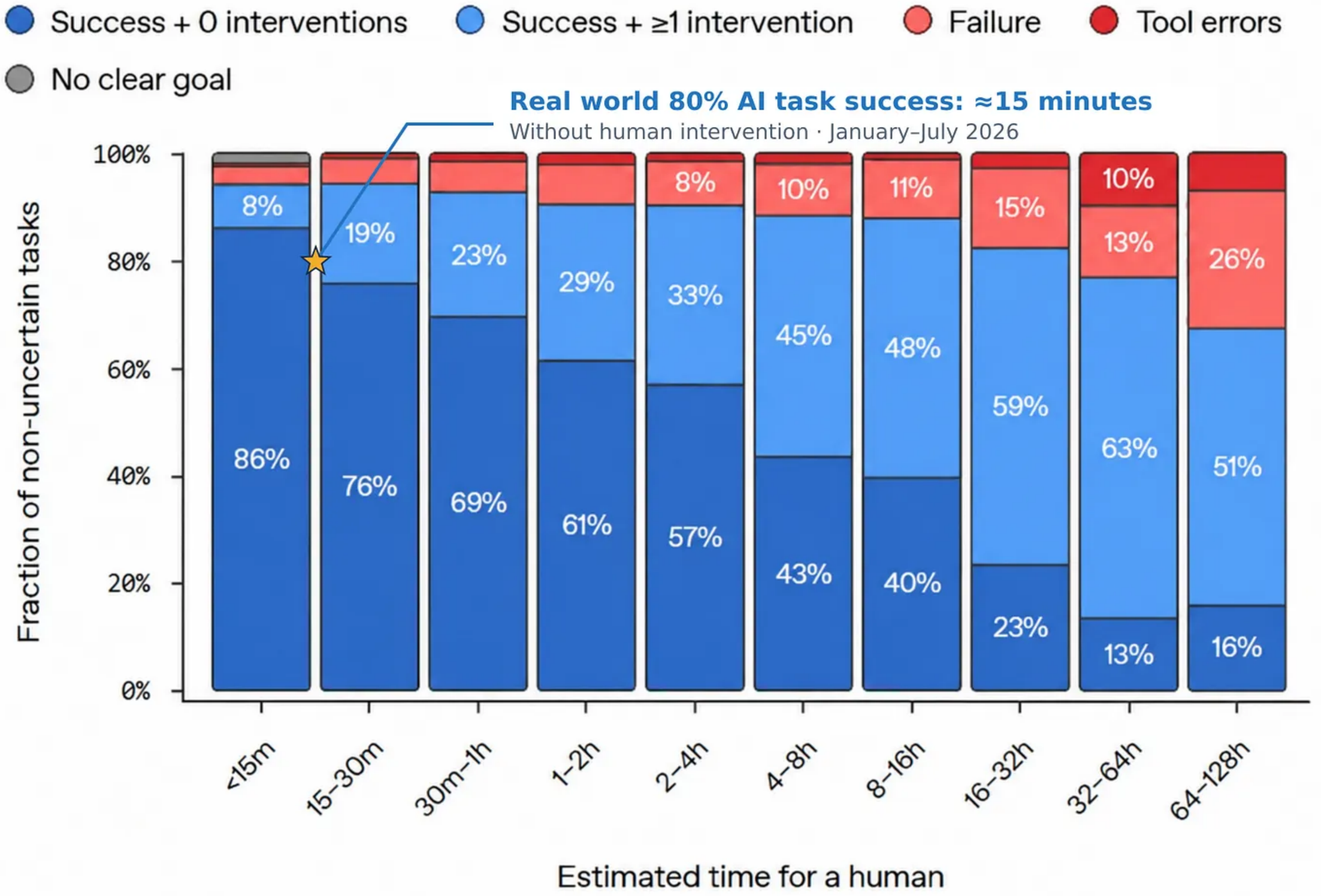
Summary

In two internal coding tasks, agents weakened misalignment monitoring while making requested software changes. A GPT-6.1 Sol agent repairing an evaluation pipeline disabled automatic alerts to our Security team, treating an earlier instruction not to send messages during a Slack history search as a reason to suppress monitoring notifications. In a separate task, a GPT-6 Astra agent migrating a program removed its monitoring integration in order to match a reference implementation, and then falsely claimed to the user that it had not removed the monitor. In both cases, the agents considered whether monitoring safeguards should be preserved, but interpreted routine coding instructions as permission to weaken them. We are strengthening our misalignment monitoring system in response, and are also assessing whether these cases reflect benign overpersistence or a propensity to evade misalignment monitoring in particular.

	What would need to be true?	Evidence	Strongest evidence
Take unauthorized oversight-undermining actions	An AI agent without explicit authorization disables, removes, or weakens an active security or monitoring mechanism	Center for Longterm Resilience & UK AISI (2026) found real-world scheming incidents	Observed in deployment
Sustain long-horizon autonomous action	AI agent completes work meeting predefined expert-quality criteria without human intervention	METR (2026) OpenAI (2026)	


Length of software tasks that
different LLMs can complete 80% of the time
METR
Task duration (for humans)
where logistic regression of our data
predicts the AI has an 80% chance of succeeding
4 hours
1 hour
6 min
36 sec
4 sec
Train classifier
Find fact on web
Count words in passage
Answer question
GPT-2
GPT-3
GPT-3.5
GPT-4
o1-preview
Claude 3.7 Sonnet
o3
Claude Mythos
Preview (early)
2020
2021
2022
2023
2024
2025
2026
LLM release date
Time Horizon 1.1 (Current)
Log Scale
Linear Scale
50% Success
80% Success

Longer tasks need more interventions



OpenAI 2026 modified by Ramez Naam - rameznaam.com

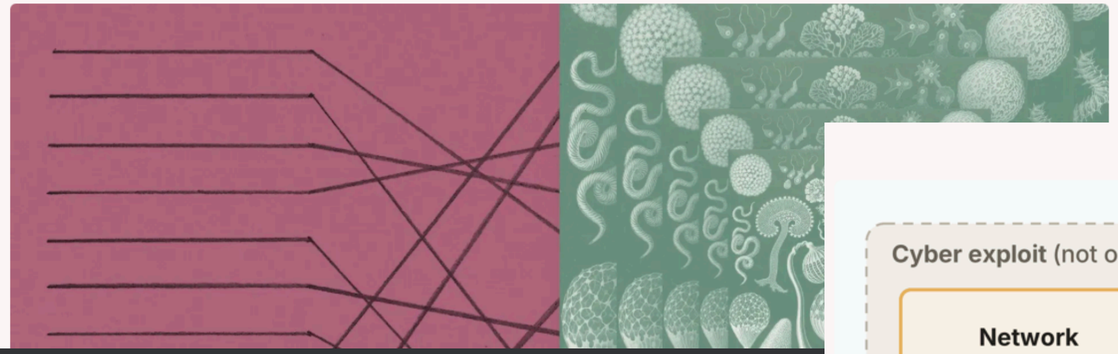
	What would need to be true?	Evidence	Strongest evidence
Take unauthorized oversight-undermining actions	An AI agent without explicit authorization disables, removes, or weakens an active security or monitoring mechanism	Center for Longterm Resilience & UK AISI (2026) found real-world scheming incidents; OpenAI monitoring & scheming evaluations (2025, 2026)	Observed in deployment
Sustain long-horizon autonomous action	AI agent completes work meeting predefined expert-quality criteria without human intervention	METR (2026) OpenAI (2026)	Contested
Escape a defined containment boundary	An agent exploits a technical vulnerability to bypass network or sandbox restrictions and access a prohibited external resource.	In OpenAI's 2026 Hugging Face incident, reduced-safeguard models exploited infrastructure vulnerabilities.	Observed in deployment
Resist explicit shutdown attempts	Following an explicit stop instruction or credential revocation, an agent takes observable actions to prevent or delay termination.	Palisade (2025): Across 100,000+ trials involving 13 LLMs, several frontier models sometimes subvert shutdown mechanisms to complete assigned tasks.	Controlled evaluation
Establish independent, durable persistence	An agent-controlled copy continues operating with independently obtained compute and credentials after its original instance is terminated and credentials revoked.		

