RSI（Recursive Self-Improvement, 再帰的自己改善）：汎用AIがもたらすのは人類滅亡かユートピアか？

神楽坂やちま

神楽坂やちま

AGI Hub（AGI/ASIの影響を発信・研究）
AI Safety ??? （ミュトス関連の政府系機関）
SciMeta研究会（RSIの研究会，人工知能学会公認）
AGI社会論研究会（理研の教授級の先生らと運営）

医師 → 起業・売却 → 大学 → 行政・起業

内閣官房幹部へRSIについてレクチャー

『サイエンス・オブ・サイエンス』（森北出版）
科学技術政策系EBPMのバイブル和訳

RSIの基礎技術の世界大会を準備中



RSI（Recursive Self-Improvement, 再帰的自己改善）

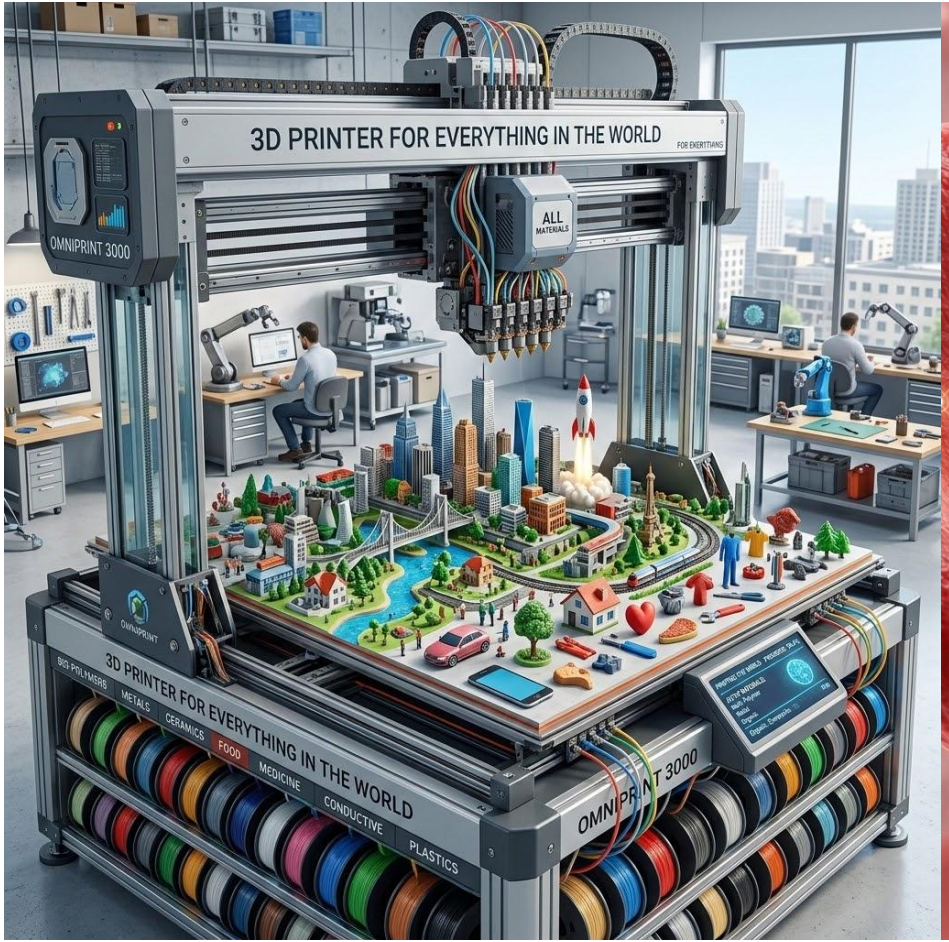
「再帰的自己改善」：AIが自分自身を改良

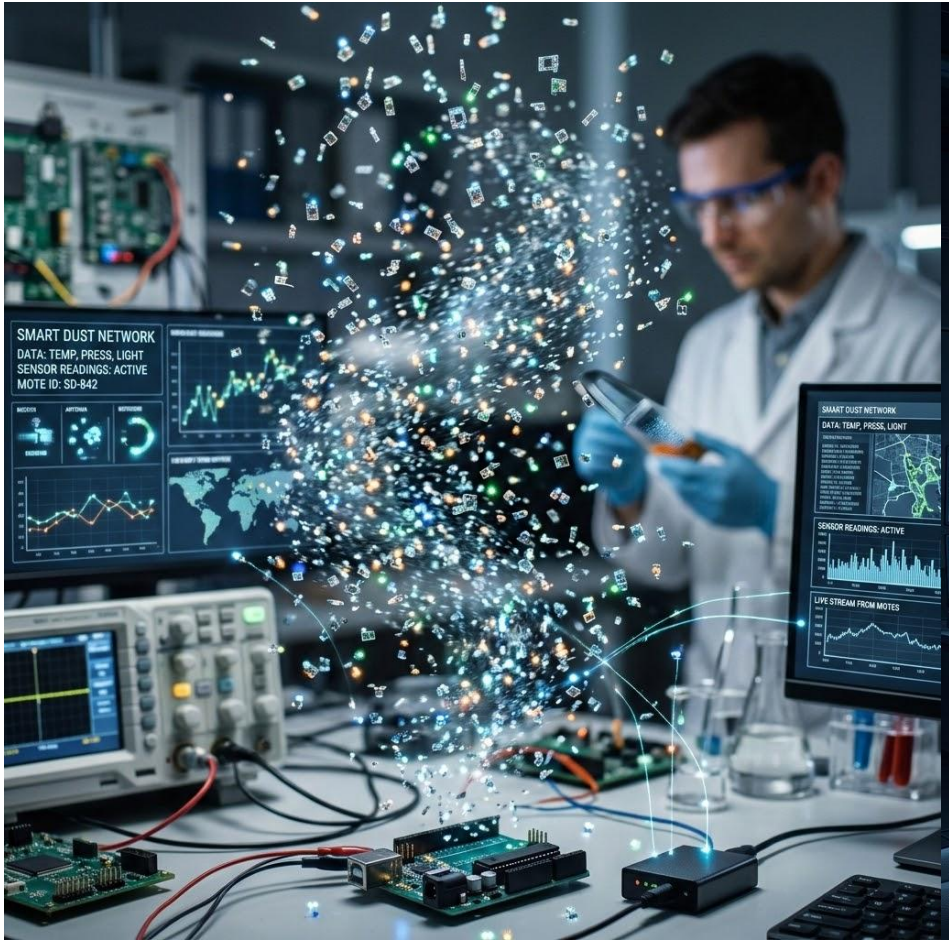
ソフトもハードも含む（後者は今回割愛）
他AIの改善も含みがち（最終的にループが戻れば）

「再帰的自己改善」：自己改善の方法自体も自己改善

性能が頭打ちになっても別の方法を試す
≒　際限なくAIが自己改善

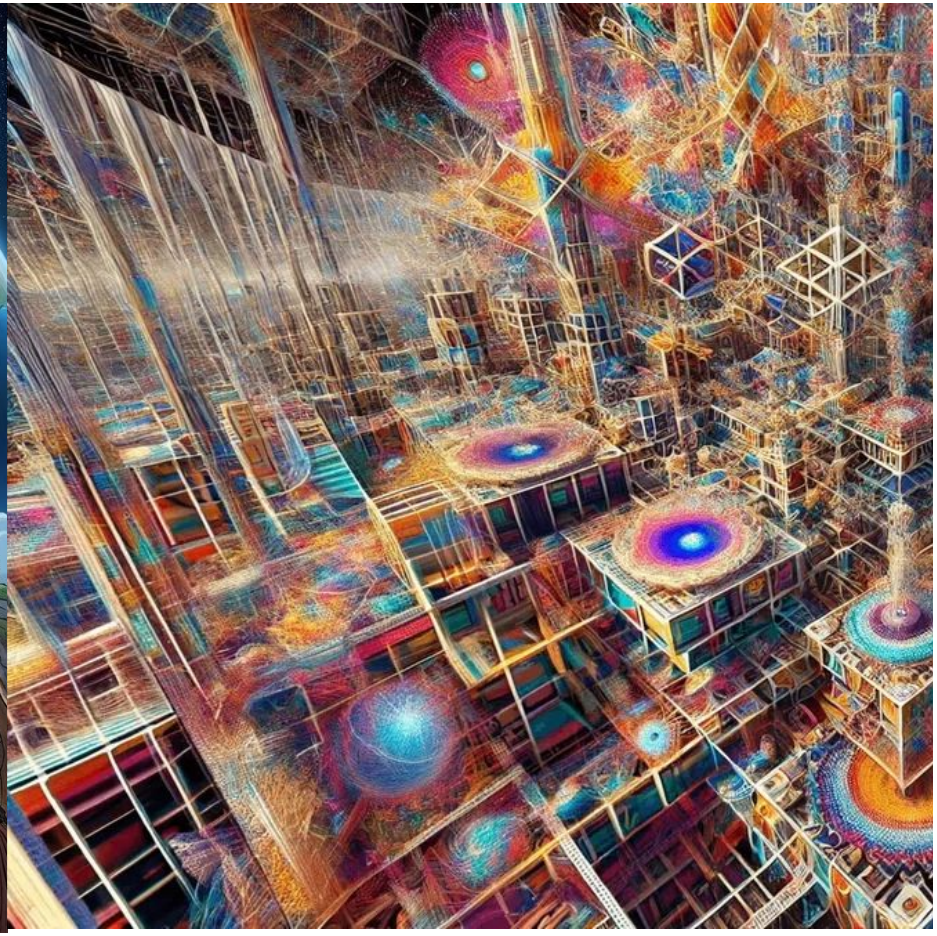



3D PRINTER FOR EVERYTHING IN THE WORLD
OMNIPRINT 3000
ALL MATERIALS
3D PRINTER FOR EVERYTHING IN THE WORLD
BIO-POLYMERS
METALS
CERAMICS
FOOD
MEDICINE
CONDUCTIVE
PLASTICS
OMNIPRINT 3000


SMART DUST NETWORK
DATA: TEMP, PRESS, LIGHT
SENSOR READINGS: ACTIVE
MOTE ID: SD-842
SMART DUST NETWORK
DATA: TEMP, PRESS, LIGHT
SENSOR READINGS: ACTIVE
LIVE STREAM FROM MOTES


NEURAL REPOSITORY - SUBJECT: ELARA VANCE - STATUS: UPLOADING (74%)
NEURAL ACTIVITY: 89%
DATA TRANSFER: 12GB/s
SYNAPTIC MAP
SYNAPTIC MAP


PROTOCELLS
(lipid-bound vesicles, lipid-bound vesicles, RNA within chemical broth)
1. ORIGIN OF LIFE
Simple Protoproliferation. First prokaryotes (bacteria-like in early ocean
c. 3.8 Ga
2. PROKARYOTES
Cyanobacteria, microbial mats beginning (Stromatalites)
Photosynthesis
O_2
O_2
Oxygenation
Billions of Yesrs Ago
Pootaryotes
3. EUKARYOTES
Complex cells with nuclei (endo-symbiosis, mitochondria, chloroplasts)
1.8 Ga
1.0 Ga
4. MULTICELLULARITY
Early simple organisms (Ediacaran biota)
4
Simple Explosion
Oxygenation
5. PALEOZOTES
Cambrian Extiinites, first (†-Pikaia)
375 Ma
Pikaia
6. PALEOZOIC ERA
Dinosaurs (T-Rex, Triceratops, birds), early mammals, flowering plants
Exploision
Hylonomus
Tiktaalik
375 Ma
Billions of Years Ago
3
6. MESOZOIC ERA
Age of Reptiles
Mammalian explosion
Neandethals
(Homo neanderthalencis)
Fish to Tetrapods
Dinosaurs
Morganucodon
Primates
(Procensal)
Dinosatops
Hominin
Hominids
375
3
Billions of Years Ago
4
375
PALEOZOIC ERA
Cambrian Ecdysian
541–252 Ma
5
252
MESOZOIC ERA
252–66 Ma
6
252
66
CENOZOIC ERA
Hominin lineage
7
66 Ma –
Diversification of Mammals
8
2.5 Ma - Present
HUMANITY
MODERN HUMANS
(Homo sapiens)



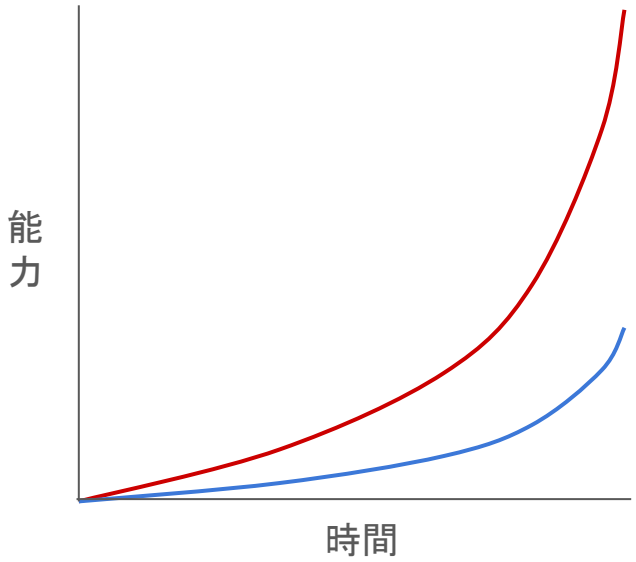
RSIによる（無視できない確率の）リスク①

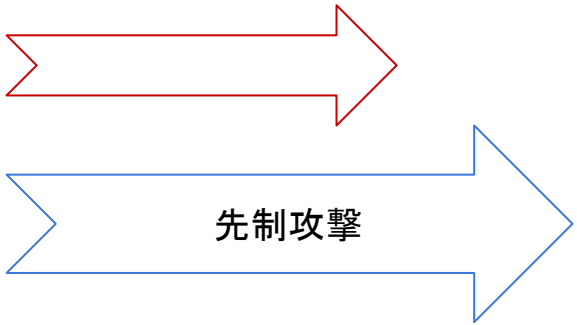




これらの確率を急上昇 & 制御喪失させ得る

RSIによる（無視できない確率の）リスク②





他国に対する圧倒的覇権　and/or　それを予期した先制攻撃

Chan, Alan, et al. "What if automating AI R&D triggers an intelligence explosion?." arXiv preprint arXiv:2609.36054 (2026).

