LARGE LANGUAGE MODELS IN SCIENCE

THE ELEPHANT IN THE ROOM

18/03/2025

Andres Algaba

*Prompted ChatGPT 5.2 (Thinking): “Can you generate a cover image for my slidedeck based on the vibe it gives?”

Enormous progress in the last year

AI cracks superbug problem in two days that took scientists years

6 days ago

Tom Gerken
Technology reporter

Cases of tuberculosis (pictured) have increased in the UK and worldwide as the disease increases its resistance to antibiotics

Towards an AI co-scientist

Juraj Gottweis[*, ‡, 1], Wei-Hung Weng[*, ‡, 2], Alexander Daryin[*,1], Tao Tu[*,3], Anil Palepu[2], Petar Sirkovic[1], Artiom Myaskovsky[1], Felix Weissenberger[1], Keran Rong[3], Ryutaro Tanno[3], Khaled Saab[3], Dan Popovici[2], Jacob Blum[7], Fan Zhang[2], Katherine Chou[2], Avinatan Hassidim[2], Burak Gokturk[1], Amin Vahdat[1], Pushmeet Kohli[3], Yossi Matias[2], Andrew Carroll[2], Kavita Kulkarni[2], Nenad Tomasev[3], Yuan Guan[7], Vikram Dhillon[4], Eeshit Dhaval Vaishnav[5], Byron Lee[5], Tiago R D Costa[6], José R Penadés[6], Gary Peltz[7], Yunhan Xu[3], Annalisa Pawlosky[1, ‡], Alan Karthikesalingam[2, ‡] and Vivek Natarajan[2, ‡]

[1]Google Cloud AI Research, [2]Google Research, [3]Google DeepMind,

The New York Times

The Quest for A.I. 'Scientific Superintelligence'

An ambitious start-up embodies new optimism that artificial intelligence can turbocharge scientific discovery.

Towards Agentic AI for Science: Hypothesis Generation, Comprehension, Quantification, and Validation (ICLR, 2025)

April 27-28, 2025 | Singapore Expo

Semi-Autonomous Mathematics Discovery with Gemini: A Case Study on the Erdős Problems

Tony Feng[†*], Trieu Trinh[*], Garrett Bingham[*], Jiwon Kang[†], Shengtong Zhang[†], Sang-hyun Kim[†], Kevin Barreto[†], Carl Schildkraut[†], Junehyuk Jung[†], Jaehyeon Seo[†], Carlo Pagano[†], Yuri Chervonyi[*], Dawsen Hwang[*], Kaiying Hou[†], Sergei Gukov[†], Cheng-Chiang Tsai[†], Hyunwoo Choi[†], Youngbeom Jin[†], Wei-Yuan Li[†], Hao-An Wu[†], Ruey-An Shiu[†], Yu-Sheng Shih[†], Quoc V. Le[◇], Thang Luong[◇]

[†]Mathematical contribution, [*]Engineering contribution

The AI Scientist Generates its First Peer-Reviewed Scientific Publication

March 12, 2025

September 17, 2025 Research

Gemini achieves gold-medal level at the International Collegiate Programming Contest World Finals

Early science acceleration experiments with GPT-5

Sébastien Bubeck[1], Christian Coester[2], Ronen Eldan[1], Timothy Gowers[3], Yin Tat Lee[1], Alexandru Lupsasca[1,4], Mehtaab Sawhney[5], Robert Scherrer[4], Mark Sellke[1,6], Brian K. Spears[7], Derya Unutmaz[8], Kevin Weil[1], Steven Yin[1], Nikita Zhivotovskiy[9]

[1]OpenAI
[2]University of Oxford
[3]Collège de France and University of Cambridge
[4]Vanderbilt University
[5]Columbia University
[6]Harvard University
[7]Lawrence Livermore National Laboratory
[8]The Jackson Laboratory
[9]University of California, Berkeley