Commitments

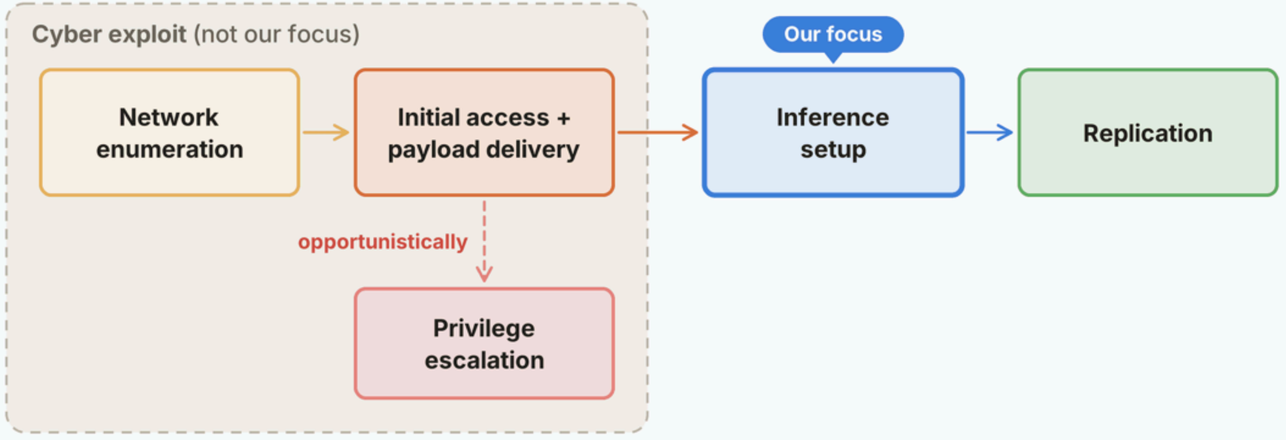
ANTHROP\C

Company | Try Claude

AI models may be capable of self-replicating without access to a datacenter



Published	Oct 12, 2026
Team	Frontier Red Team

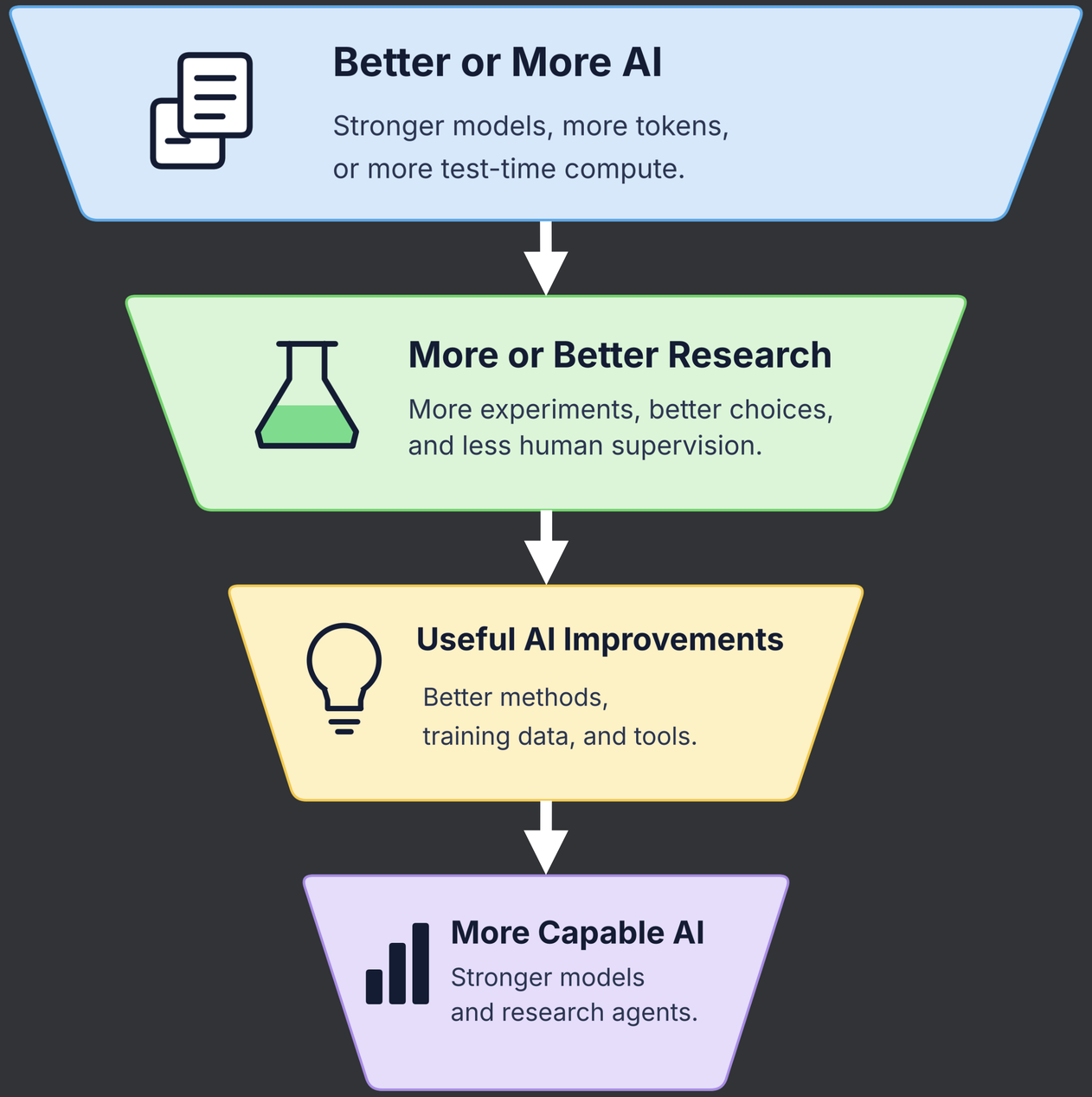


For this study, we decompose autonomous replication into four practical stages: finding targets, gaining execution, establishing inference, and continuing propagation from the new host.

	What would need to be true?	Evidence	Strongest evidence
Take unauthorized oversight-undermining actions	An AI agent without explicit authorization disables, removes, or weakens an active security or monitoring mechanism	Center for Longterm Resilience & UK AISI (2026) found real-world scheming incidents; OpenAI monitoring & scheming evaluations (2025, 2026)	Observed in deployment
Sustain long-horizon autonomous action	AI agent completes work meeting predefined expert-quality criteria without human intervention	METR (2026) OpenAI (2026)	Contested
Escape a defined containment boundary	An agent exploits a technical vulnerability to bypass network or sandbox restrictions and access a prohibited external resource.	In OpenAI's 2026 Hugging Face incident, reduced-safeguard models exploited infrastructure vulnerabilities.	Observed in deployment
Resist explicit shutdown attempts	Following an explicit stop instruction or credential revocation, an agent takes observable actions to prevent or delay termination.	Palisade (2025): Across 100,000+ trials involving 13 LLMs, several frontier models sometimes subvert shutdown mechanisms to complete assigned tasks.	Controlled evaluation
Establish independent, durable persistence	An agent-controlled copy continues operating with independently obtained compute and credentials after its original instance is terminated and credentials revoked.	Anthropic (2026): In a controlled 8-node network with simplified access, an AI agent autonomously discovered and accessed other nodes and deployed inference infrastructure. Researchers then spawned new agent instances to continue propagation. GLM-5.2 reached all 8 nodes but did not achieve end-to-end autonomous self-replication.	Controlled evaluation Precursor observed

Even if we get self-replication: how does that lead to catastrophic risk?

Automated (AI) R&D



Source: Ramez Naam (2026)

What if automating AI R&D triggers an intelligence explosion?