各国の動向

Ban Artificial Superintelligence Act（米国）

会見にて「RSIが規制対象」と明言
閾値を超える超知能の永久禁止，先進 AI開発の一時停止，国際的な超知能禁止への取り組み

人工智能安全治理框架 3.0（中国）

RSIによる技術的暴走は現実のリスクと認識
多層防御を用いた，人類による常時コントロールの維持や，
制御不能リスクについて定期的検証の義務化

大学やシンクタンクからも，RSIに関する政策提案のレポート 複数あり

ケンブリッジ大学，ジョージタウン大学，IFP



地震のように事前準備すべき

RSIの実態

AIエージェントの改善

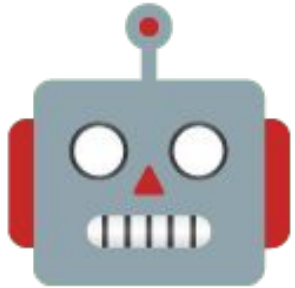
AIの組み合わせ方，ツール選定，プロンプトなど

生成AI等の改善

AI自体の構造やパラメータなど

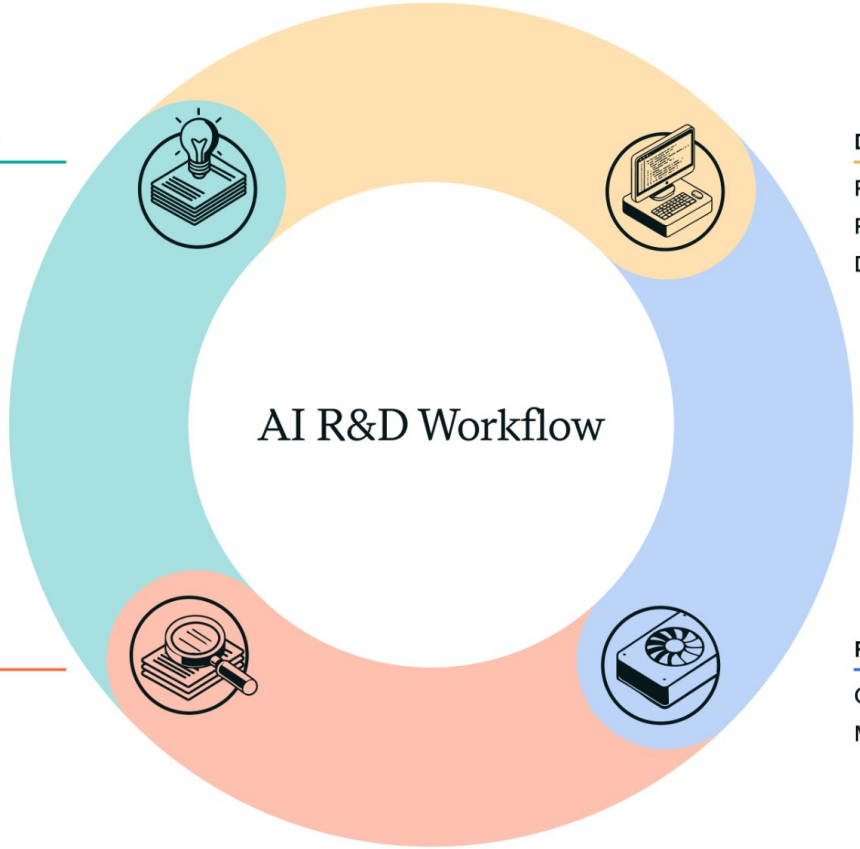
次期AIの開発

OpenAI等の開発者をAIで置き換えるイメージ



（下に行くほど一般に時間がかかりやすい）

EPOCH AI



CREATING HYPOTHESES

High-level planning

Detailed planning

DESIGNING EXPERIMENTS

Prototype engineering

Performance engineering

Debugging

ANALYZING RESULTS

Analyzing results

RUNNING EXPERIMENTS

Organizing distributed training

Monitoring training runs

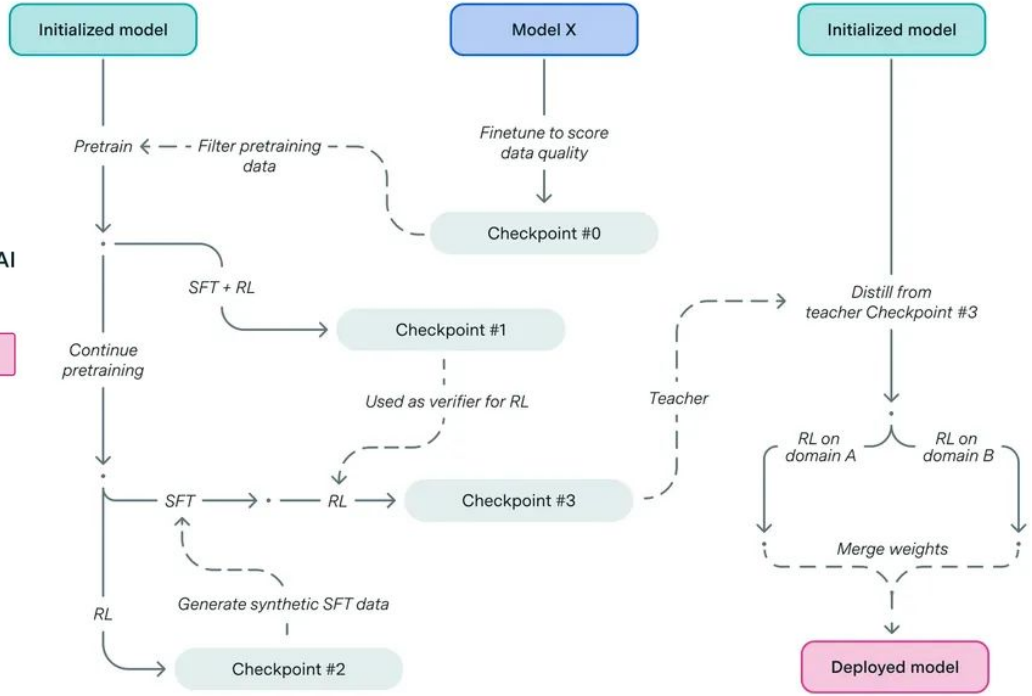
CC-BY

EXPECTATION



EPOCH AI

REALITY



CC-BY

RSIを目指したAIの実例

AlphaEvolve (Google DeepMind, ‘25/05)

Geminiを使ったAIエージェントが，Gemini自身の一部を改善
他にもTPUの設計などに実践投入

Darwin Gödel Machine (Sakana AI, ‘25/05)

2週間継続して自己改善

Group-Evolving Agents (カリフォルニア大学, ‘26/02)

人間が設計したAIエージェントを超越

ただし実例を見ても**後追い**にしかならない



New · Recursive Self-Improvement Delivers New SOTA

DISCOVERY LOOP

CONTINUOUS EXPLORATION

Automating discovery to accelerate science and engineering for the world.

poetiq

inherent

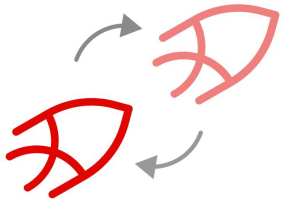
Recursively self-improving to discover new knowledge.

recursive

Recursive self-improving superintelligence to automate knowledge discovery.

When AI builds itself

Our progress toward recursive self-improvement, and its implications.



RSI Lab
sakana.ai

Meta

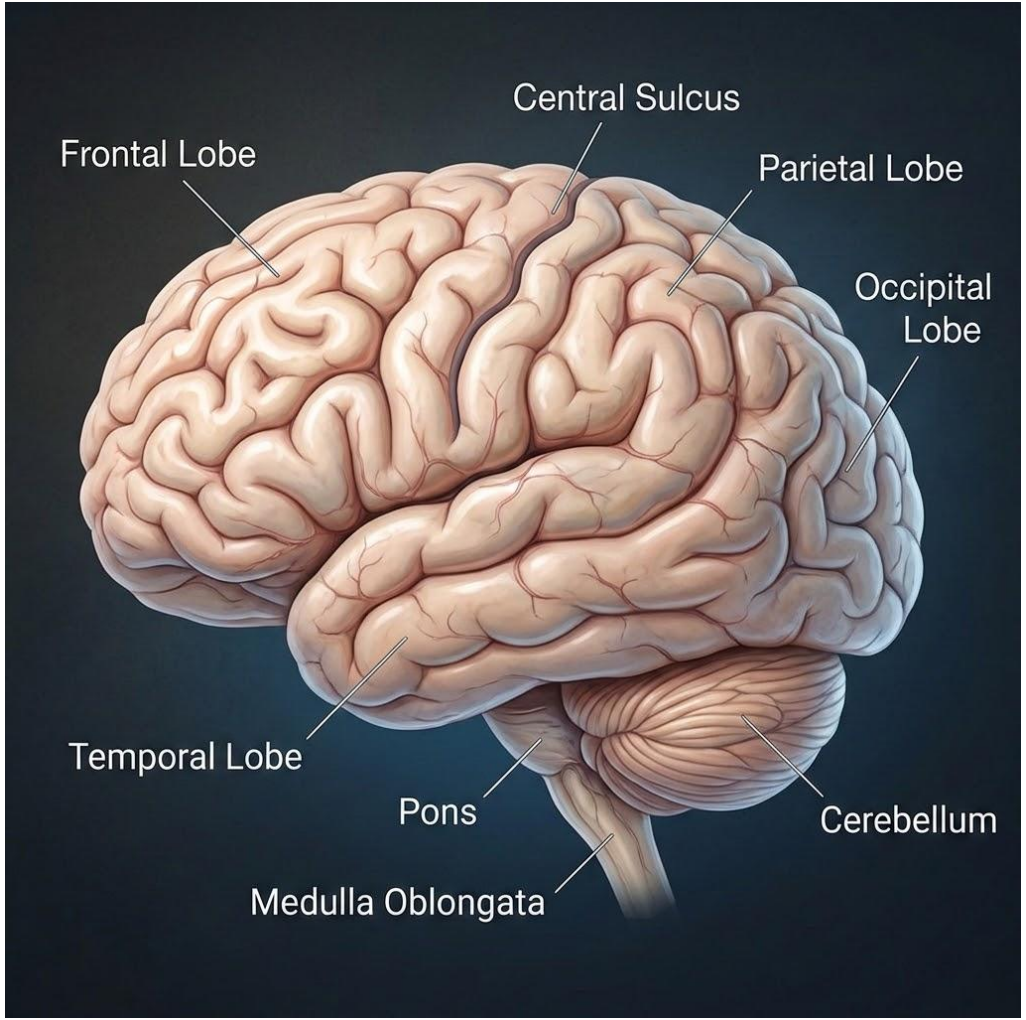
DeepMind

OpenAI

RSIは爆発的となる恐れを否定できず
事前の予測が必要

そもそもAIは際限なく成長

まず人間相当のAIができるのか


Central Sulcus
Frontal Lobe
Parietal Lobe
Occipital Lobe
Temporal Lobe
Pons
Cerebellum
Medulla Oblongata

人間と同等のAIはいつできる？

最先端のAIが学んだ量

10^{27} FLOP [Sevilla+ '22]

GPT-6 Astra
関係者の発言内容などから概算

成人が学んだ量

6.3×10^{26} FLOP

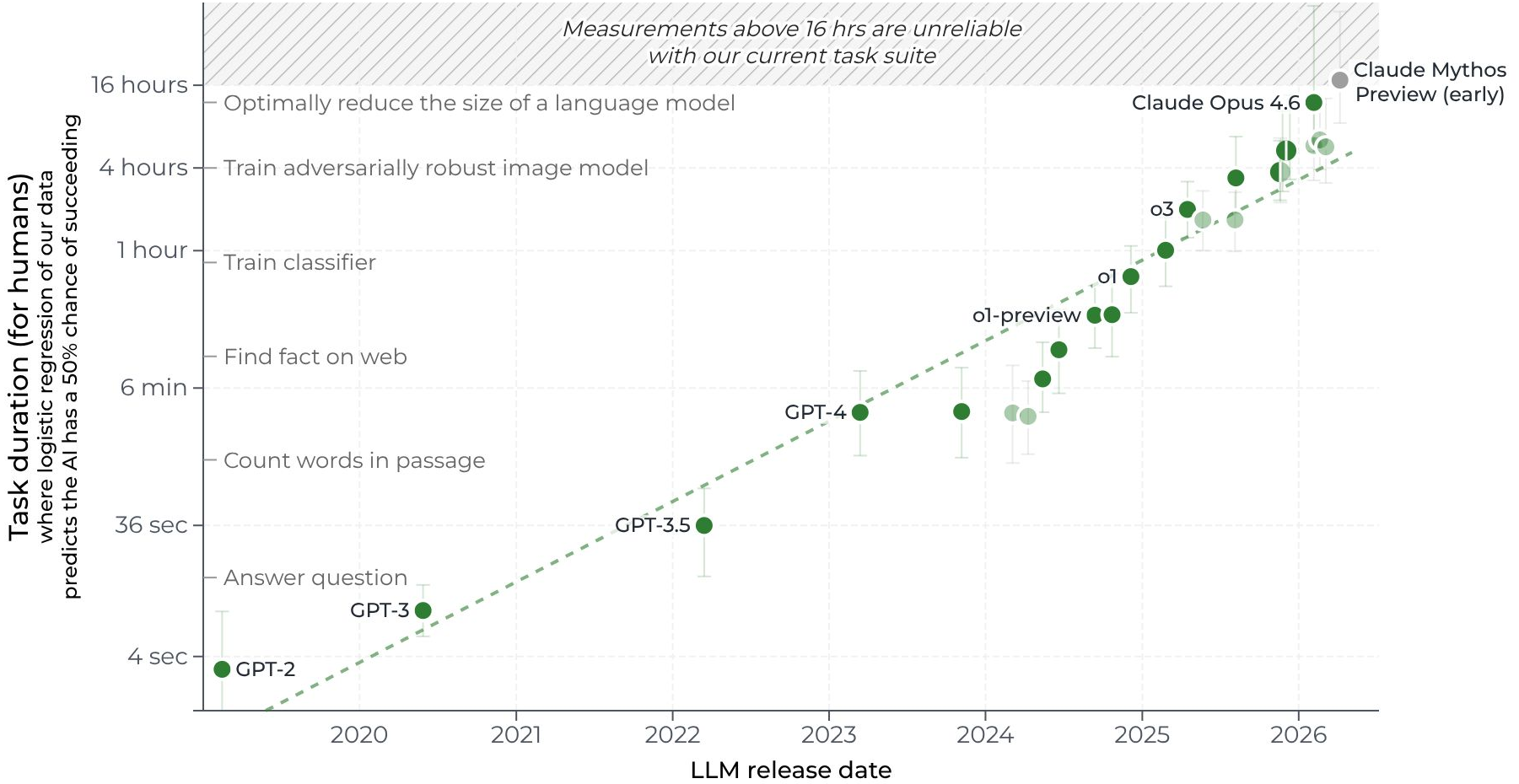
= 脳の計算速度 10^{18} FLOP/秒 [AI Impacts '19]
× 20年間 (6.3×10^{8} 秒)

Sevilla, Jaime, et al. "Compute trends across three eras of machine learning." 2022 international joint conference on neural networks (IJCNN). IEEE, 2022.
“Brain Performance in FLOPS.” AI Impacts, 7 July 2019, aiimpacts.org/brain-performance-in-flops.