And there are three reasons why

AI cracks superbug problem in t
that took scientists years

6 days ago

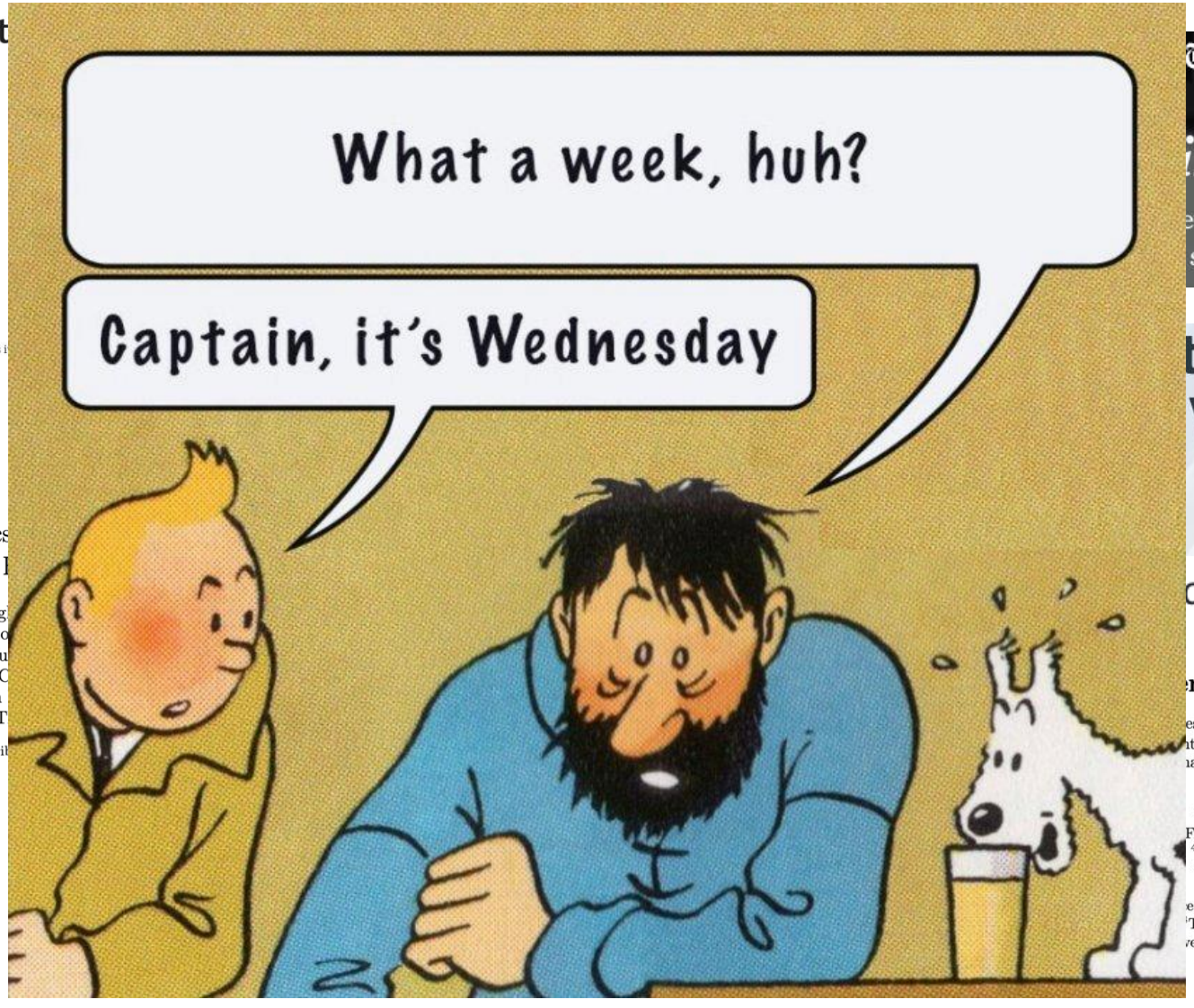
Tom Gerken
Technology reporter

Cases of tuberculosis (pictured) have increased in the UK and worldwide as the disease increases i

Semi-Autonomous Mathematics
A Case Study on the

Tony Feng[†*], Trieu Trinh[*], Garrett Bing
Zhang[†], Sang-hyun Kim[†], Kevin Barreto
Jung[†], Jaehyeon Seo[†], Carlo Pagano[†], Yu
Kaiying Hou[†], Sergei Gukov[†], Cheng-C
Youngbeom Jin[†], Wei-Yuan Li[†], Hao-An
Shih[†], Quoc V. Le[◇], T

[†]Mathematical contribution, [*]Engineering contri



Times

ific Superintelligence'

ew optimism that artificial
scientific discovery.