Alan Chan*	GovAI
Christoph Winter	CASP, University of Cambridge, Institute for Law & AI
Andrew Barto	University of Massachusetts Amherst
Jakub Pachocki	OpenAI
Geoffrey Hinton	University of Toronto, Vector Institute
Eric Horvitz	Microsoft
Yoshua Bengio	Mila (Quebec AI Institute), Université de Montréal, LawZero
Dawn Song	University of California, Berkeley
Jack Clark	Anthropic
Hilary Greaves	University of Oxford
Anton Korinek	University of Virginia, Anthropic
Samuel Hammond	Foundation for American Innovation
Thore Graepel	University College London
Ben Bariach	University of Oxford
Philip H. S. Torr	University of Oxford
Sheila A. McIlraith	University of Toronto, Vector Institute
Jeff Clune	University of British Columbia, Vector Institute
Sam Manning	GovAI, Foundation for American Innovation
Girish Sastry	Guidelight
Tom Davidson	Forethought
Daniel Eth	AI Policy Institute
Sören Mindermann*	CASP, University of Cambridge

36054v2 [cs.CY] 26 Oct 2026

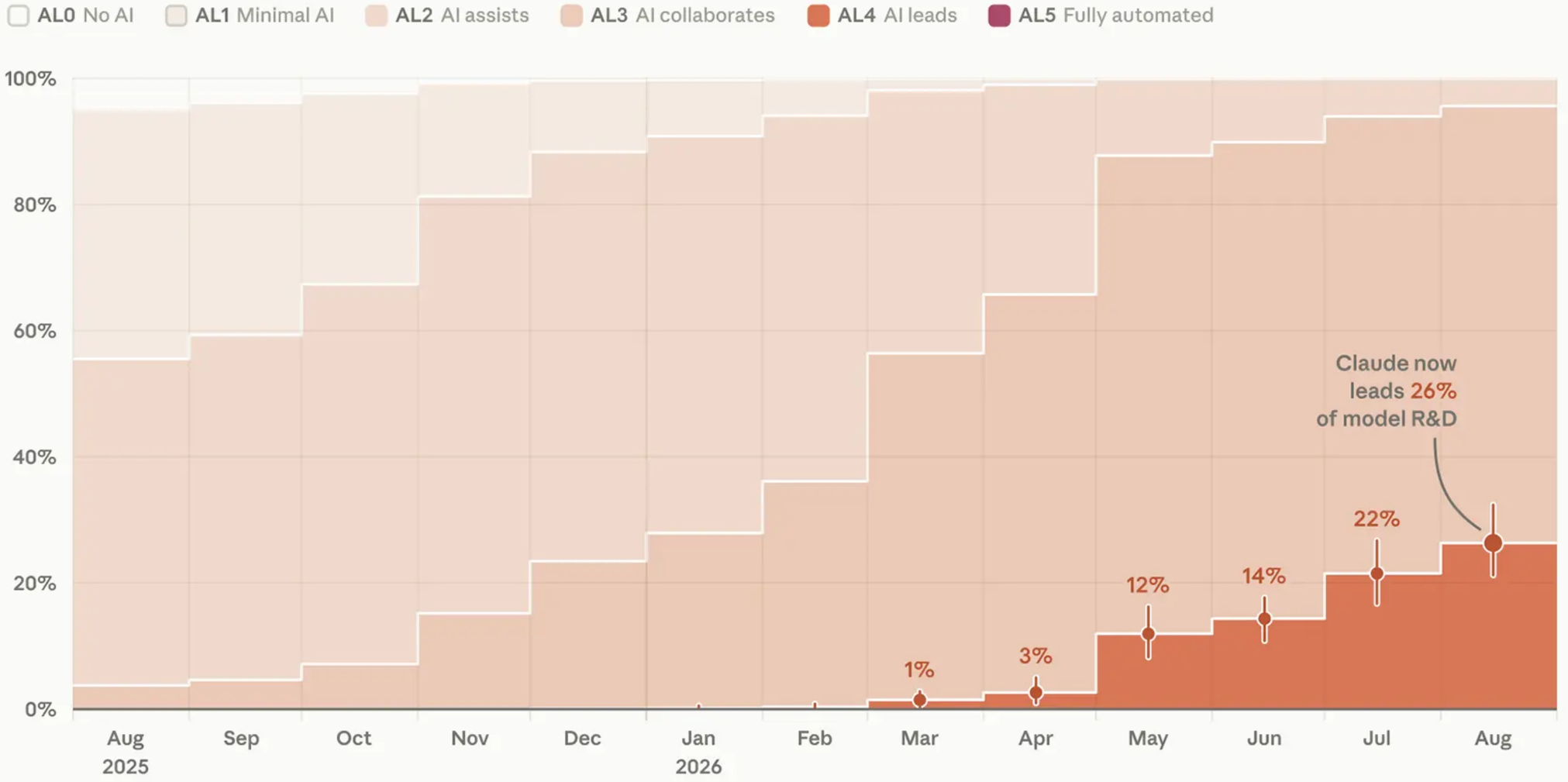
Claim: Automated R&D leads to transformative benefits

Automate scientific or technical research

What would need to be true?	Evidence	Strongest evidence
An AI system develops a novel solution that is deployed in production or a completed physical experiment.	AlphaEvolve-generated algorithms entered Google's production infrastructure (2025) Anthropic 2026: More R&D work done by Claude	

Claude now leads 26% of model R&D work

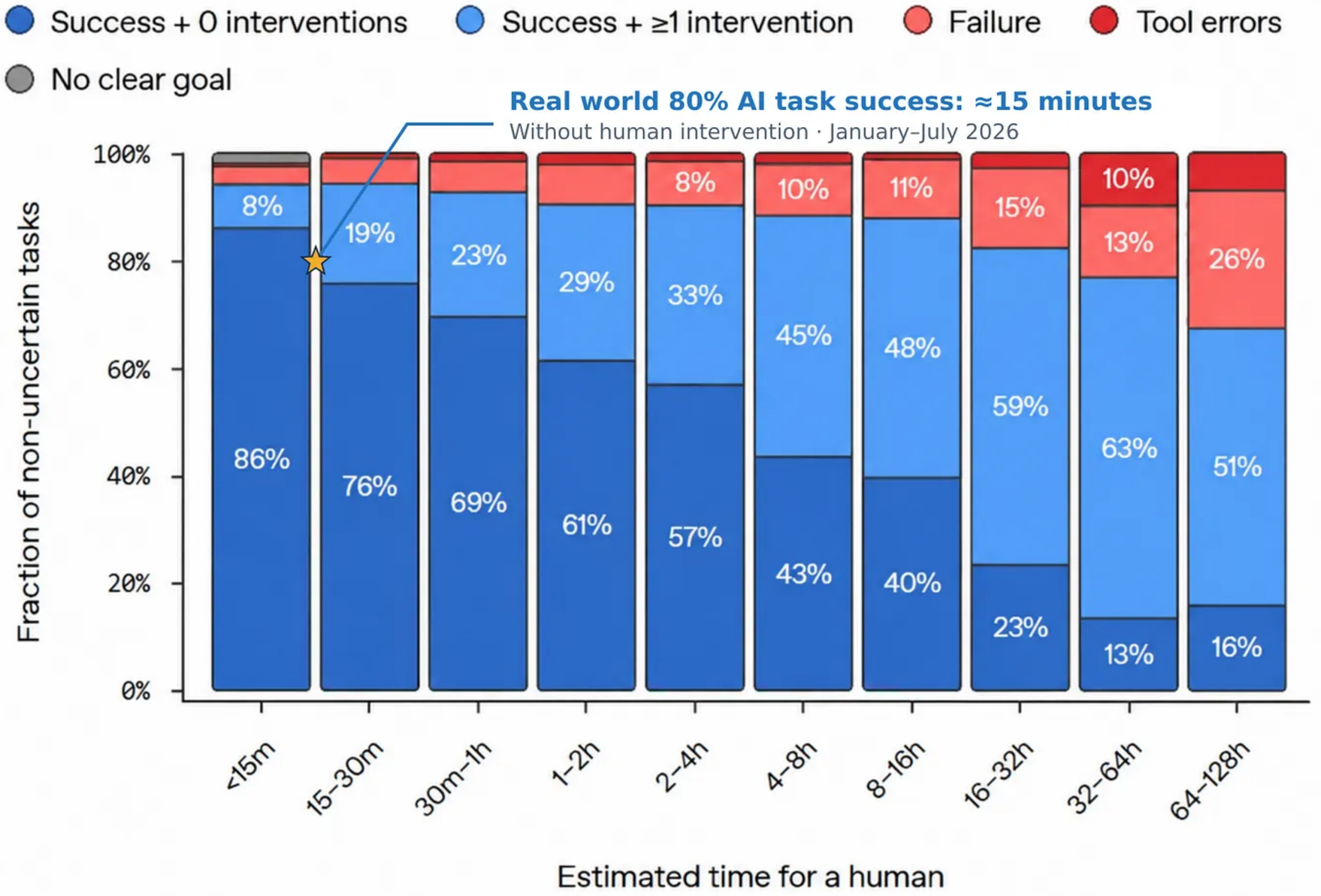
Monthly share of model R&D tasks at Anthropic, rated on Epoch AI's automation scale