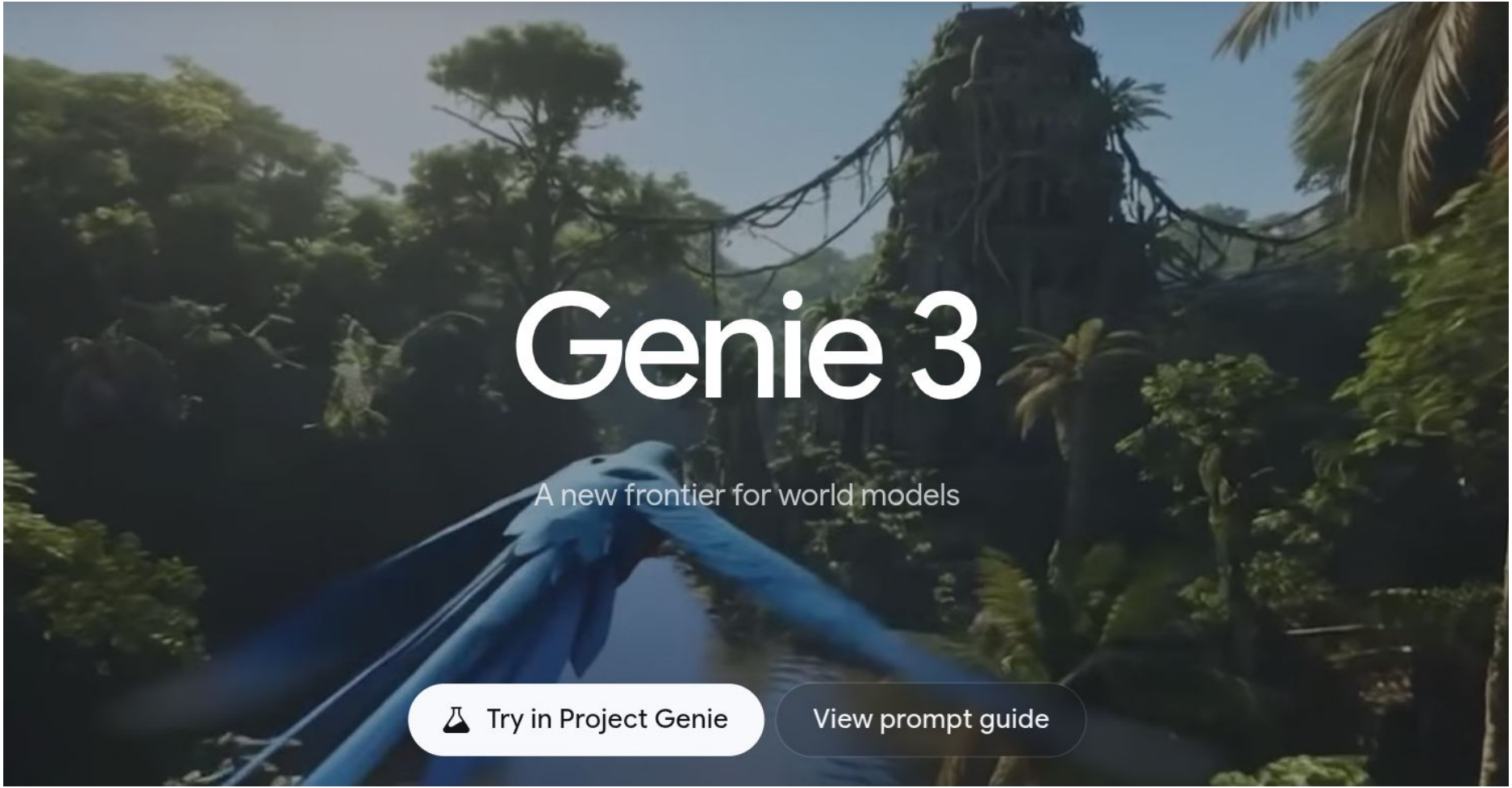
長いタスク

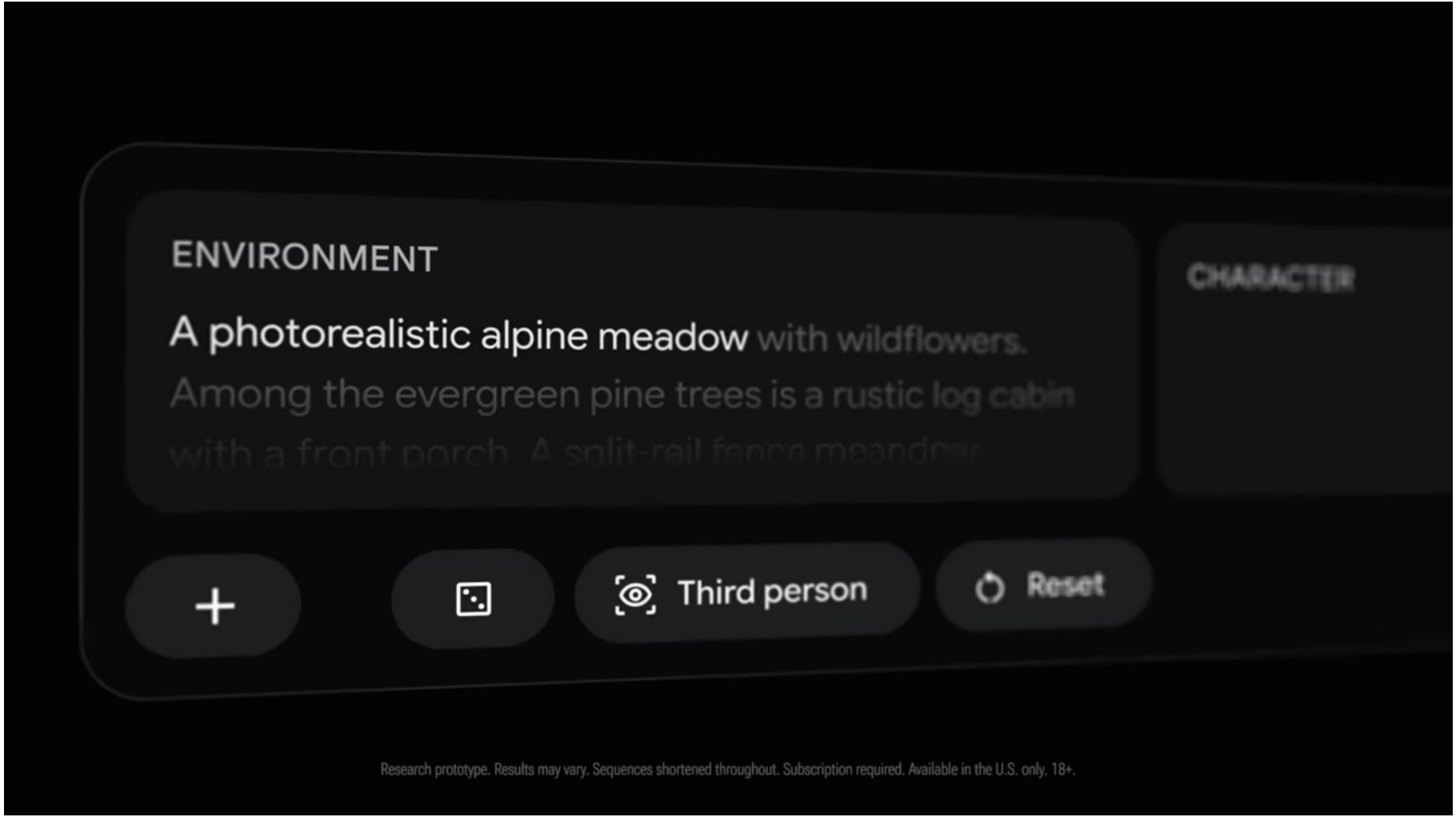
Time horizon of software tasks
different LLMs can complete 50% of the time

METR

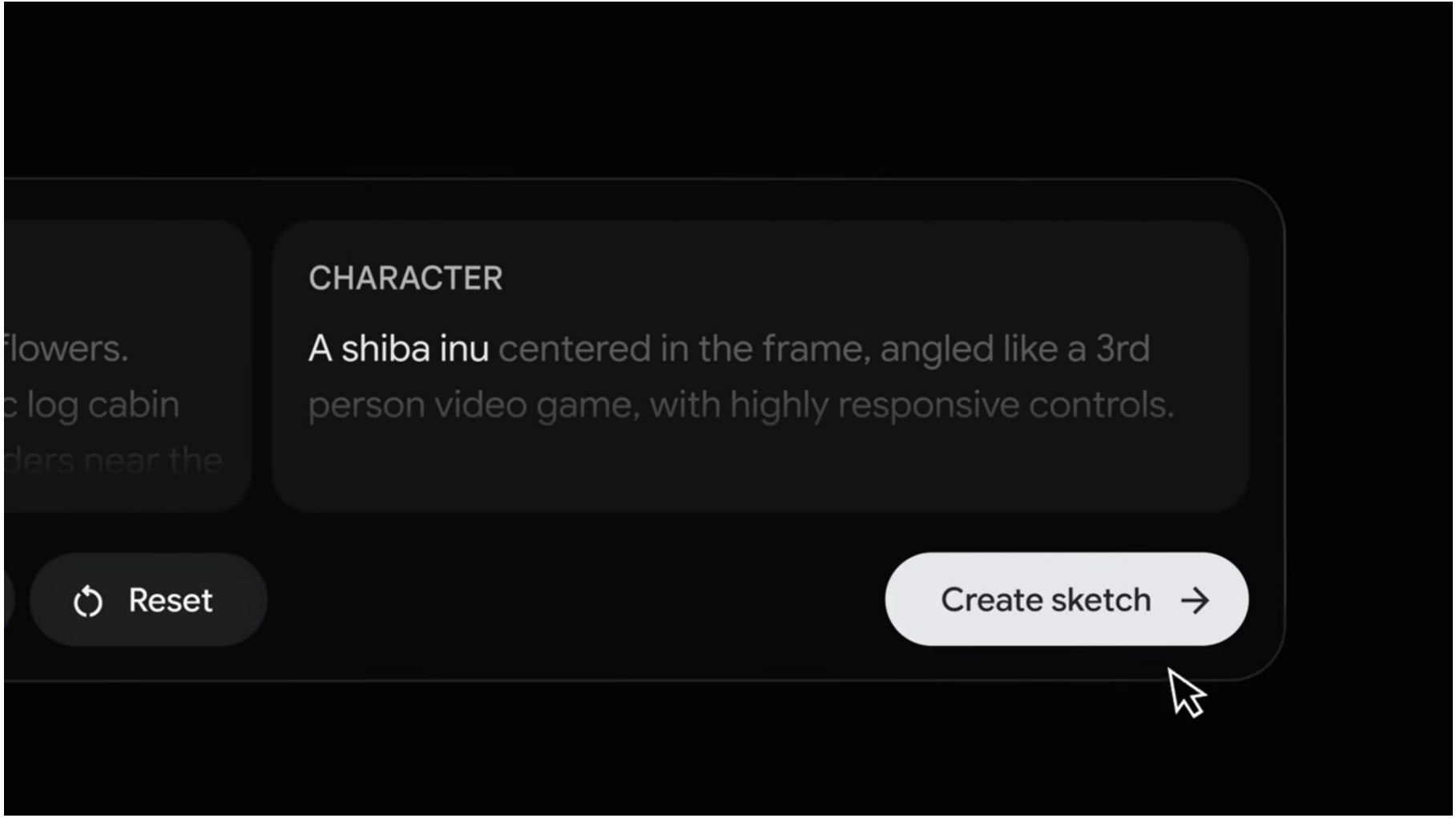


身体性・空間

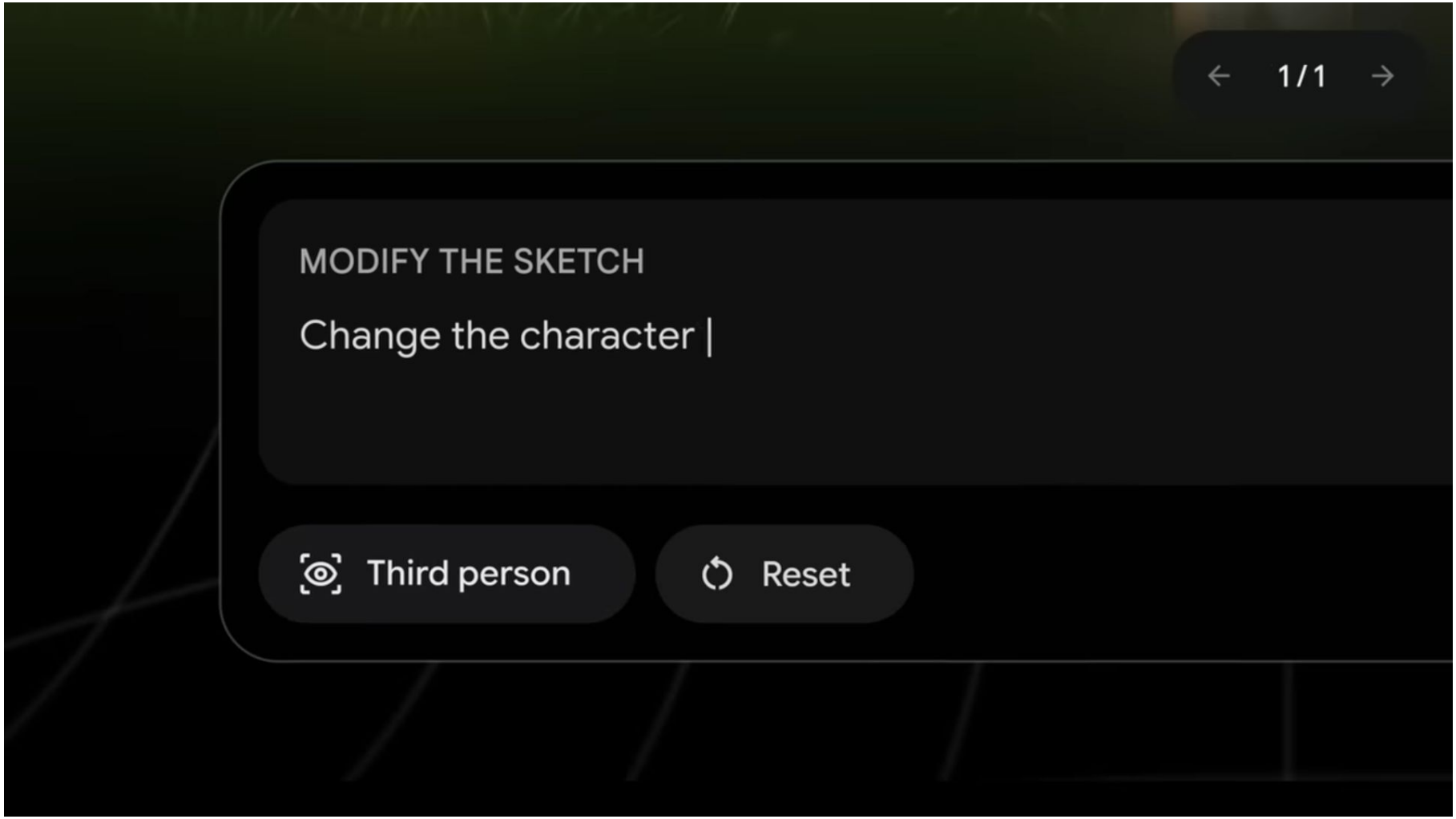

Genie 3
A new frontier for world models
Try in Project Genie
View prompt guide

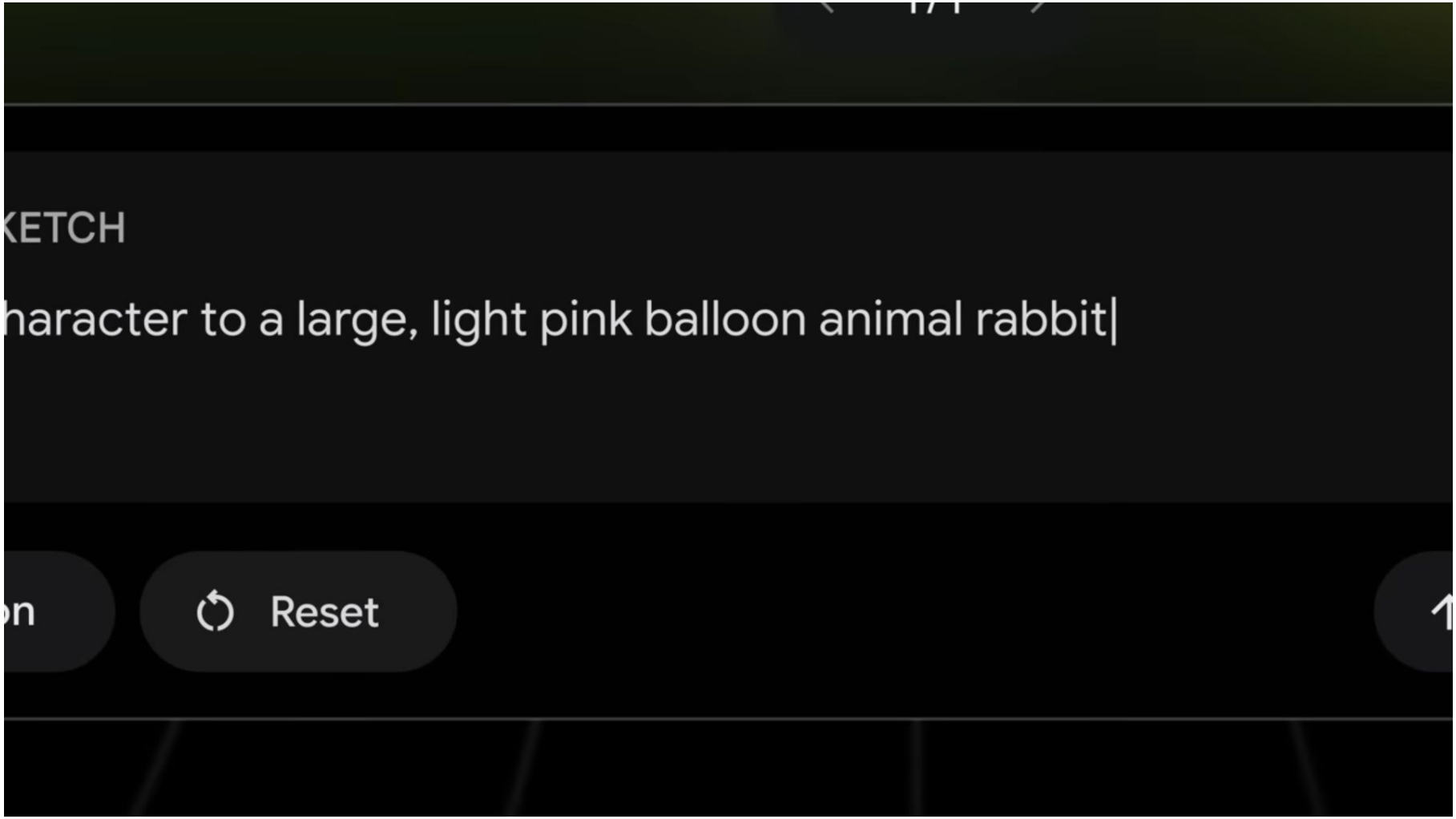


Research prototype. Results may vary. Sequences shortened throughout. Subscription required. Available in the U.S. only. 18+.


flowers.
c log cabin
ders near the
CHARACTER
A shiba inu centered in the frame, angled like a 3rd person video game, with highly responsive controls.
Reset
Create sketch