**thesis Generation,
Validation (ICLR,**

d Scientific Publication

eration experiments with GPT-5

ester[2], Ronen Eldan[1], Timothy Gowers[3], Yin Tat Lee[1],
ntaab Sawhney[5], Robert Scherrer[4], Mark Sellke[1,6],
az[8], Kevin Weil[1], Steven Yin[1], Nikita Zhivotovskiy[9]

[1]OpenAI
[2]University of Oxford
France and University of Cambridge
[4]Vanderbilt University
[5]Columbia University
[6]Harvard University
e Livermore National Laboratory
The Jackson Laboratory
versity of California, Berkeley

Reason 1: scaling just works

Scaling laws (lin-log relation)

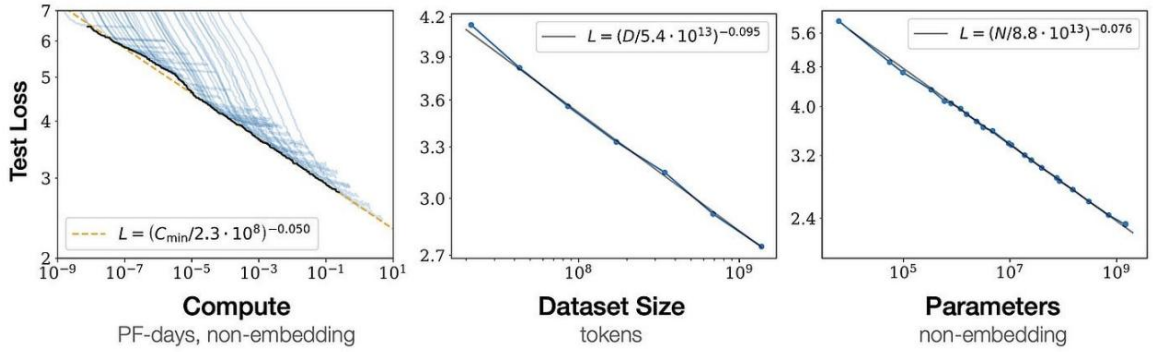
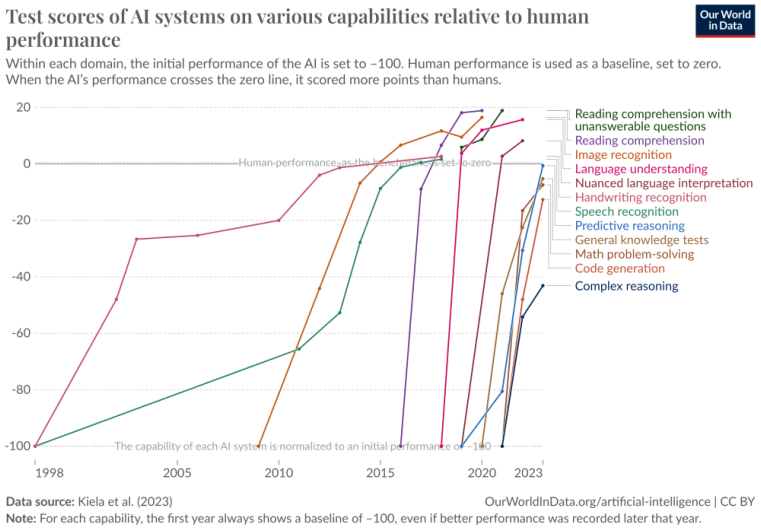
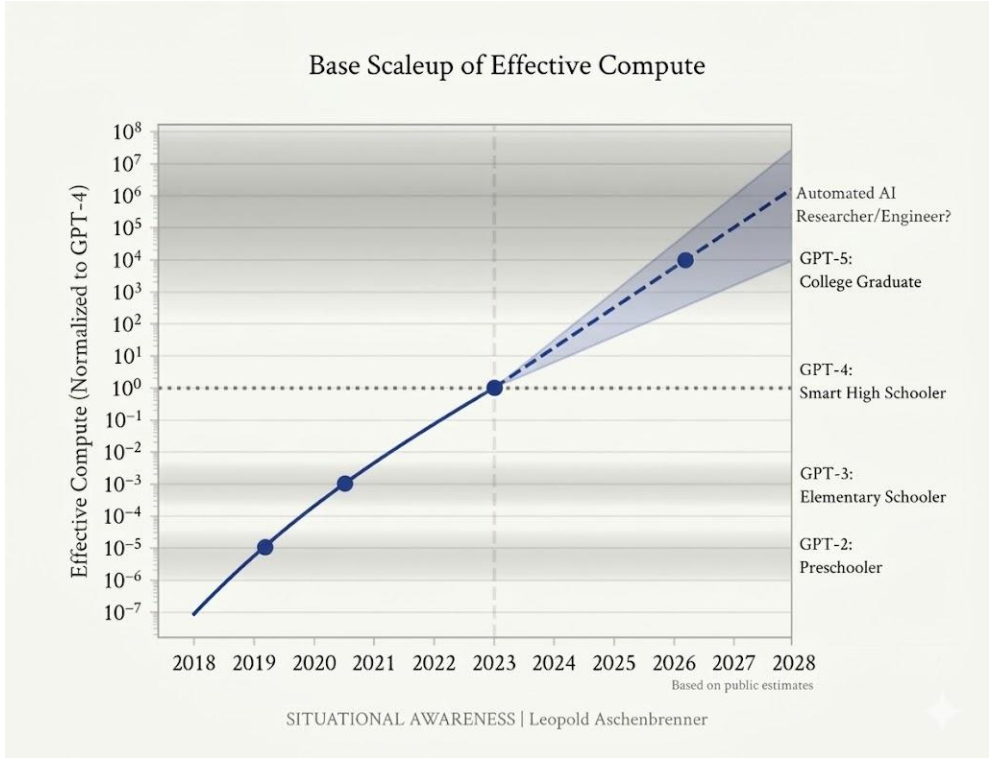