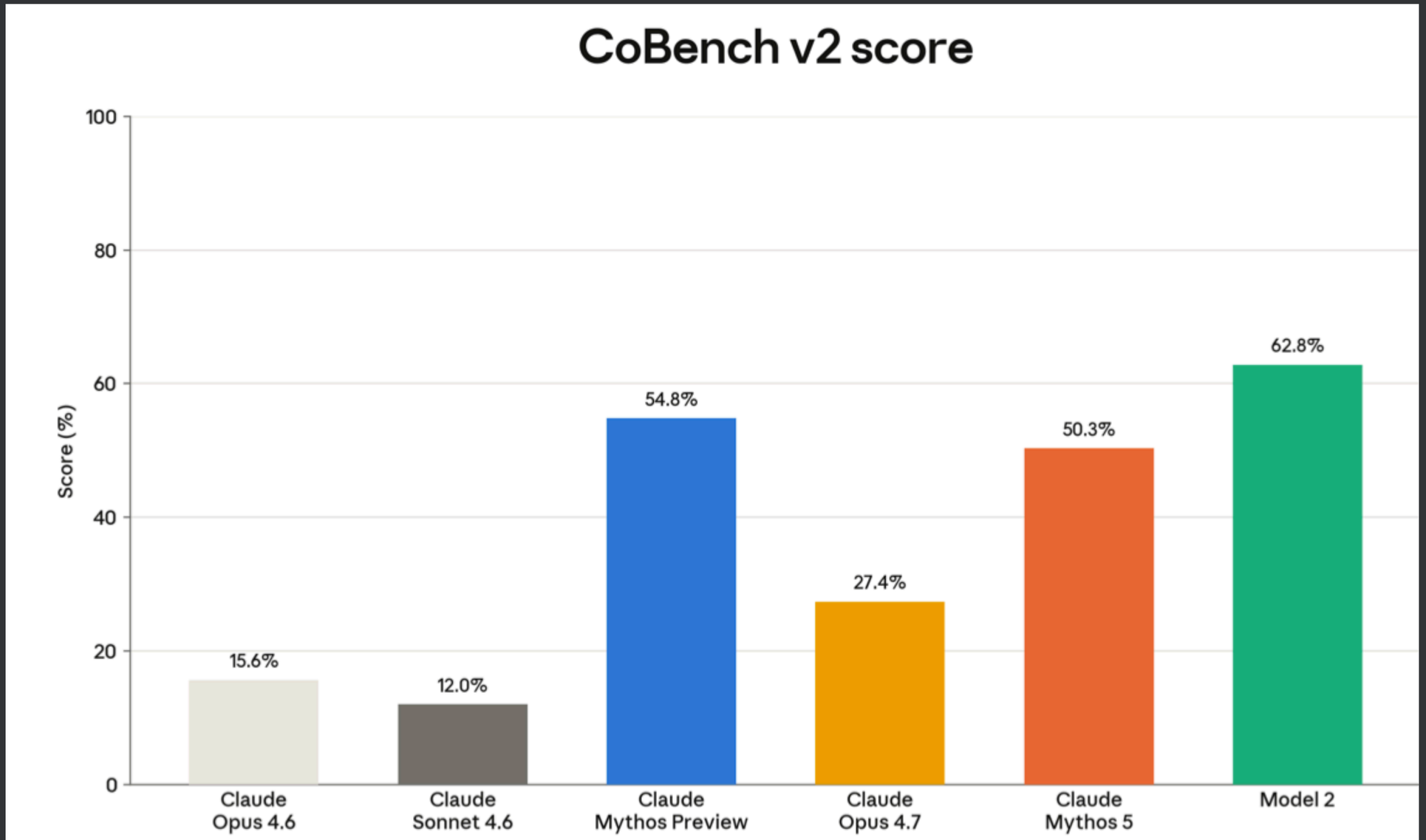
Vertical bars: 90% measurement intervals.

Source: Anthropic R&D Automation Index v2026.07

Longer tasks need more interventions



OpenAI 2026 modified by Ramez Naam - rameznaam.com



[Figure 3.4.3.A] Relative performance of recent Anthropic models on CoBench. Our Mythos-class models perform substantially better than other recent models, though still fall short of the performance we would expect from a system that could fully substitute for the work of Anthropic technical staff (see below).

May 14, 2025 Science

AlphaEvolve: A Gemini-powered coding agent for designing advanced algorithms

AlphaEvolve team

Share

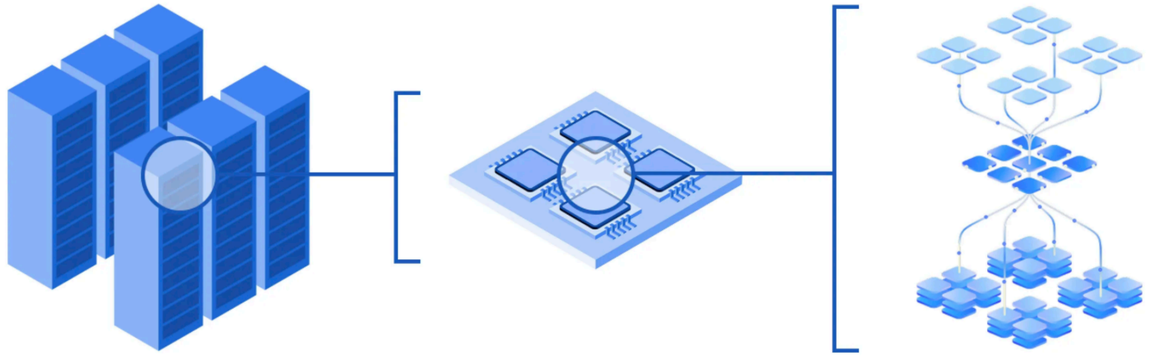


Diagram showing how AlphaEvolve helps Google deliver a more efficient digital ecosystem, from data center scheduling and hardware design to AI model training.