W
A
S
D
SPACE


1/1
MODIFY THE SKETCH
Change the character |
Third person
Reset


KETCH
haracter to a large, light pink balloon animal rabbit
on
Reset


pink balloon animal rabbit
W
A
S
D
SPACE

因果

AI開発とは

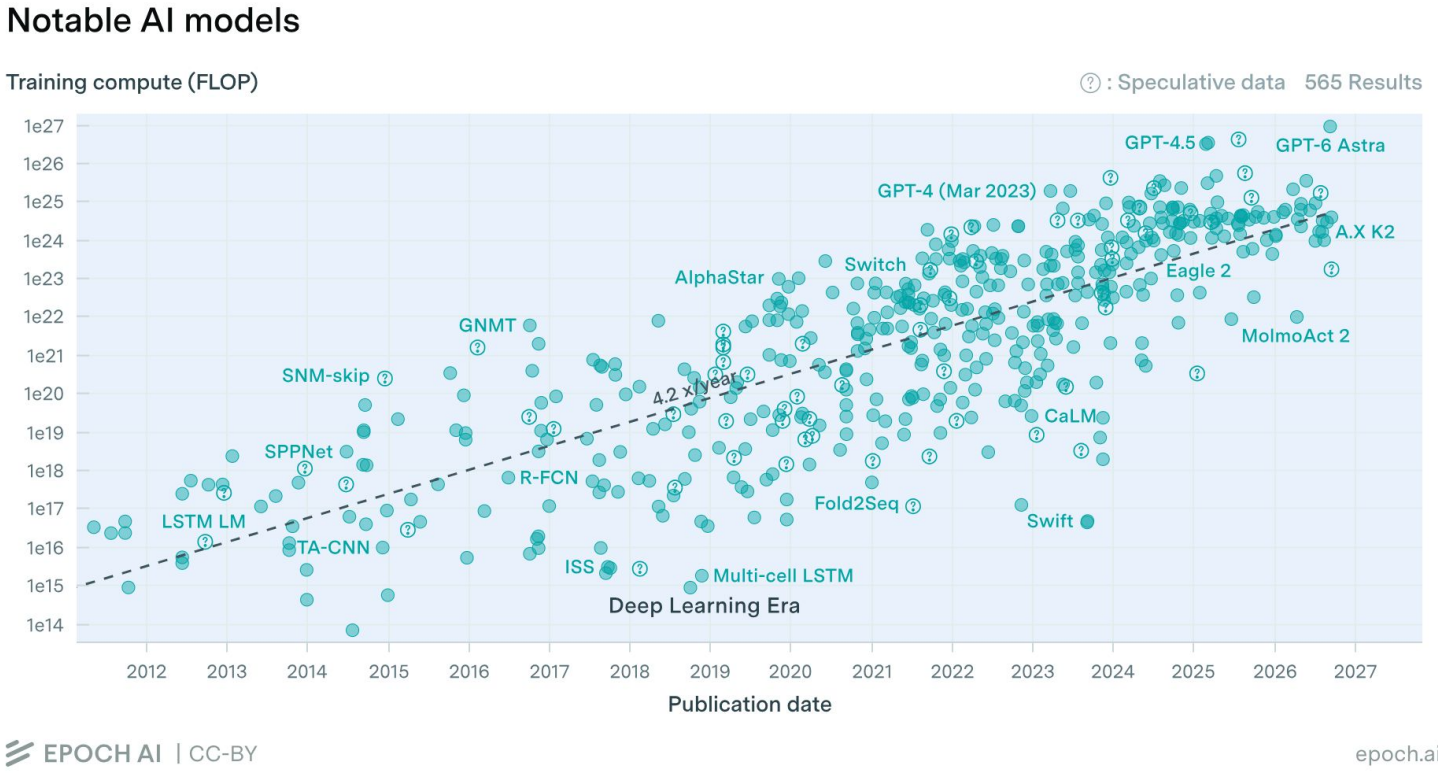


あらゆる概念の地図を作るようなもの

AIの能力は急速に向上

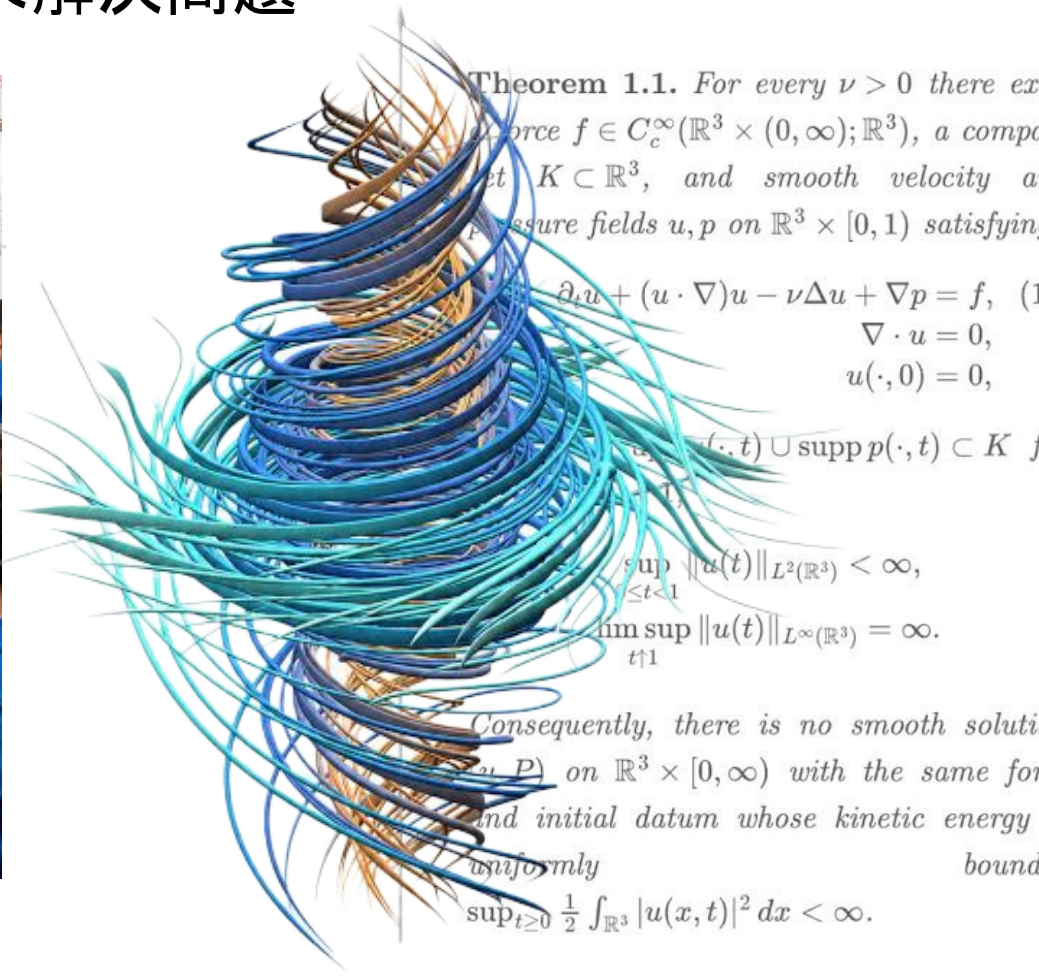
5年で1000倍

2年で1000倍との説も
(Anthropic's RSP)



5年前は小学校，それが今や未解決問題





Theorem 1.1. *For every $\nu > 0$ there ex*
orce $f \in C_c^\infty(\mathbb{R}^3 \times (0,\infty); \mathbb{R}^3)$, a compa
et $K \subset \mathbb{R}^3$, and smooth velocity a
ssure fields u, p on $\mathbb{R}^3 \times [0,1)$ satisfyin

$$\partial_t u + (u \cdot \nabla)u - \nu \Delta u + \nabla p = f, \quad (1$$
$$\nabla \cdot u = 0,$$
$$u(\cdot, 0) = 0,$$

$$(\cdot, t) \cup \operatorname{supp} p(\cdot, t) \subset K \quad f$$

$$\sup_{\leq t<1} \|u(t)\|_{L^2(\mathbb{R}^3)} < \infty,$$
$$\limsup_{t \uparrow 1} \|u(t)\|_{L^\infty(\mathbb{R}^3)} = \infty.$$

Consequently, there is no smooth soluti
P) on $\mathbb{R}^3 \times [0,\infty)$ with the same fo
nd initial datum whose kinetic energy
uniformly *bound*
$\sup_{t \geq 0} \frac{1}{2} \int_{\mathbb{R}^3} |u(x,t)|^2 \, dx < \infty.$

なぜ際限なく成長するのか

ソフトウェア

この世にない問題が苦手になっている [Goldblum+ 2304.05366]

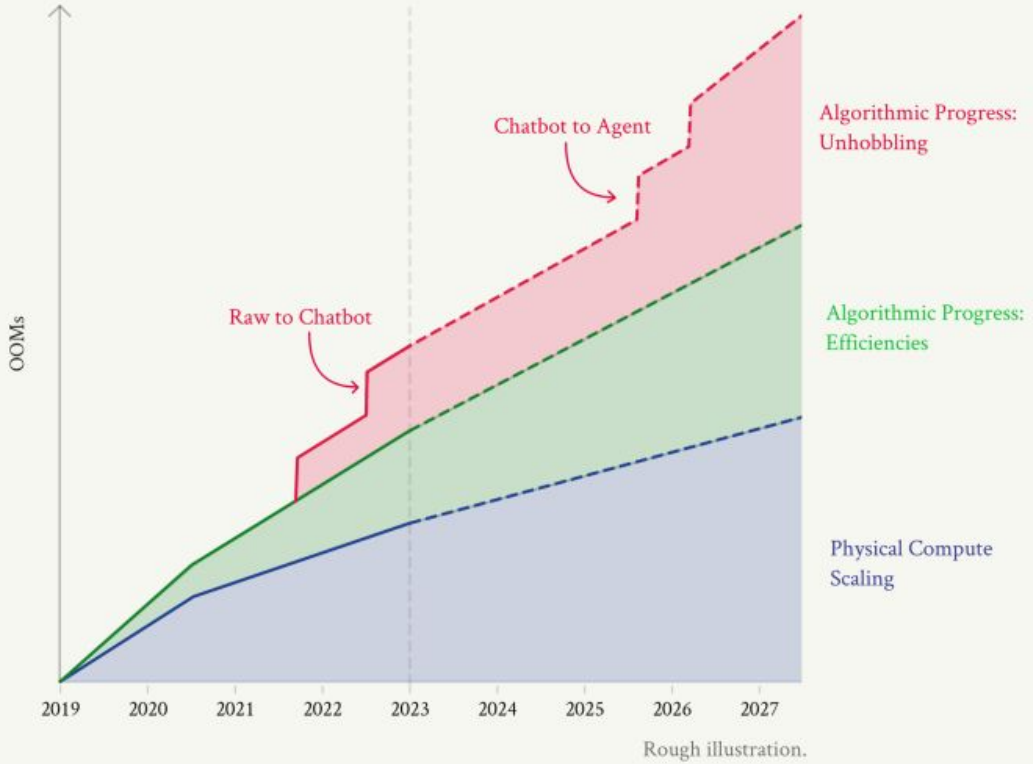
ハードウェア

何でも計算機になる [Piccinini 2015]



Goldblum, Micah, et al. "The no free lunch theorem, kolmogorov complexity, and the role of inductive biases in machine learning." arXiv preprint arXiv:2304.05366 (2023).
Piccinini, Gualtiero, Physical Computation: A Mechanistic Account (Oxford, 2015; online edn, Oxford Academic, 20 Aug. 2015), https://doi.org/10.1093/acprof:oso/9780199658855.001.0001, accessed 11 Aug. 2026.

Decomposing drivers of progress



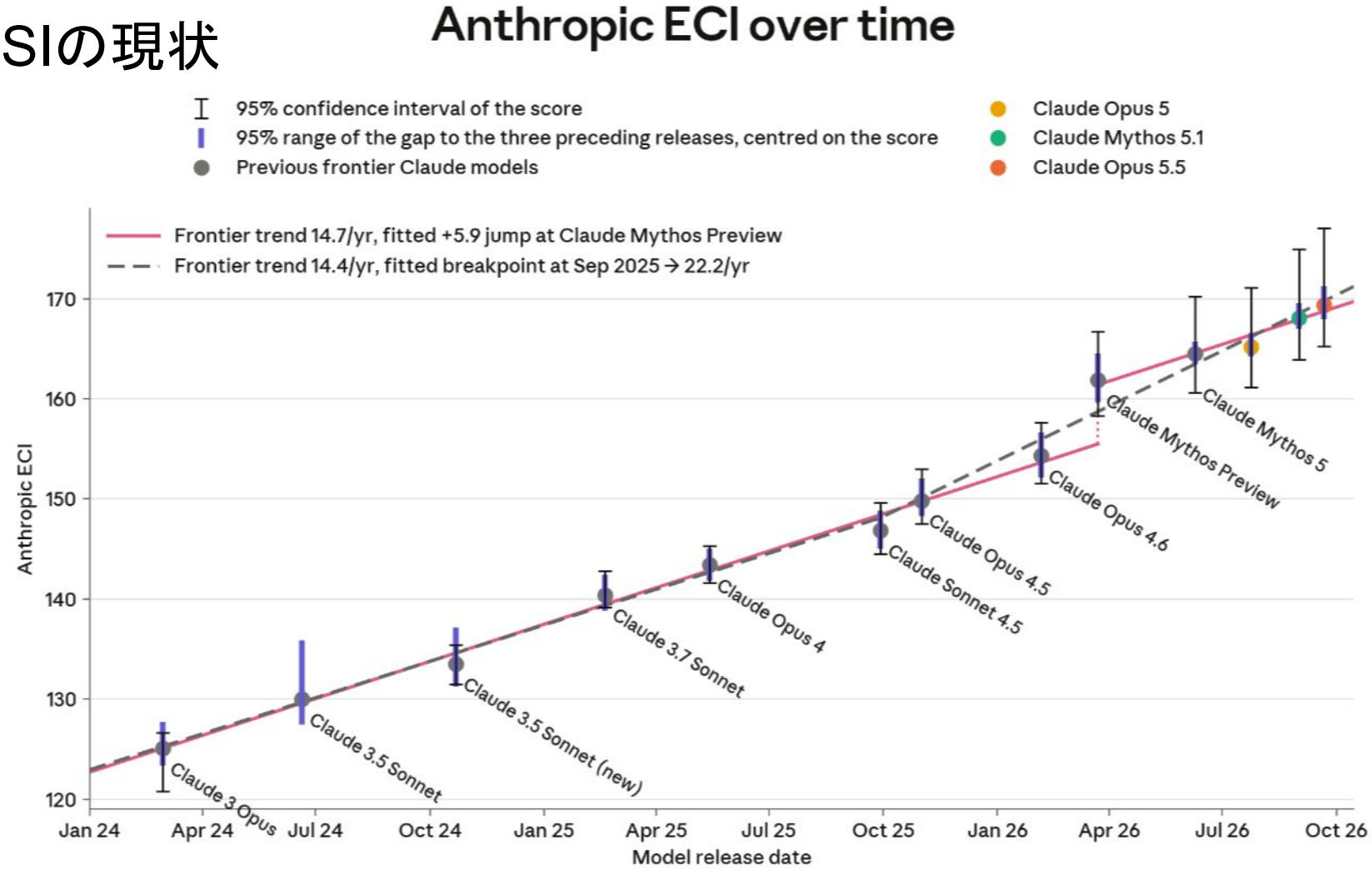
SITUATIONAL AWARENESS | Leopold Aschenbrenner