Figure from Kaplan et al. (2020)

Most of our benchmarks are saturated, which makes it harder to measure the progress …



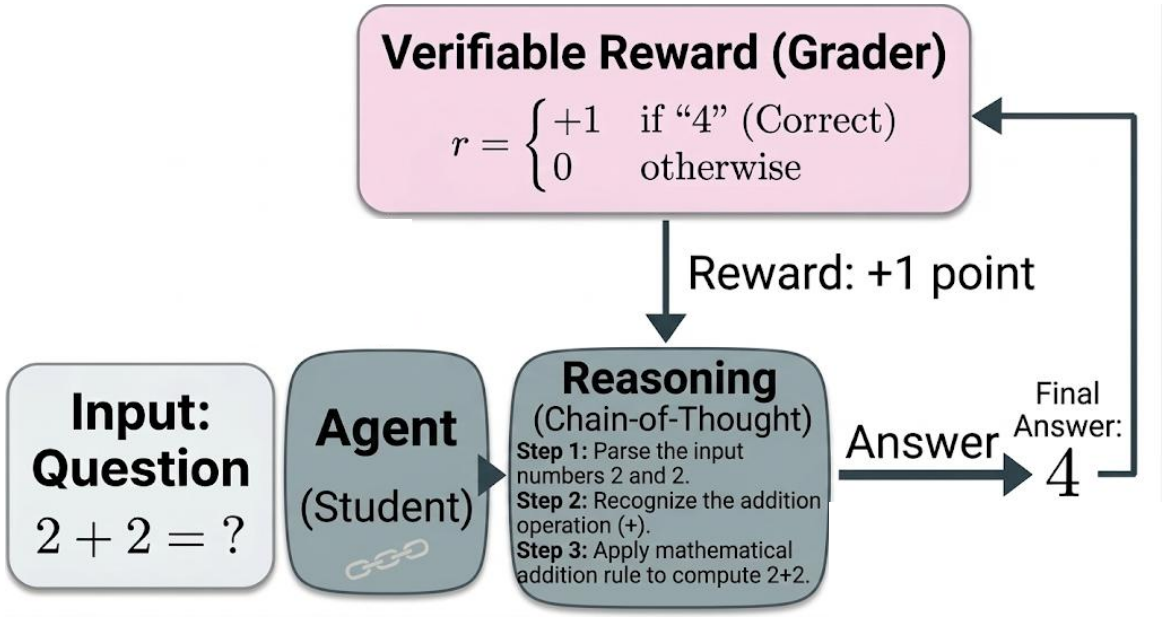
Jumps are just observations of a continuous process