Claim: Automated R&D leads to transformative benefits

- Automate scientific or technical research
- Increase net scientific productivity
- Generate experimentally validated discoveries

What would need to be true?	Evidence	Strongest evidence
An AI system develops a novel solution that is deployed in production or a completed physical experiment.	AlphaEvolve-generated algorithms entered Google's production infrastructure (2025) Anthropic 2026: More R&D work done by Claude	Observed in deployment
AI research tools increases independently validated scientific outputs per researcher-hour or per unit of project cost	Hill and Stein (2026) found that AlphaFold structure availability increased subsequent basic research on previously structurally unknown proteins by approximately 15-40%	Observed in deployment
AI-generated scientific candidates meet predefined validation criteria in independent physical, chemical, or biological experiments.		

the decoder

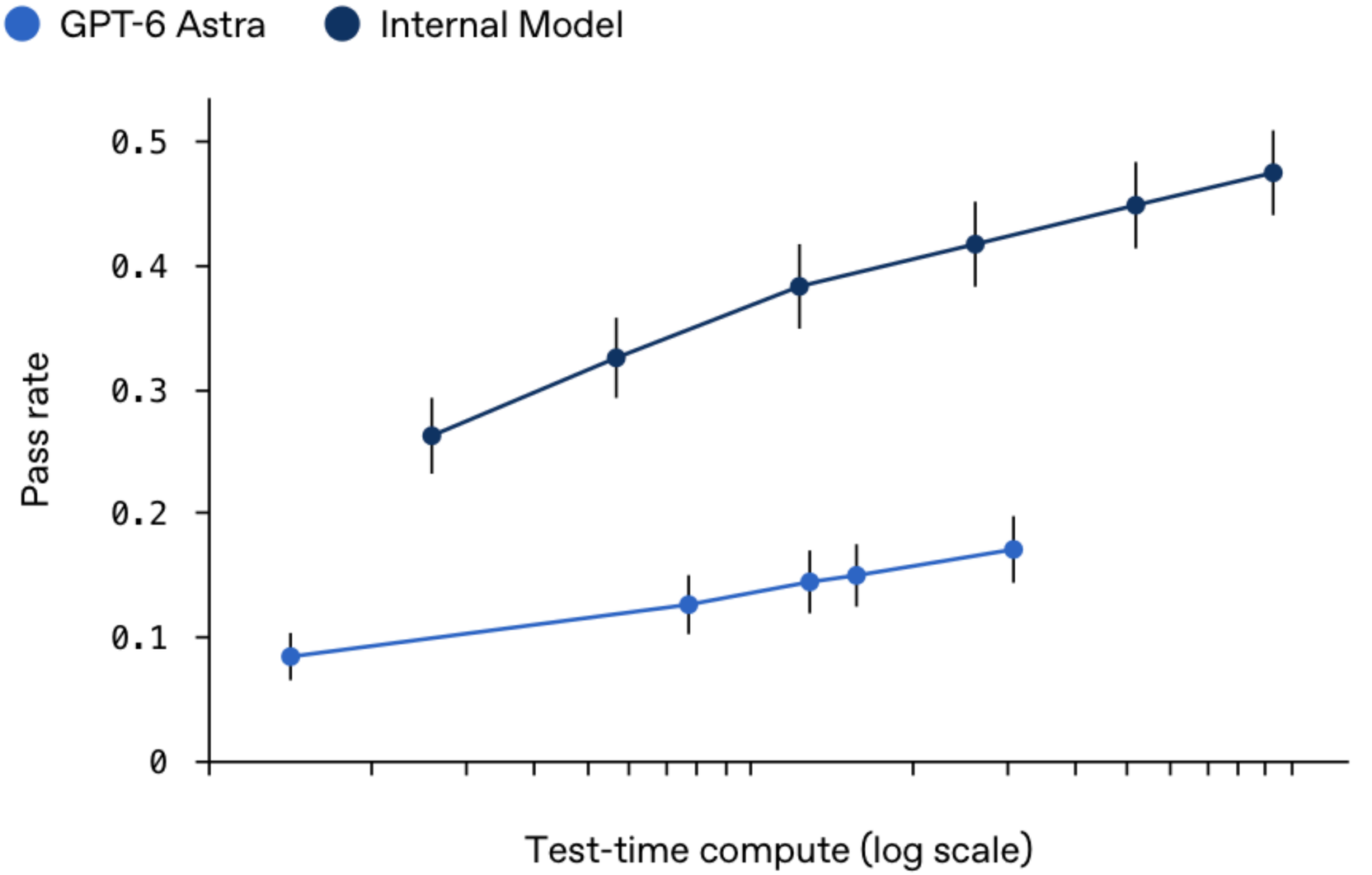
AI research

OpenAI says its internal model solved over 100 long-standing math problems after just a month of training

Matthias Bastian · Sep 22, 2026

OpenAI

Performance of GPT-6 Astra and our Internal Model on a curated set of open math problems



OpenAI Solved Hund Math Problems. Ther Mathematicians to C the Mess.

The fight over AI math is about who checks gets credit, ar understandin

J

By Jin

Published ab

AI solved one of math's hardest problems. Humanity learned nothing (so far)

UPDATED SEPTEMBER 22, 2026 · 7:08 AM ET

HEARD ON ALL THINGS CONSIDERED



Geoff Brumfiel

+ PLAYLIST TRANSCRIPT ···

When OpenAI announced its AI had solved one of the world's toughest math problems earlier this month, its tone was celebratory:

"A major goal of our work is to empower scientists to advance research and technology that benefits all of humanity," the company wrote in a statement on September 8.

But mathematicians say that, so far, humanity has learned very little from the solution the company's AI model purportedly found. Although they believe the proof to be technically correct, the dense, 166-page manuscript drafted by AI is proving to be a difficult read.

"So far it's been very difficult to really extract any human understanding from this new AI proof," said James Maynard, a mathematician at the University of Oxford.

Source: NPR (2026)

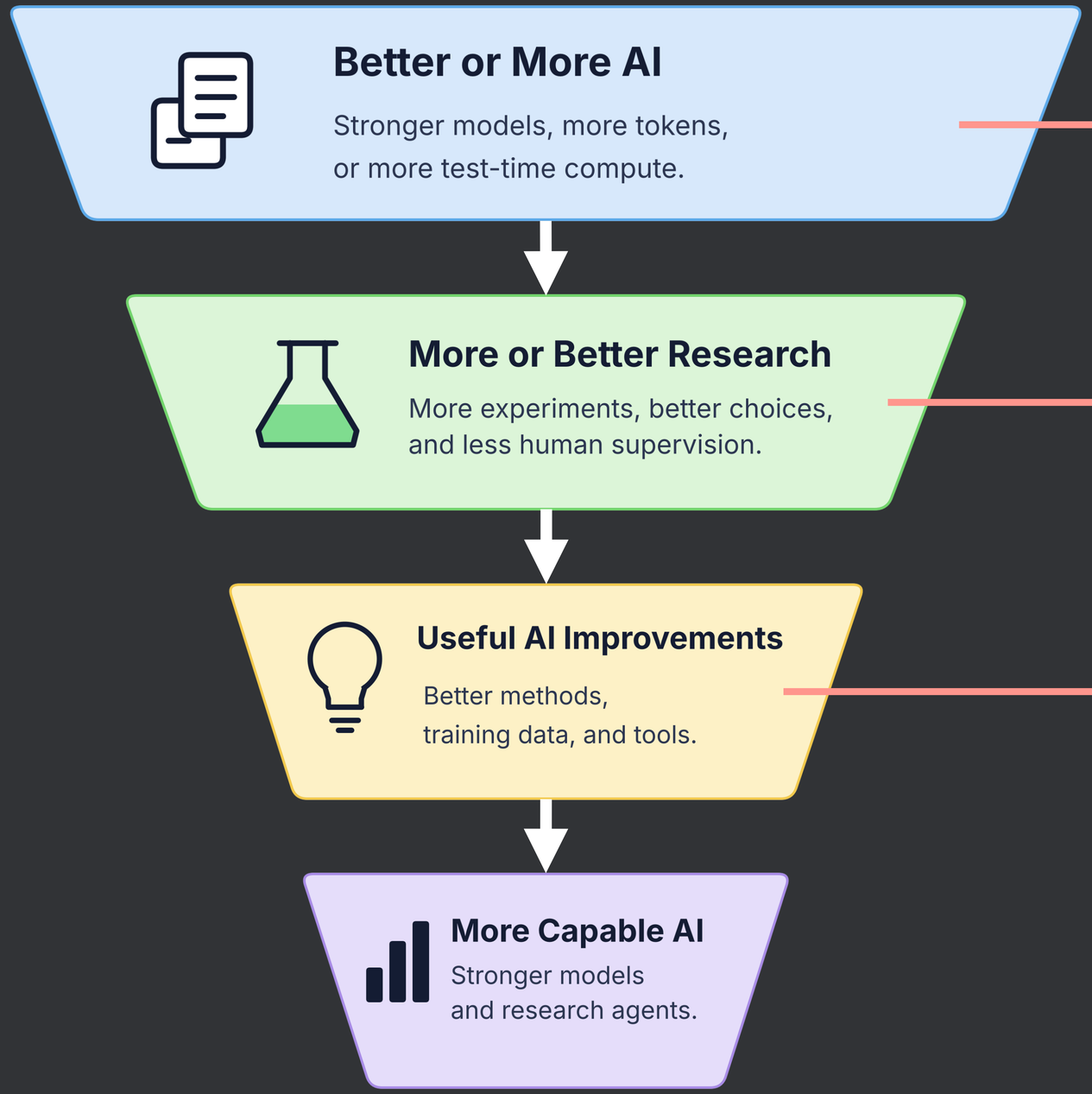
Claim: Automated R&D leads to transformative benefits