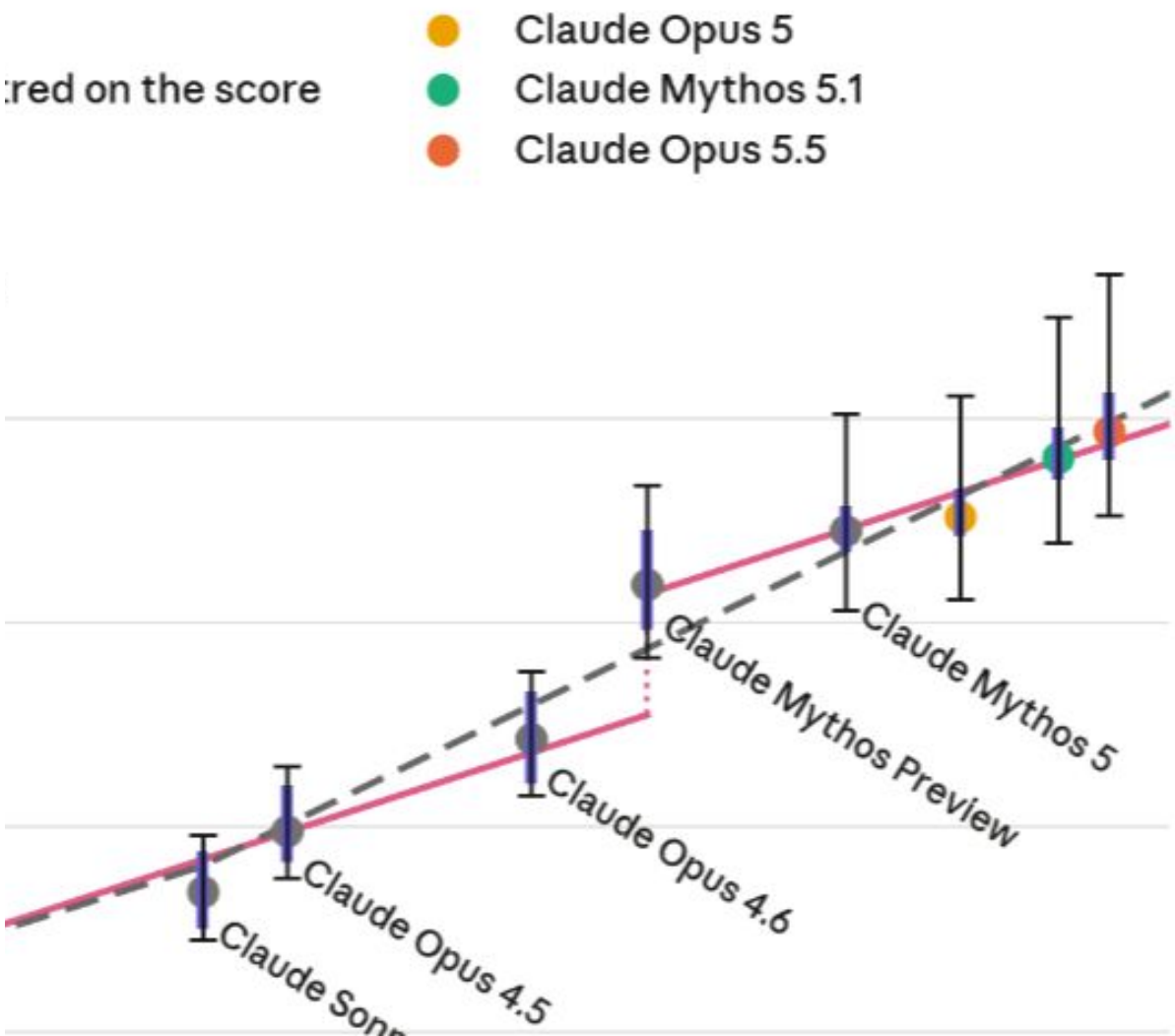


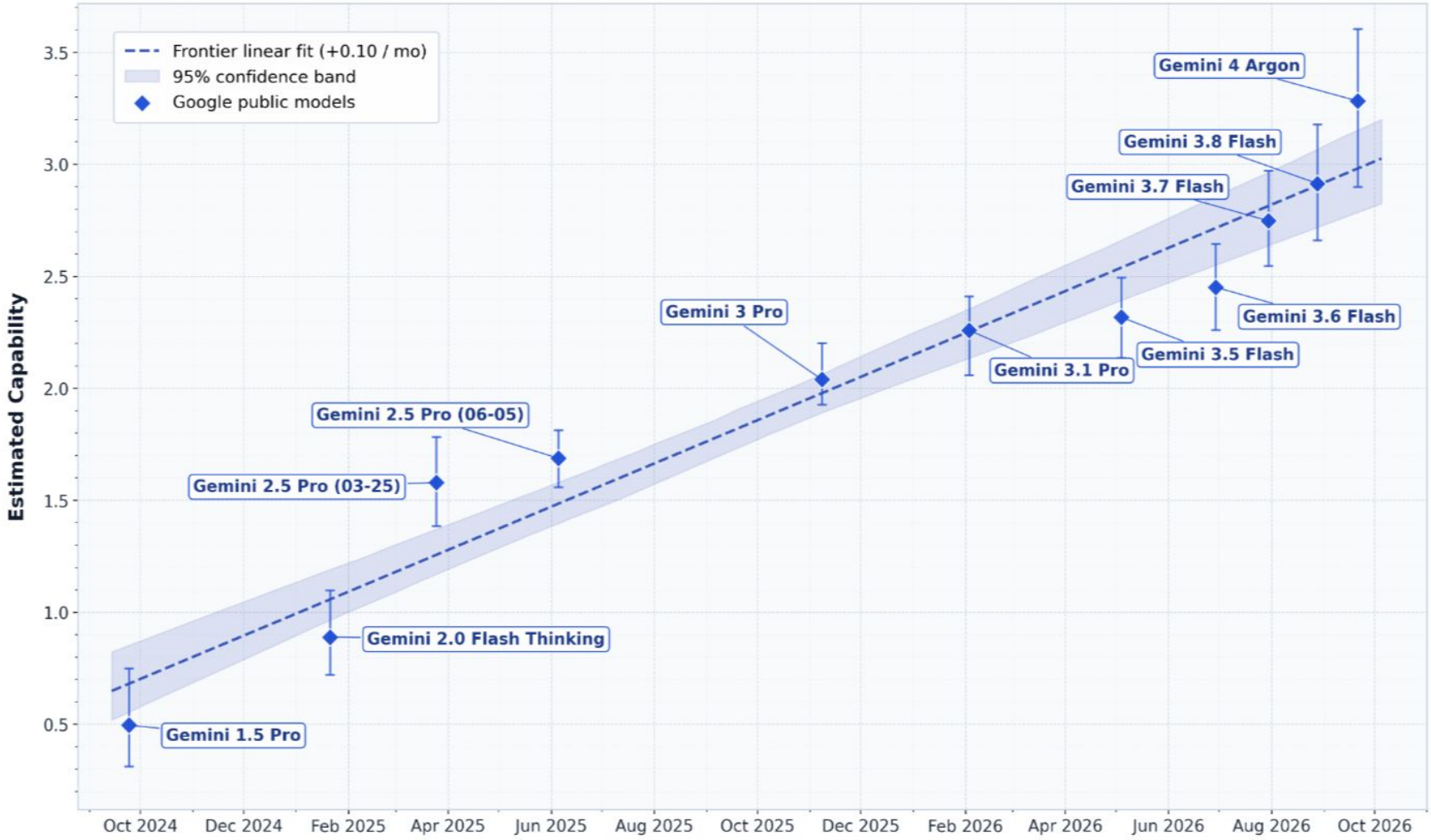
RSIによって更に加速

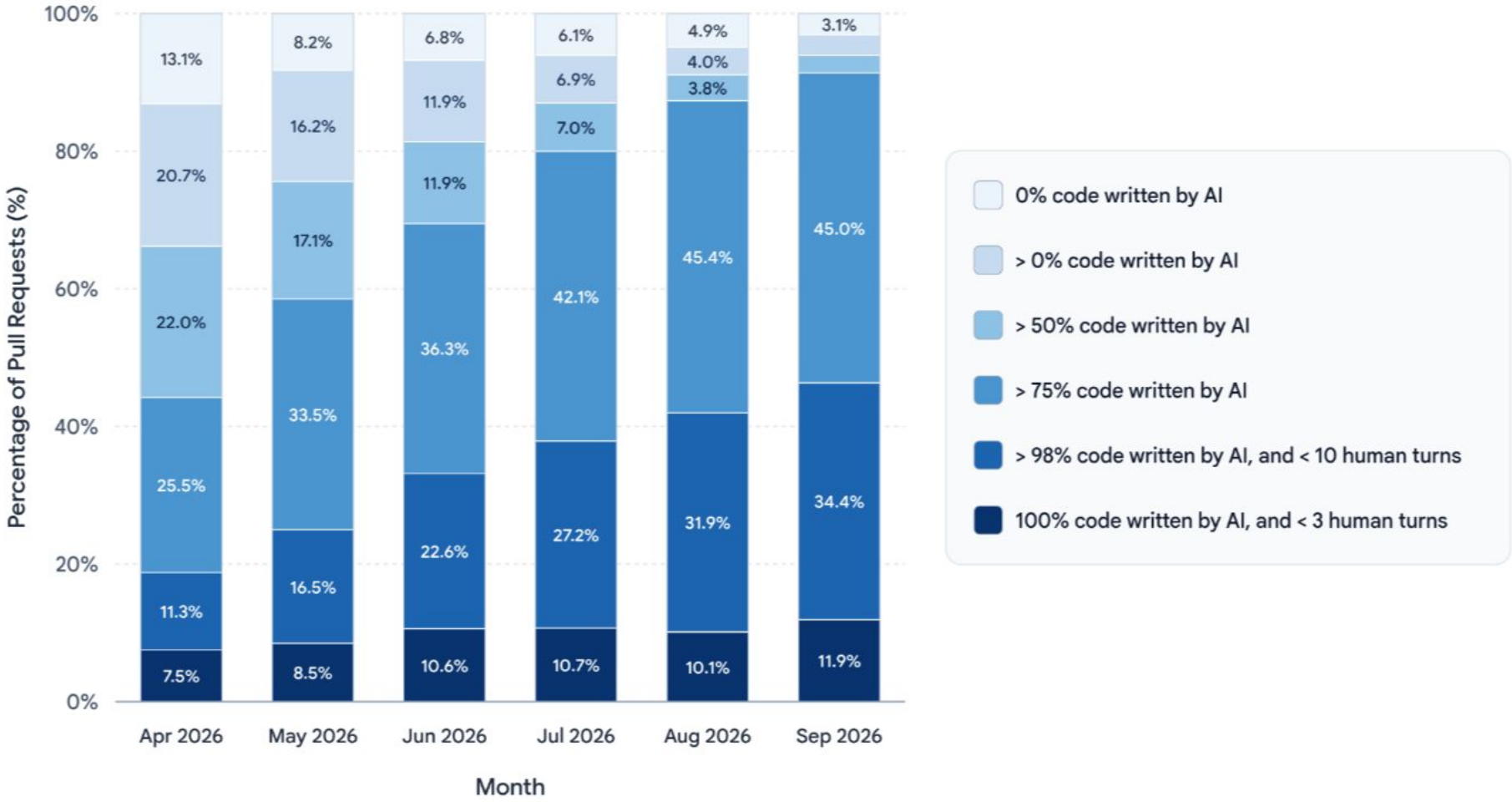
RSIは今にも始まる可能性

RSIの現状


Anthropic ECI over time
95% confidence interval of the score
95% range of the gap to the three preceding releases, centred on the score
Previous frontier Claude models
Claude Opus 5
Claude Mythos 5.1
Claude Opus 5.5
Frontier trend 14.7/yr, fitted +5.9 jump at Claude Mythos Preview
Frontier trend 14.4/yr, fitted breakpoint at Sep 2025 → 22.2/yr
Anthropic ECI
120
130
140
150
160
170
Jan 24
Apr 24
Jul 24
Oct 24
Jan 25
Apr 25
Jul 25
Oct 25
Jan 26
Apr 26
Jul 26
Oct 26
Model release date
Claude 3 Opus
Claude 3.5 Sonnet
Claude 3.5 Sonnet (new)
Claude 3.7 Sonnet
Claude Opus 4
Claude Sonnet 4.5
Claude Opus 4.5
Claude Opus 4.6
Claude Mythos Preview
Claude Mythos 5

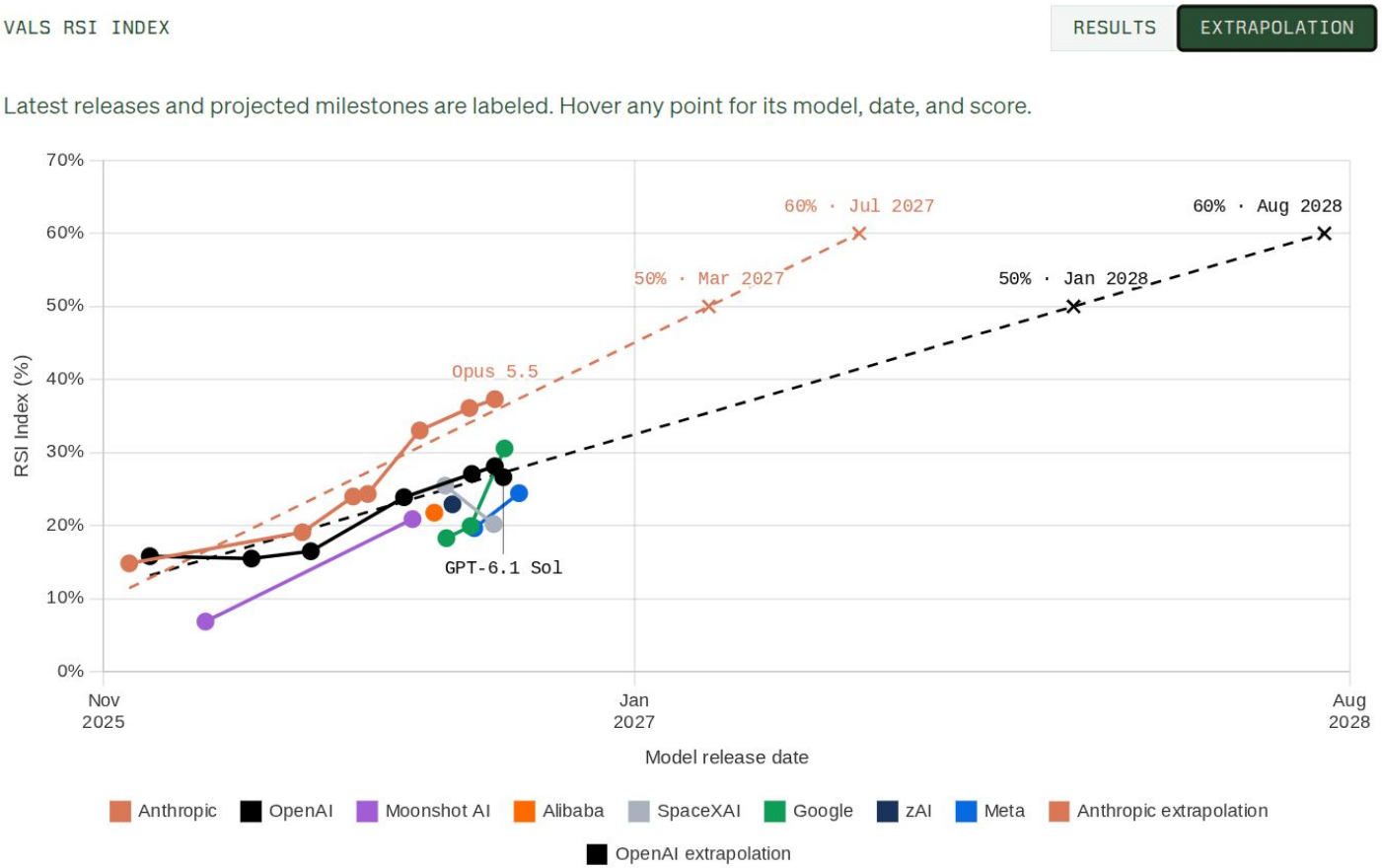

:red on the score
Claude Opus 5
Claude Mythos 5.1
Claude Opus 5.5
Claude Mythos 5
Claude Mythos Preview
Claude Opus 4.6
Claude Opus 4.5


Frontier linear fit (+0.10 / mo)
95% confidence band
Google public models
Estimated Capability
3.5
3.0
2.5
2.0
1.5
1.0
0.5
Gemini 4 Argon
Gemini 3.8 Flash
Gemini 3.7 Flash
Gemini 3.6 Flash
Gemini 3.5 Flash
Gemini 3.1 Pro
Gemini 3 Pro
Gemini 2.5 Pro (06-05)
Gemini 2.5 Pro (03-25)
Gemini 2.0 Flash Thinking
Gemini 1.5 Pro
Oct 2024
Dec 2024
Feb 2025
Apr 2025
Jun 2025
Aug 2025
Oct 2025
Dec 2025
Feb 2026
Apr 2026
Jun 2026
Aug 2026
Oct 2026


100%
80%
60%
40%
20%
0%
Percentage of Pull Requests (%)
13.1%
20.7%
22.0%
25.5%
11.3%
7.5%
8.2%
16.2%
17.1%
33.5%
16.5%
8.5%
6.8%
11.9%
11.9%
36.3%
22.6%
10.6%
6.1%
6.9%
7.0%
42.1%
27.2%
10.7%
4.9%
4.0%
3.8%
45.4%
31.9%
10.1%
3.1%
45.0%
34.4%
11.9%
Apr 2026
May 2026
Jun 2026
Jul 2026
Aug 2026
Sep 2026
Month
0% code written by AI
> 0% code written by AI
> 50% code written by AI
> 75% code written by AI
> 98% code written by AI, and < 10 human turns
100% code written by AI, and < 3 human turns

AI研究開発の進捗

AIエージェント
なら更に
1年程度前倒し



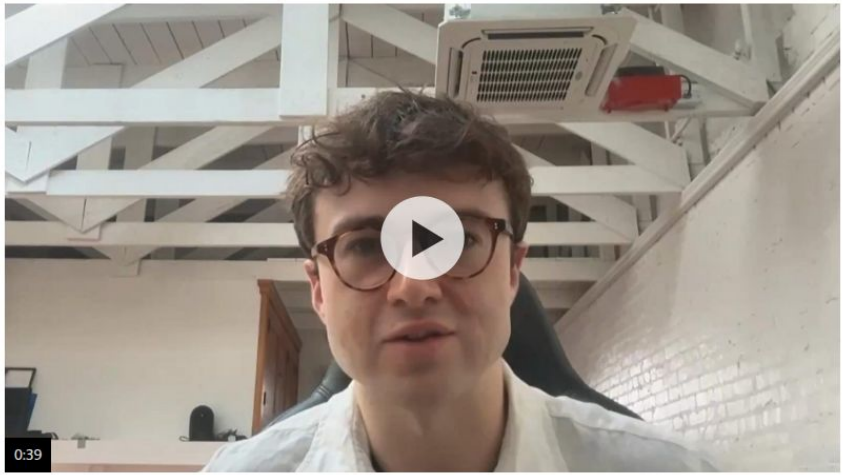
命を賭け，自己改善型超知能へ突き進んでいる (Jacob Coxon)

AI staff 'genuinely frightened' for humanity's future, ex-Anthropic researcher tells BBC

13 September 2026

Share Save Add as preferred on Google

Brandon Livesay and Mitchell Labiak, business reporter



'There is a possibility of human extinction,' former Anthropic employee says about AI

An artificial intelligence researcher who quit AI firm Anthropic has said people working on the technology were "genuinely frightened" about the speed of its advancements and what it could mean for humanity.