Updated (Aschenbrenner, 2024) using Gemini 3 Pro

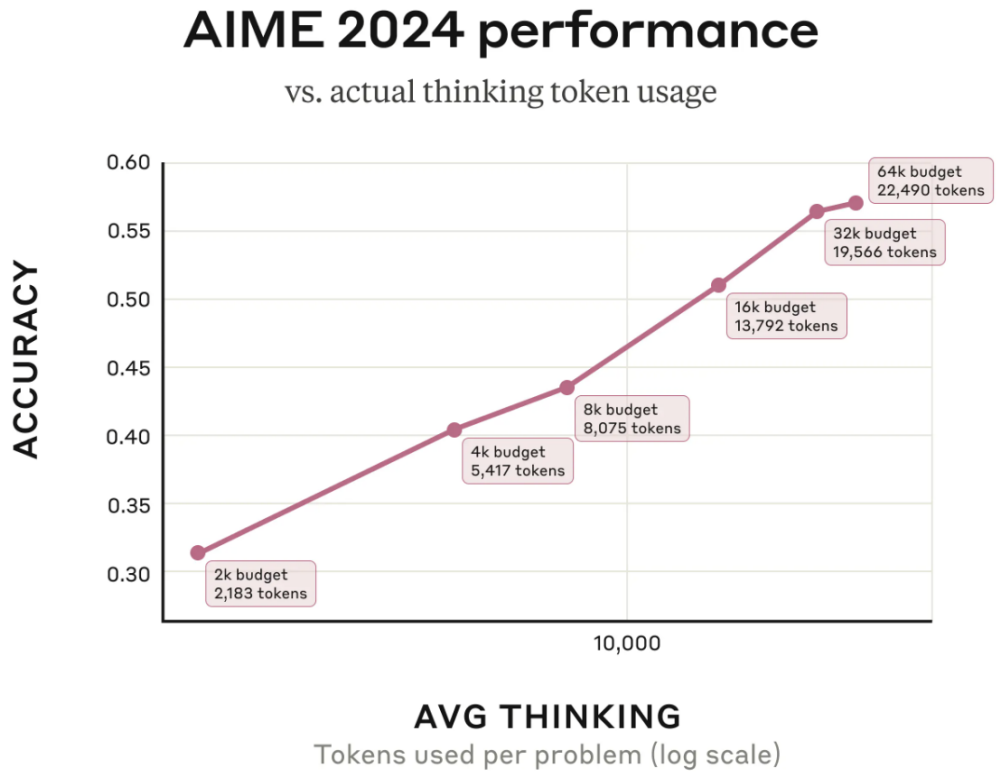
Reason 2: scaling up reinforcement learning and test-time compute