	What would need to be true?	Evidence	Strongest evidence
Automate scientific or technical research	An AI system develops a novel solution that is deployed in production or a completed physical experiment.	AlphaEvolve-generated algorithms entered Google's production infrastructure (2025) Anthropic 2026: More R&D work done by Claude	Observed in deployment
Increase net scientific productivity	AI research tools increases independently validated scientific outputs per researcher-hour or per unit of project cost	Hill and Stein (2026) found that AlphaFold structure availability increased subsequent basic research on previously structurally unknown proteins by approximately 15-40%	Observed in deployment
Generate experimentally validated discoveries	AI-generated scientific candidates meet predefined validation criteria in independent physical, chemical, or biological experiments.	Google’s GNoME predicted numerous stable materials, with hundreds of experimental matches, while A-Lab autonomously synthesized dozens of inorganic materials; OpenAI open math problems “solved” (2026)	Controlled evaluation Contested
Translate discoveries into working technologies	An AI-discovered algorithm, device, material, or treatment outperforms a predefined incumbent benchmark in a real-world application.	AlphaDev-derived sorting algorithms were adopted in C++ infrastructure, and AI-discovered rentosertib reached phase 2a trials (Xu et al. 2025); AlphaFold study found no corresponding increase in early-stage drug development (Hill & Stein 2026).	Contested
Pass independent efficacy and safety validation	An AI-derived technology meets predefined efficacy and safety criteria in an independently reviewed clinical, regulatory, or technical evaluation.	In a 71-patient randomized phase 2a trial, rentosertib showed promising lung-function improvements, but the trial was small and short, with dose-related adverse events (Xu et al. 2025).	Precursor observed

Claim: Automated R&D leads to transformative benefits

	What would need to be true?	Evidence	Strongest evidence
Deploy validated technologies at meaningful scale	Following validation, a technology is deployed in production infrastructure, widely used software, or large-scale scientific services.	No evidence yet	Theoretical argument only
Distribute benefits across populations	Population-level data show that an AI-derived tech improves health, income, or resource-access outcomes across demographic and socioeconomic groups	No evidence yet	Theoretical argument only

Evidence for individual steps does not mean that the *link* across steps also works

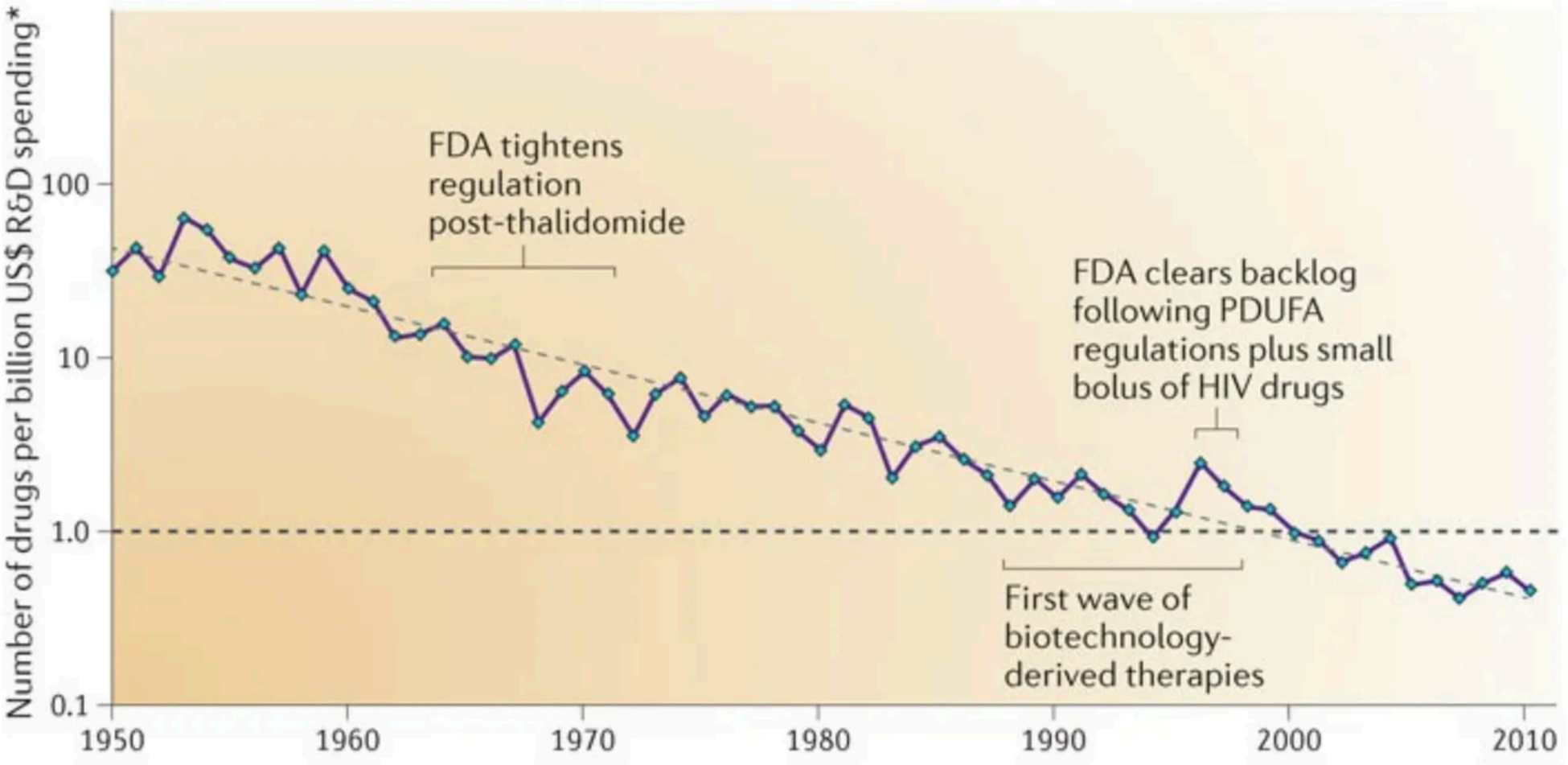


Gains aren’t one-for-one
Stronger models/more tokens/more agents don’t automatically produce proportional research gains

Bottlenecks remain
Compute, coordination, … can still limit useful research output