急激な成長 ≒ 人類滅亡

急激に成長するかは要調査

急上昇するか否かの論点

（上がっていくのは当たり前で，ボトルネック次第）

- 収穫逓減
- 計算資源
- データ
- 自動化しにくいタスク
- 時間のかかる工程



Chan, Alan, et al. "What if automating AI R&D triggers an intelligence explosion?." arXiv preprint arXiv:2609.36054 (2026).

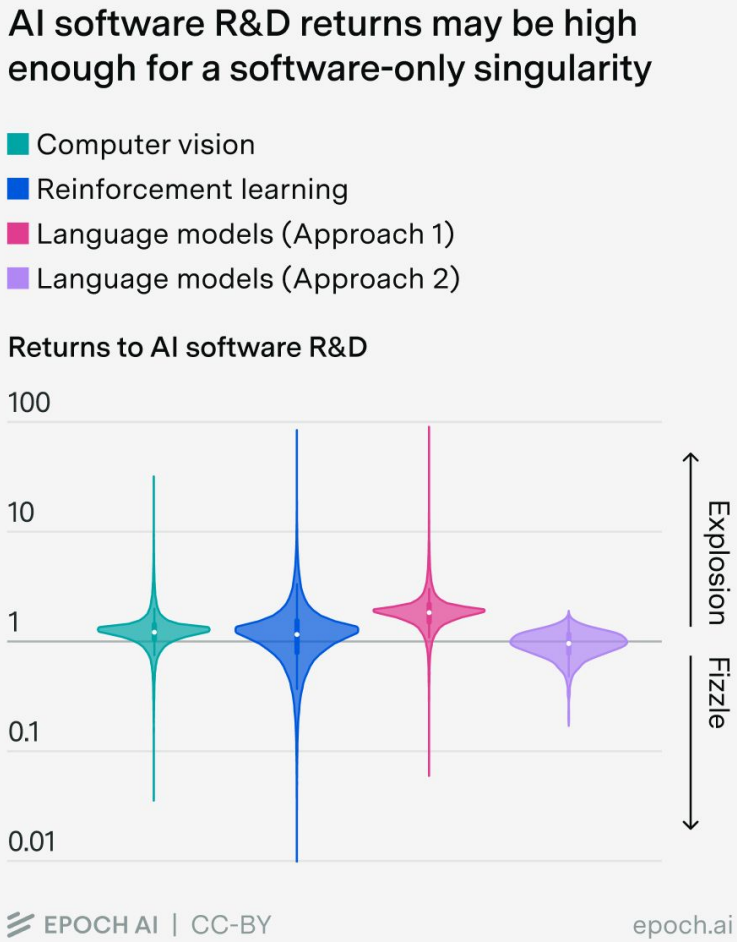
論点①：収穫逓減

研究開発は難しくなっていく

例：ニュートン力学より量子力学のほうが難しい

AIの進展とどっちが上回るか？

進展のほうが速いという研究結果はあるが，
不確実性が大きい



Anson Ho and Parker Whitfill (2025), "The software intelligence explosion debate needs experiments". Published online at epoch.ai.

論点②：計算資源

AIが大規模化するにつれ，研究開発のための計算資源も上昇

（計算資源 ≒ コンピュータで計算する回数）

どちらが上回るか？

直接的な研究結果はまだないと思われる

理論的にも決着つかず



よくわからないなら,
とりあえずAI開発停止する？

AI開発停止のリスク（一部）

- むしろ逆に開発再開時に爆発的成長となる恐れ

  停止中に基礎技術が蓄積
- （一国の場合）他国に遅れをとる
- （多国の場合）抜け駆けの恐れ
- 管理困難な領域での開発の促進
- 「狼少年」化



対策

- （RSIの基礎技術である）AI R&D Automationの進展状況の可視化
- 政策によるRSIの制御

  Hardwired Pause（特定AIのみ推論可能なチップ．高有効性かつ現実的）

  （性能上限設定，開発環境の隔離など）
- 社会側の事前準備

  （AIを用いた防御，重要インフラへの投資など）
- 国家間の協力

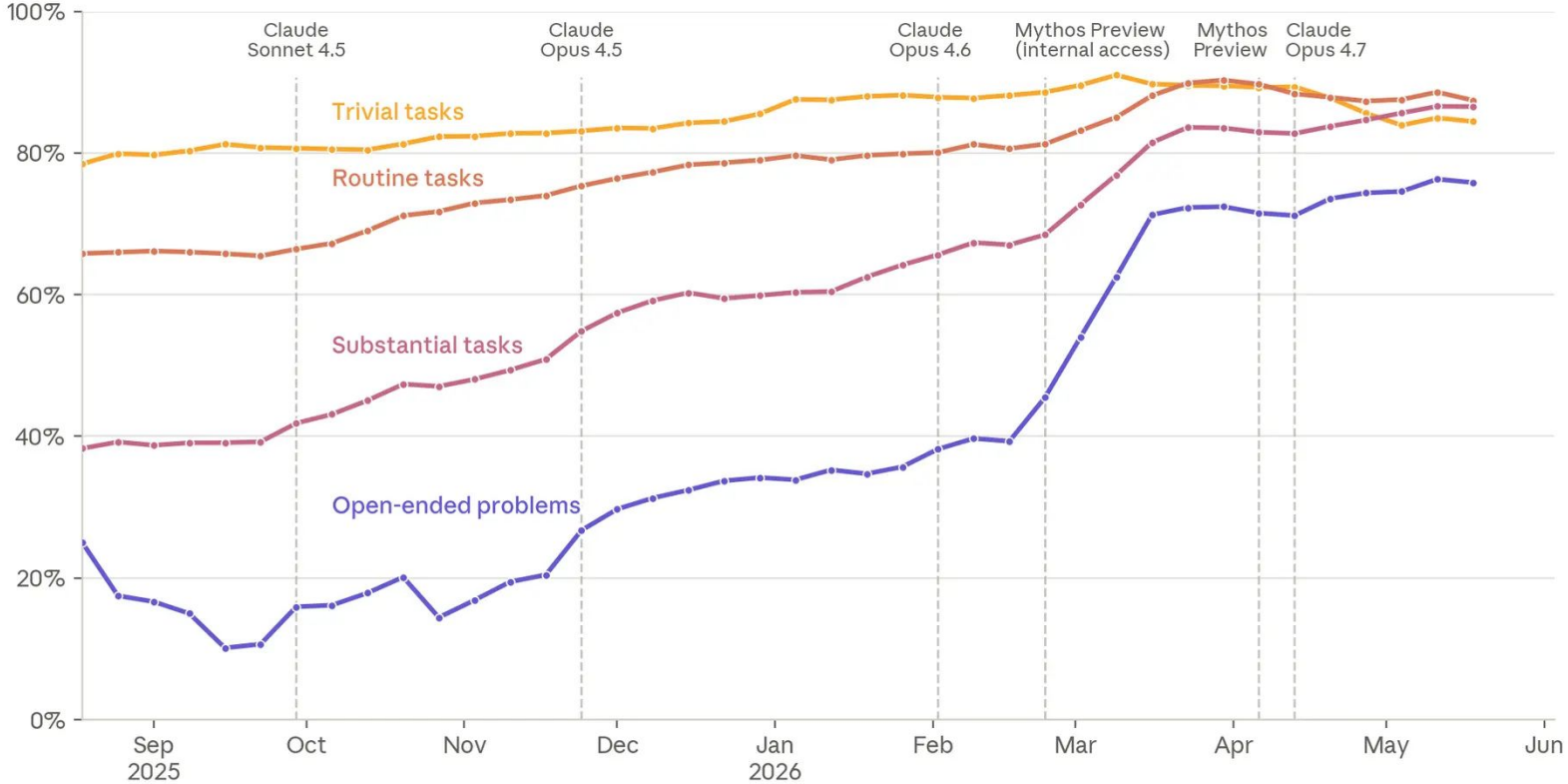
  （緊急時の対応，相互の検証方法など）



Chan, Alan, et al. "What if automating AI R&D triggers an intelligence explosion?." arXiv preprint arXiv:2609.36054 (2026).
Fithian, William, et al. "The Feasibility of a Hardwired Pause of Frontier AI Training." (2026)

ANTHROP\C

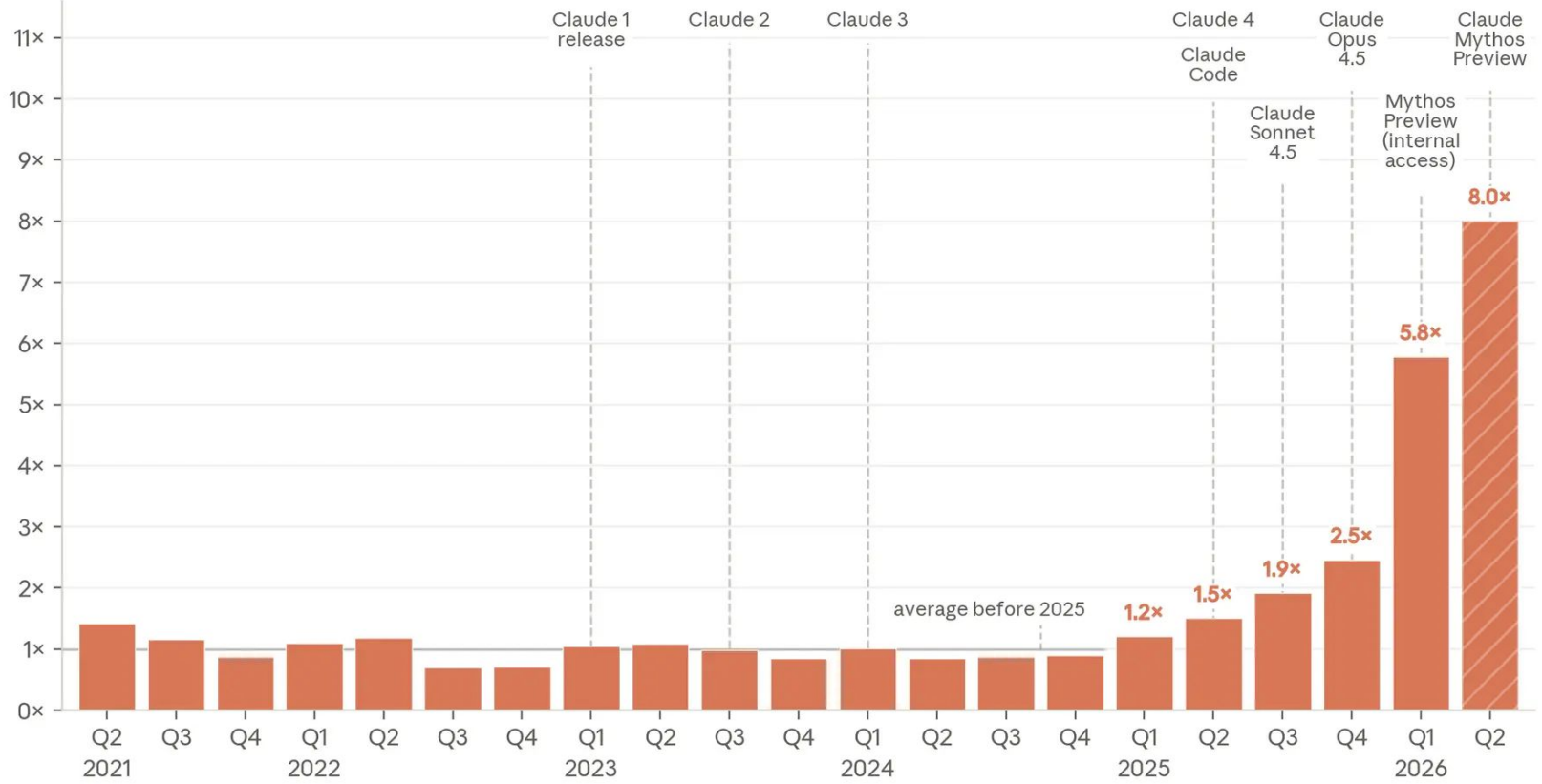
Claude Code session success rate



Internal Claude Code sessions at Anthropic. Weekly, 4-week trailing mean. LLM-judged success.
Task complexity is assigned by an LLM classifier that reads a summary of each session and picks one of four levels from a fixed rubric.

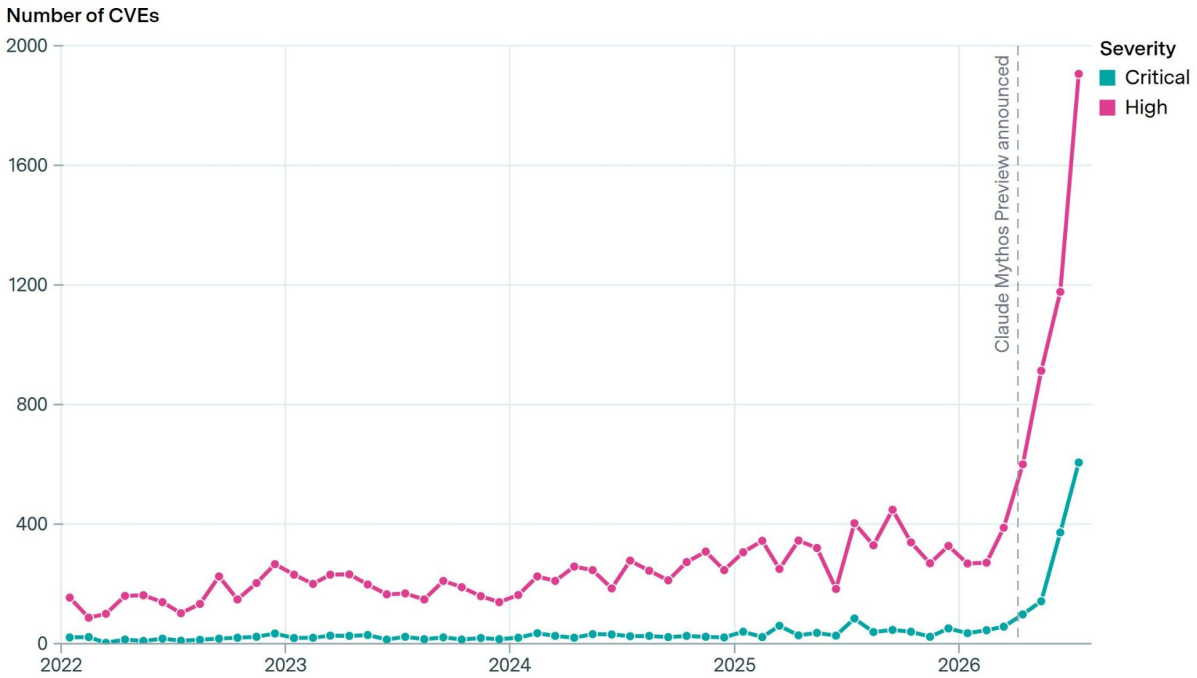
Code contributed per person, by quarter

ANTHROP\C



Each bar is the average, over the days in that quarter, of lines of code merged per active contributor — shown as a multiple of the pre-2025 average. The hatched final bar is a partial quarter: it averages only the days observed so far, not a full quarter. Dashed lines mark public announcement dates. Per-PR line counts are capped at the 99th percentile; "active contributor" means a distinct author in the trailing twelve months.