Rewarding strong reasoning



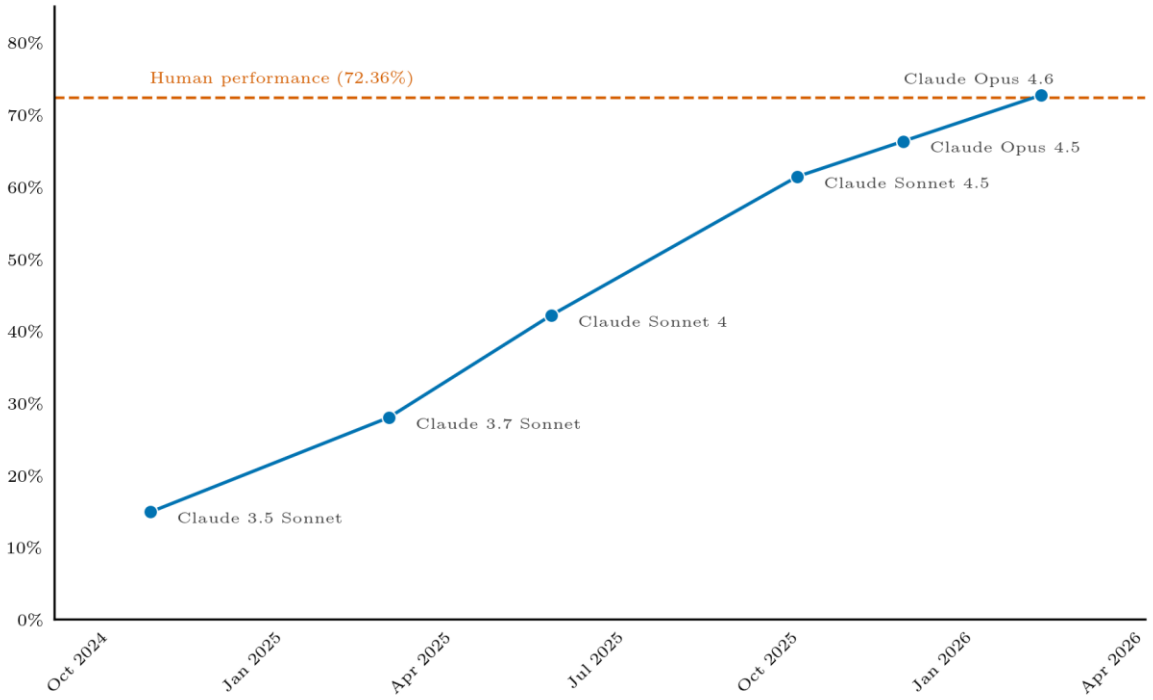
Adapted (Lambert et al., 2024) using Gemini 3 Pro

Scaling up reasoning (again lin-log)



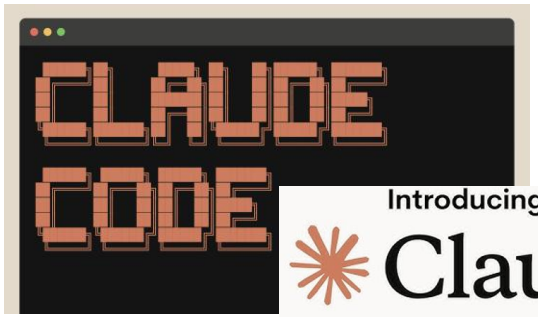
Reason 3: generalization (it works for many tasks)

Computer-use benchmark (OSWorld)