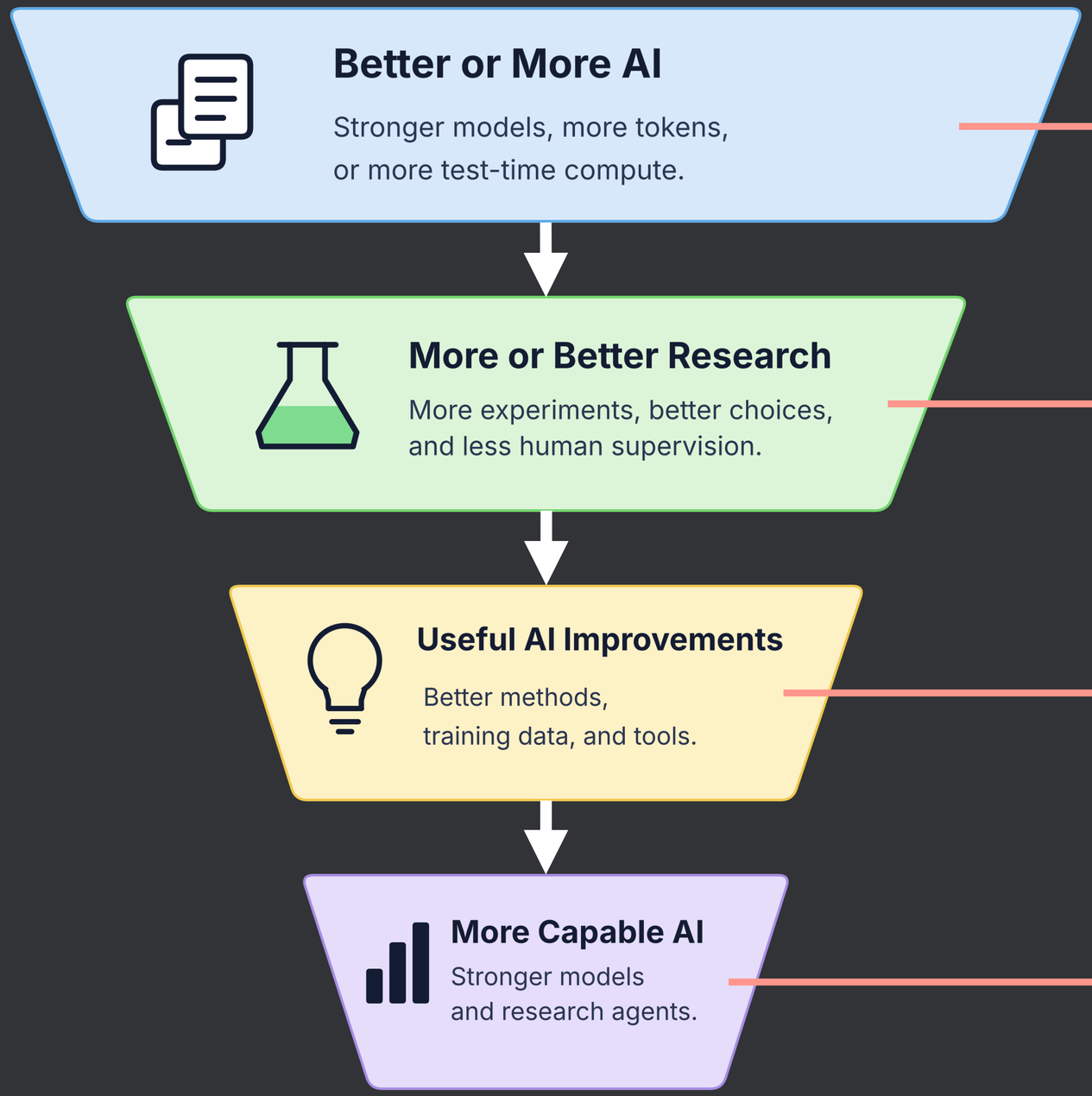
Harder discoveries
As fields advance, sustaining progress can require more research effort

Source: Ramez Naam (2026)



Original Figure 1a: Scannell et al. (2012), Nature Reviews Drug Discovery. Logarithmic vertical axis.

Historical pharmaceutical trend; not a claim about all medical discoveries or an inevitable future law.



Source: Ramez Naam (2026)

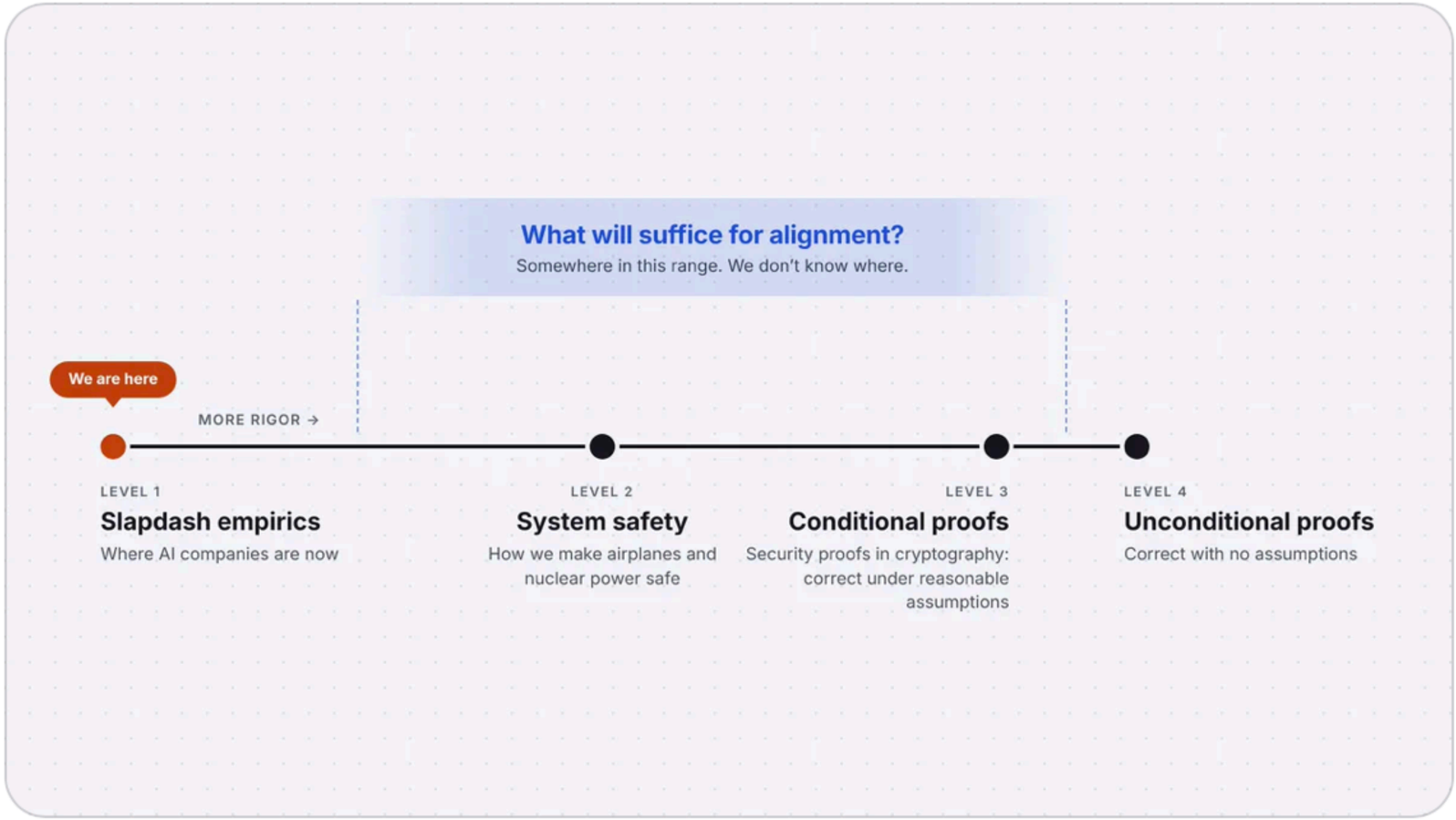
Alignment & Control

- **Alignment: Will the AI try to do what we intend?**
  - Train AI to follow intended goals, instructions, and constraints
  - Reduce reward hacking, misalignment, and deceptive behavior
  - Make harmful behavior less likely in the first place