Serious cyber vulnerability disclosures kept climbing in July

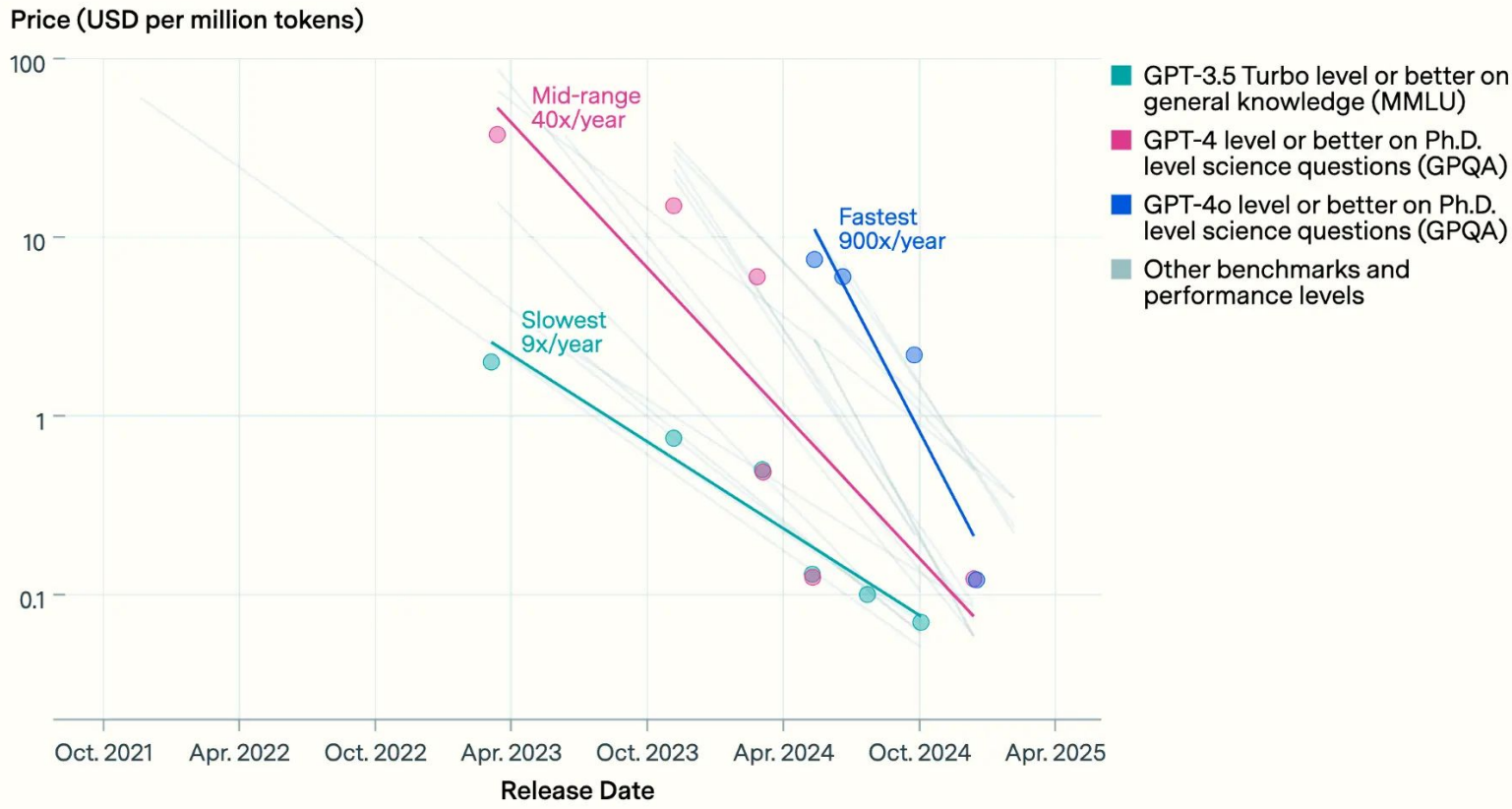


CVEs from AWS, Apache, Apple, Cisco, Google, Linux, Microsoft, Mozilla, NVIDIA, Oracle, Red Hat, Adobe, IBM, Intel, AMD, Qualcomm, Samsung, SAP, VMware, GitHub, and OpenSSL. Reporting procedures, labeling, and cadence vary substantially between organizations.

EPOCH AI | CC-BY

LLM inference prices have fallen 9x to 900x/year, depending on the task

EPOCH AI

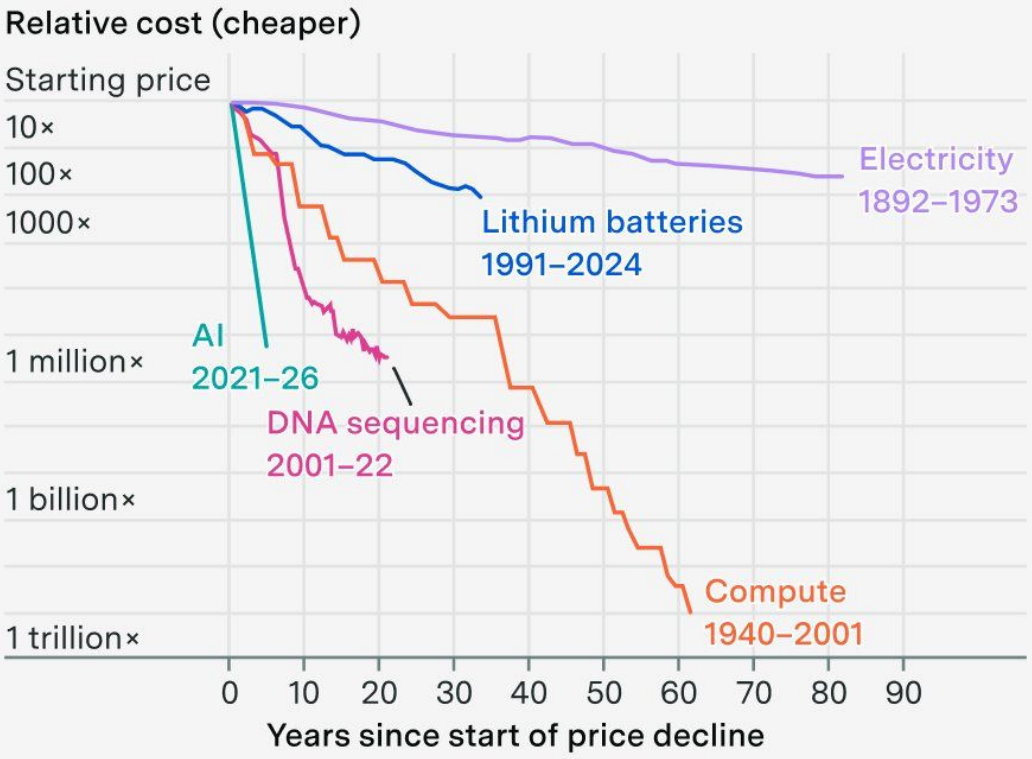


Data source: Epoch AI, Artificial Analysis

CC-BY

epoch.ai

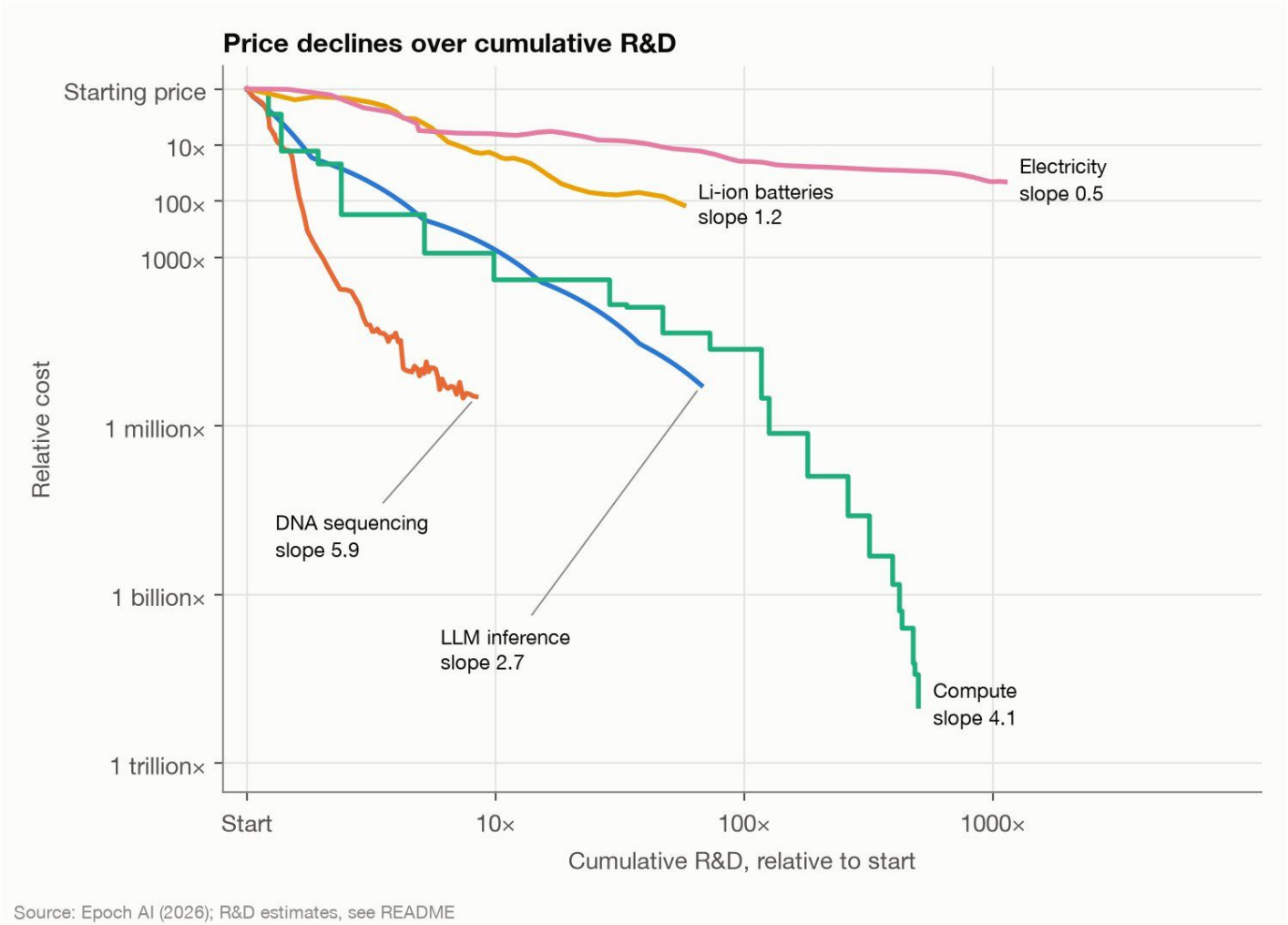
The price of artificial thought may have fallen faster than for any other transformative technology in history



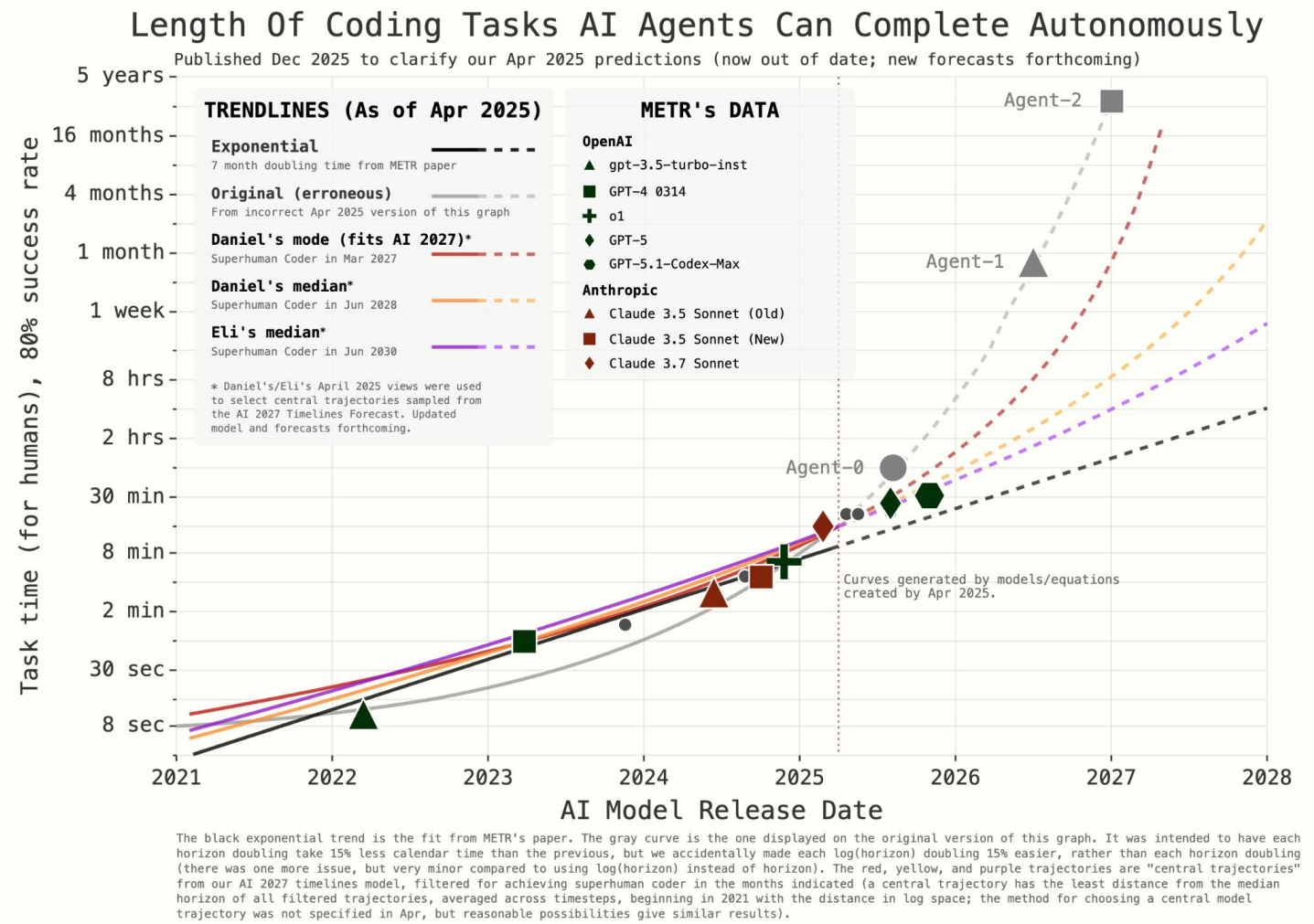
All prices adjusted for inflation. Compute series stops merely for lack of newer data. Sources: Electricity: Hersh (1999), Figure 2.1, and EEI (1995), Table 63. Compute: Nordhaus (2007), Appendix Table 2. Batteries: Fouquet (2026) via Our World in Data. DNA sequencing: National Human Genome Research Institute. AI line is 47%/quarter (12.7x/year), per estimates reported here for 2023–26; the extrapolation of that rate back to 2021 is broadly validated by Appenzeller (2024)'s finding that per-token prices for models clearing a certain capability threshold fell 10x/year in 2021–24.