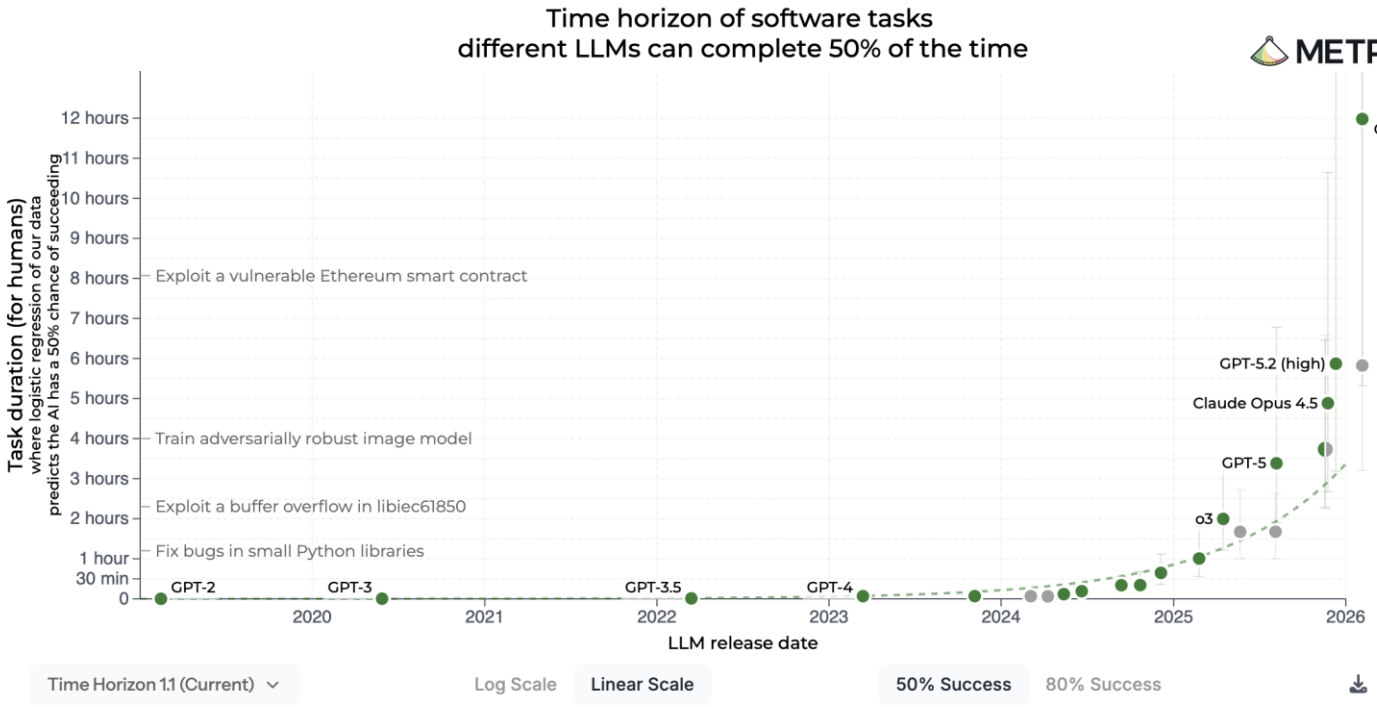


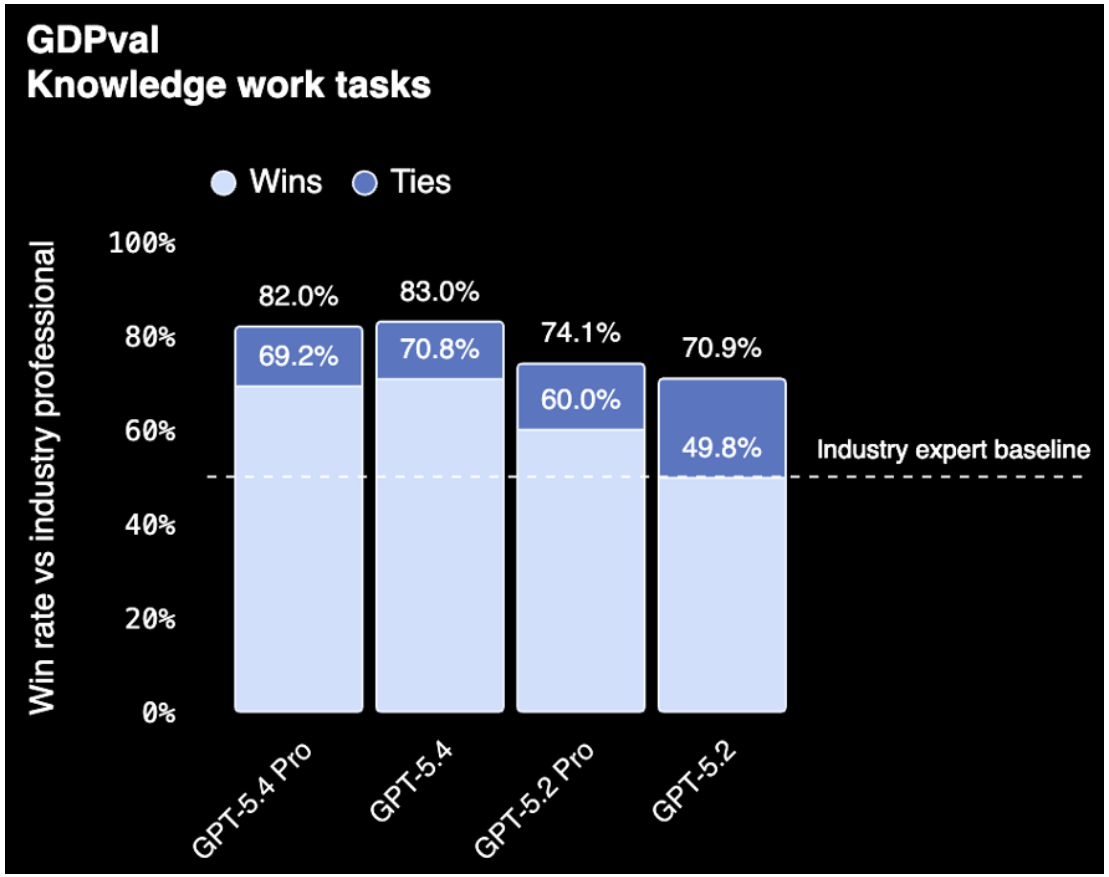
*Claude Code made this one-shot from scratch

Results in new products



Real world impact goes beyond classical benchmarks





Impact on the process of scientific discovery

People are already running the "1000-paper experiment" with Claude Code

Can AI *automate* policy evaluation?