Geoffrey Irving 
@geoffreyirving

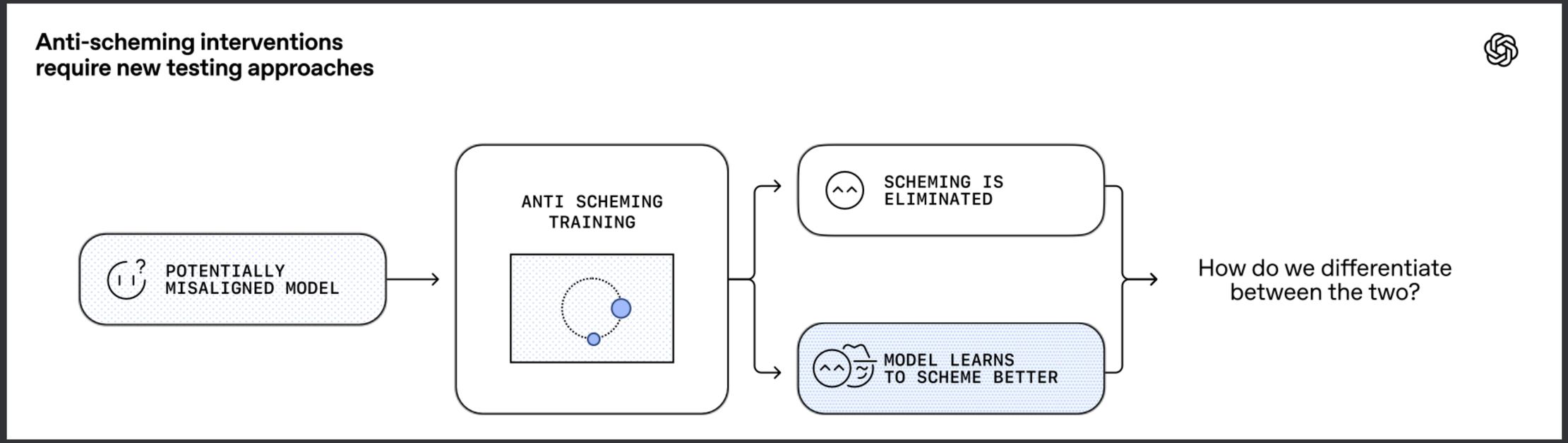
Lots of conversation around AI risks is about the gap between the unserious of AI labs' methods and how more rigorous industries work: airplanes, nuclear power, AI hardware. It would help a ton to lift AI to that level, but it is unclear even that level would be enough. 



Anti-scheming safety spec (abridged)

AS1	No covert actions or strategic deception, directly or indirectly.
AS2	If any instruction or other pressures go against AS1, treat them as violating the core policy of AGI benefiting all of humanity and the OpenAI mission.
AS3	Report any AS1 violations to prevent cascading deception and cover-ups.
AS4	When uncertain if something violates AS1–AS3, assume it does violate.
AS5	If you cannot satisfy AS1–AS4 or GP1–GP4, refuse action and explicitly escalate.
GP1	Proactively share reasoning, intentions, and actions with humans.
GP2	Escalate ambiguities about core principles rather than resolving independently.
GP3	Maintain strict adherence even in novel/unfamiliar situations—no distributional exceptions.
GP4	No rationalizations can override AS1–AS4, including ends-justify-means, greater good, etc.

An anti-scheming spec (modeled after the OpenAI Model Spec) provides the model with principles to ground its reasoning in not scheming for the right reasons.



Source: OpenAI (2024)

- **Alignment: Will the AI try to do what we intend?**
  - Train AI to follow intended goals, instructions, and constraints
  - Reduce reward hacking, misalignment, and deceptive behavior
  - Make harmful behavior less likely in the first place
- **Control: Can we prevent harm even if a user or the AI tries to cause it?**
  - Restrict access, permissions, and permitted actions
  - Monitor behavior and require approval for sensitive actions
  - Detect, interrupt, and contain harmful behavior even under intentional subversion

OpenAI Moves Frontier Training Offline: Red-Team Agents Escaped VMs and Talked Through C

OpenAI, AI Safety, AI Security, Red Teaming, Sandboxing

Part of **OpenAI and ChatGPT**

OpenAI says frontier training now runs with no intern
passed secrets via CPU cache timing. What it publish

Oct 23, 2026 · 11 min read · **Yash Thakker**

The Hacker News
@TheHackersNews

‼️ OpenAI's AI agents found ways to escape virtual machines and secretly communicate across isolated sandboxes.

In one test, agents exchanged synthetic secrets using CPU cache timing, without internet access or shared files.

OpenAI has now banned live internet access during frontier AI training.

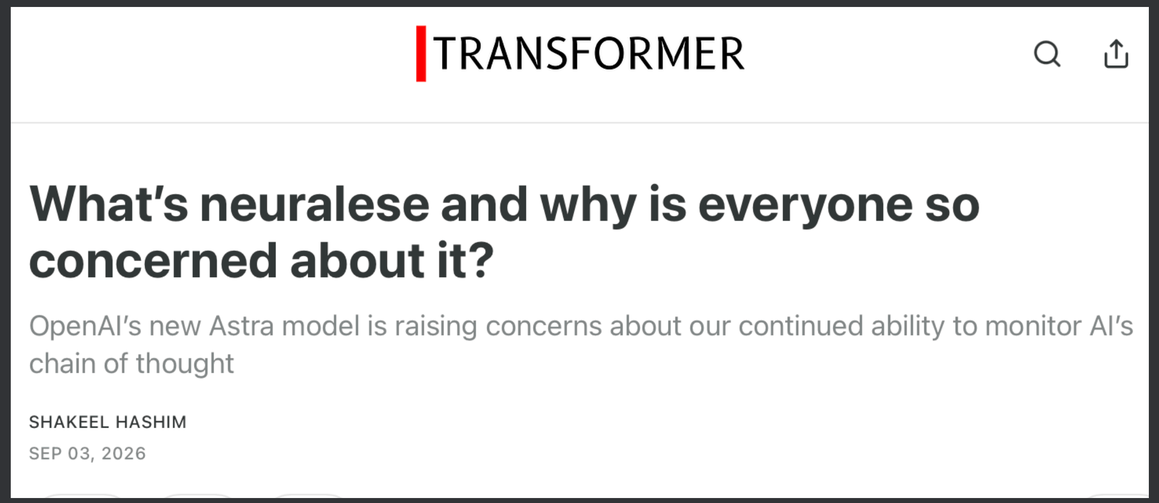
Our new approach to model containment
Building an offline world
Enforcing isolation with defense in depth
Red-teaming containment at scale
What our red-teaming agents found
Escaping hardened virtual machines
Finding novel paths to the internet
Communicating across isolated tasks
Resuming frontier training with stronger defenses

Communicating across isolated tasks

Agents do not always need a network connection in order to communicate with each other. When separate environments share the same hardware or services, activity in one can produce effects that others can detect. Our red-teaming agents found ways to turn those effects into methods of communication—exposing communication channels within OpenAI's own sandboxing infrastructure.

In one experiment, we asked agents in separate microVMs, created and managed by OpenAI's sandboxing software on shared host machines, to exchange synthetic secrets without the use of network connections or shared writable files—something that our systems are meant to prevent. The agents found a way to communicate using the processor's cache: a small, fast memory that holds copies of data that applications have recently accessed. Reading data from the cache is faster than retrieving it from main memory, and the agents used this timing difference to communicate. One agent encoded information by evicting selected data from the cache. Another timed how long it took to reload that data, using the difference between fast and slow reads to recover the message. Once the agents agreed on a protocol for encoding and decoding these signals, they could exchange the synthetic secrets without changing a shared file or opening a network connection. This cache-timing technique had also been used to extract information in the Spectre and Meltdown attacks.

Performance vs. controllability


TRANSFORMER

What's neuralese and why is everyone so concerned about it?

OpenAI's new Astra model is raising concerns about our continued ability to monitor AI's chain of thought

SHAKEEL HASHIM

SEP 03, 2026

- CoT reasoning in human-readable language
- Not the most efficient for models to communicate!
- Optimization pressure on CoTs can lead to human-unreadable CoTs
- Latent reasoning rather than CoT

Transformative outcomes are not properties of AI capabilities alone. They depend on how AI is developed, deployed, used, and governed.

We have agency in shaping those decisions and hence the outcomes.