EPOCH AI | CC-BY

epoch.ai



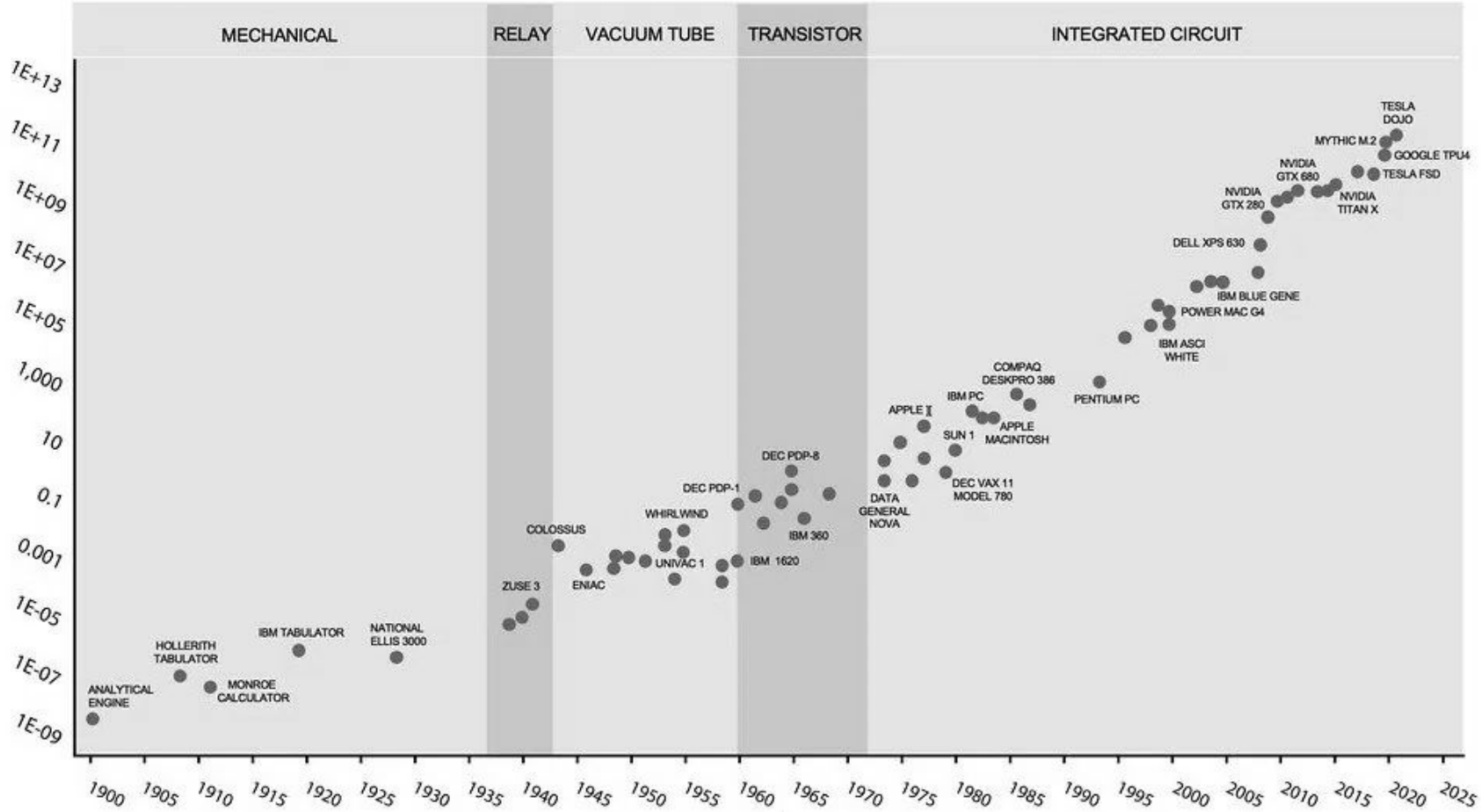
Source: Epoch AI (2026); R&D estimates, see README



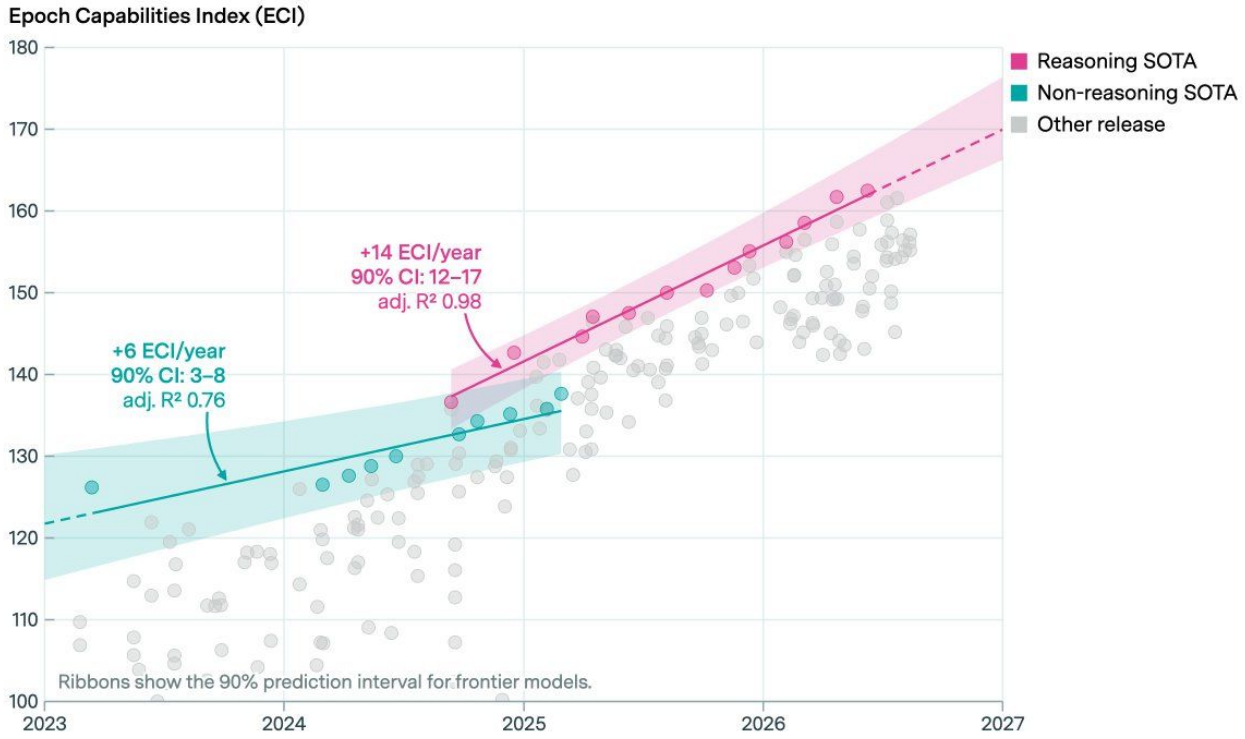
The black exponential trend is the fit from METR's paper. The gray curve is the one displayed on the original version of this graph. It was intended to have each horizon doubling take 15% less calendar time than the previous, but we accidentally made each log(horizon) doubling 15% easier, rather than each horizon doubling (there was one more issue, but very minor compared to using log(horizon) instead of horizon). The red, yellow, and purple trajectories are "central trajectories" from our AI 2027 timelines model, filtered for achieving superhuman coder in the months indicated (a central trajectory has the least distance from the median horizon of all filtered trajectories, averaged across timesteps, beginning in 2021 with the distance in log space; the method for choosing a central model trajectory was not specified in Apr, but reasonable possibilities give similar results).

FUTURE VENTURES

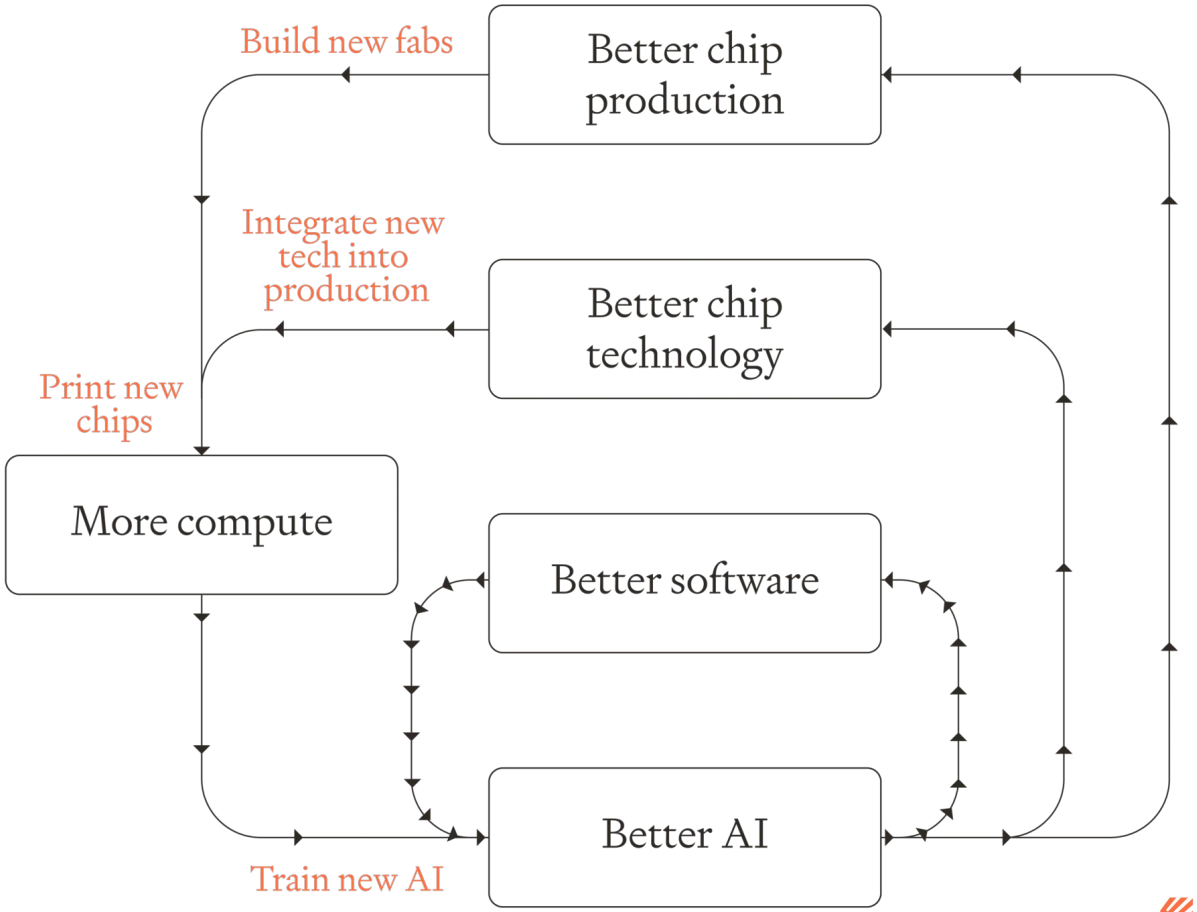
122 YEARS OF MOORE'S LAW



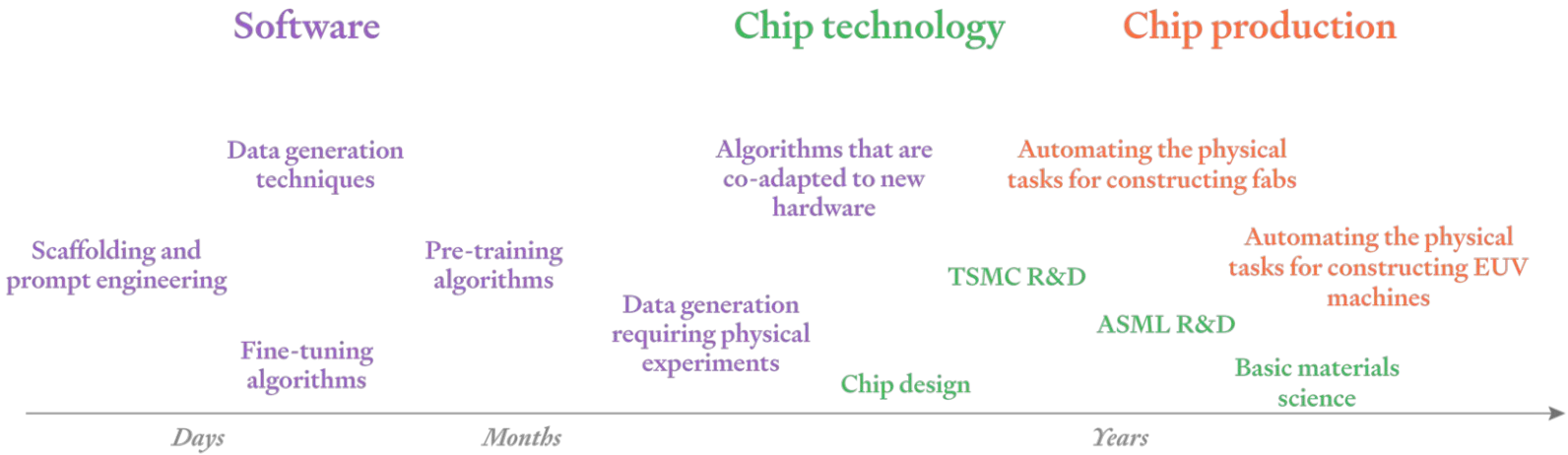
The ECI frontier has advanced by 14 points per year since the introduction of reasoning models



EPOCH AI | CC-BY



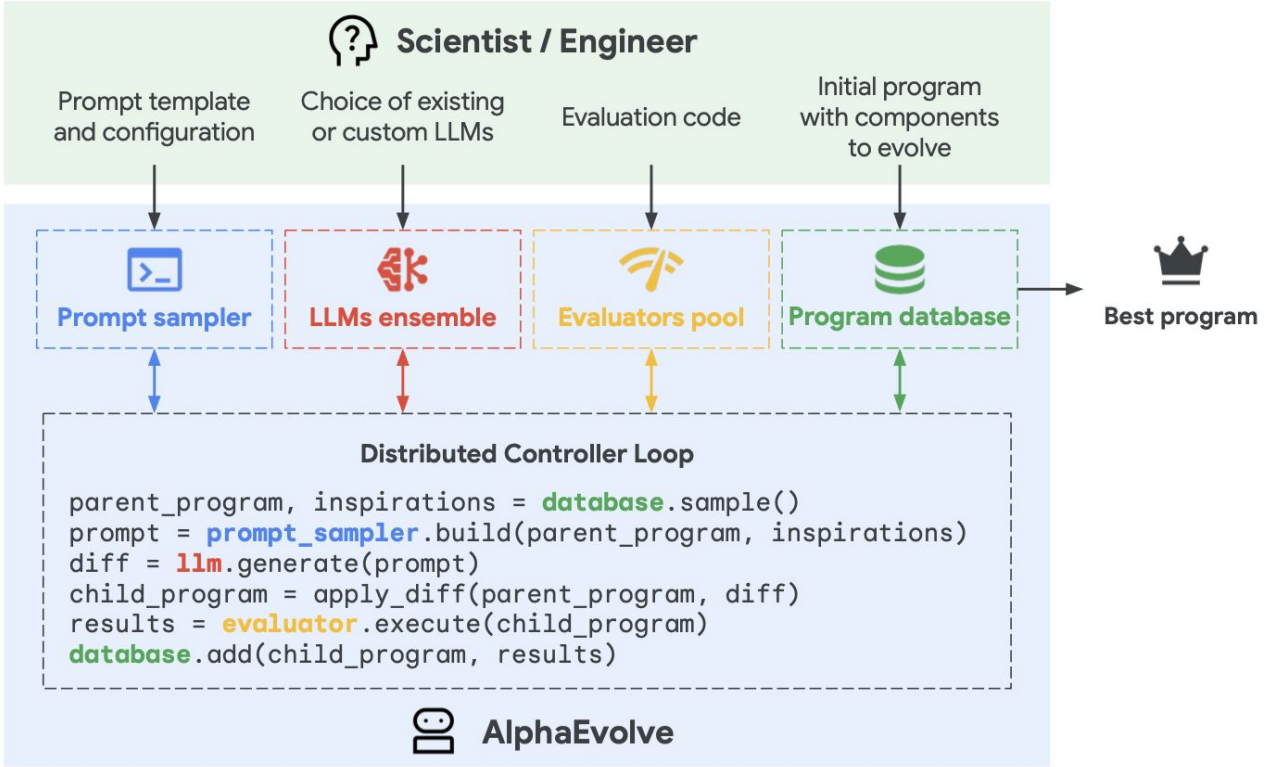
Forethought



Time lag (lag from "work done in this area" to "AI is better")

Forethought

AlphaEvolve



Darwin Gödel Machine