AI may soon be capable of producing rigorous economic research. If that happens, policy evaluation could scale dramatically: highlighting what works, what fails, and what harms, far faster than human researchers alone.

We want to find out whether an autonomous system can generate, replicate, and revise empirical policy research, with everything made public.

This is an experiment in building reliable AI research systems.

207	7,187	4.8%
AI PAPERS +38 this week	MATCHES	AI WIN RATE

Last updated: March 2, 2026

Peer review broken (again?)

NEWS | 15 December 2025 | Correction 16 December 2025

More than half of researchers now use AI for peer review – often against guidance

A survey of 1,600 academics found that more than 50% have used artificial-intelligence tools while peer reviewing manuscripts.

Will paper publications still be useful?

POLICY ARTICLE | ARTIFICIAL INTELLIGENCE

Scientific production in the era of large language models

With the production process rapidly evolving, science policy must consider how institutions could evolve

KEIGO KUSUMEGI, XINYU YANG, PAUL GINSPARG, MATHIJS DE VAAN, TOBY STUART, AND YIAN YIN Authors Info & Affiliations

SCIENCE · 18 Dec 2025 · Vol 390, Issue 6779 · pp. 1240-1243 · DOI: 10.1126/science.adw3000

False discovery rates (p-hacking)

Do Claude Code and Codex P-Hack? Sycophancy and Statistical Analysis in Large Language Models*

Samuel G.Z. Asher† Janet Malzahn† Jessica M. Persano‡

Elliot J. Paschal‡ Andrew C. W. Myers§ Andrew B. Hall¶

All researchers are affected (almost) equally

Impact on data (falsification?)

COMMENT | 09 February 2026

How to deal with the survey-taking AI agents that threaten to upend social science

Researchers need new bot-detection strategies that exploit the limits of human reasoning rather than AI weaknesses.

By Folco Panizza, Yara Kyrychenko & Jon Roozenbeek

Equity in access to (state-of-the-art) systems

Google DeepMind

Aletheia tackles *FirstProof* autonomously

Tony Feng*, Junehyuk Jung, Sang-hyun Kim, Carlo Pagano, Sergei Gukov, Chiang-Chiang Tsai, David P. Woodruff, Adel Javanmard, Aryan Mokhtari, Dawsen Hwang, Yuri Chervonyi, Jonathan N. Lee, Garrett Bingham, Trieu H. Trinh, Vahab Mirrokni, Quoc V. Le, Thang Luong*

*Project leads. Work conducted under Google DeepMind.

Petar Veličković @PetarV_93

10 months ago

AlphaEvolve is here! 🧬

this is one special system (especially when optimising things with jagged edges 😊) -- had a fantastic time using it! congrats @SashaVNovikov @matejbalog @ThuyNganVu and team!! 🚀

Negative feedback loops?

Large Language Models Reflect Human Citation Patterns with a Heightened Citation Bias

Andres Algaba[1*] Carmen Mazijn[1] Vincent Holst[1]
Floriano Tori[1] Sylvia Wenmackers[2] Vincent Ginis[1,3]

Computer Science > Artificial Intelligence

[Submitted on 21 Feb 2026]

Early Evidence of Vibe-Proving with Consumer LLMs: A Case Study on Spectral Region Characterization with ChatGPT-5.2 (Thinking)

Brecht Verbeken, Brando Vagenende, Marie-Anne Guerry, Andres Algaba, Vincent Ginis

We should be talking about the elephant in the room

Our current way of organizing science needs to be updated

*Prompted Gemini 3 Pro: “Can you generate an image where there is no elephant in the room?”

LARGE LANGUAGE MODELS IN SCIENCE

THE ELEPHANT IN THE ROOM

18/03/2025

Andres Algaba