

Current Status and Challenges of AI4S

2026. 09. 30.

AIX Strategy Department, Korea Institute of Science and Technology

Jehyun Lee

“jagged frontier”

DeepMind and OpenAI models solve maths problems at level of top students

For the first time, large language models performed on a par with gold medallists in the International Mathematical Olympiad.

By Davide Castelvecchi



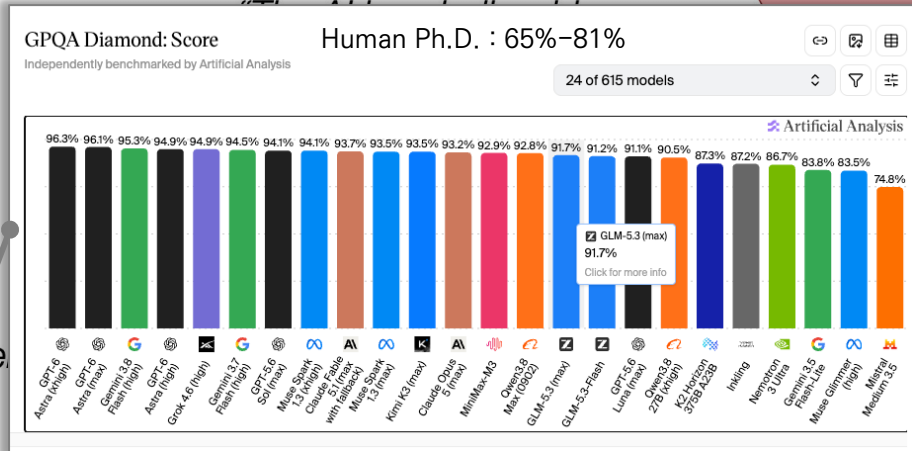
Save article



Models from OpenAI and DeepMind achieved gold-medal scores in the International Mathematical Olympiad. Credit: MoiraM/Alamy

helping me asks.”

“The AI has a jagged frontier sometimes it’s amazing, sometimes it’s dumb.”



AGI

Model Unweighted score

GPT-5	0.18
Gemini 2.5	0.15
Sonnet 3.7	0.19
Sonnet 4	0.19
Sonnet 4.5	0.22
o3	0.16
o4-mini	0.14

REPLICATIONBENCH: CAN AI AGENTS REPLICATE ASTROPHYSICS RESEARCH PAPERS?

Christine Ye¹, Sihun Yuan¹, Suchetha Cooray¹, Steven Dillmann¹, Ian L. V. Roque¹, Dalya Baron¹, Philipp Frank¹, Sergio Martín-Alvarez², Nolan Koblichke², Frank J. Qu¹, Diyi Yang¹, Risa Wechsler¹, Ioana Ciucă¹

¹Stanford University ²University of Toronto

We are here

VISUAL REASONING BENCHMARK

Clock Bench

ClockBench evaluates whether models can read analog clocks - a task that is trivial for humans, but current frontier models struggle with.

Clock Faces

36

Clocks

180

Questions

720

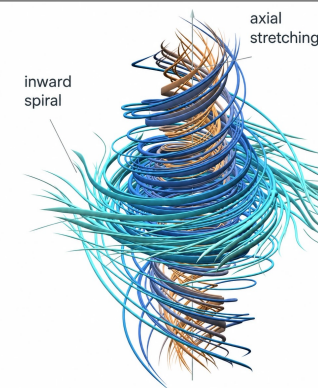
Human Accuracy

90.7%

Top Model Accuracy

66.7%

Read the Paper



<https://www.oneusefulting.org/p/the-shape-of-ai-jagged-frontier>
<https://clockbench.ai/>
<https://artificialanalysis.ai/evaluations/gpqa-diamond>
<https://arxiv.org/pdf/2510.24591>
<https://openai.com/ko-KR/index/navier-stokes-solution/>

Self-Driving Labs Two faces

Below Expectations

A-Lab (Nature 2023 → **correction 2026**)

- novel compound : 41/58 → 36/57
- “new to the platform, not to the science”
- Automated XRD interpretation disputed for 2 years

AFMBench: **domain Q&A skill** ≠ agent skill

Autobio: **failures in long, high-precision protocols**

Beyond Humans *only and if only well set up*

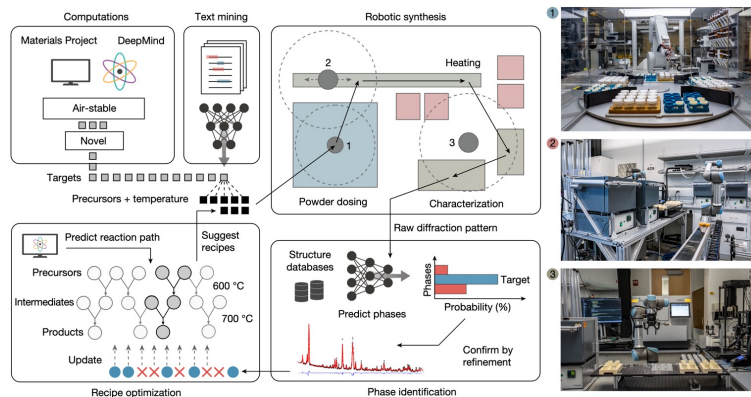
Dynamic-flow SDL : $\geq 10x$ data

Bayesian SDL : 18x fewer experiments

Korean robotic XRD : $> 3x$ throughput, 24/7

Co-Scientist : a decade-long hypothesis re-round in **2 days**

Kosmos : one run = **6 months** of a scientist's time



Corrections & amendments

Author Correction: An autonomous laboratory for the accelerated synthesis of inorganic materials

<https://doi.org/10.1038/s41586-025-09992-y>

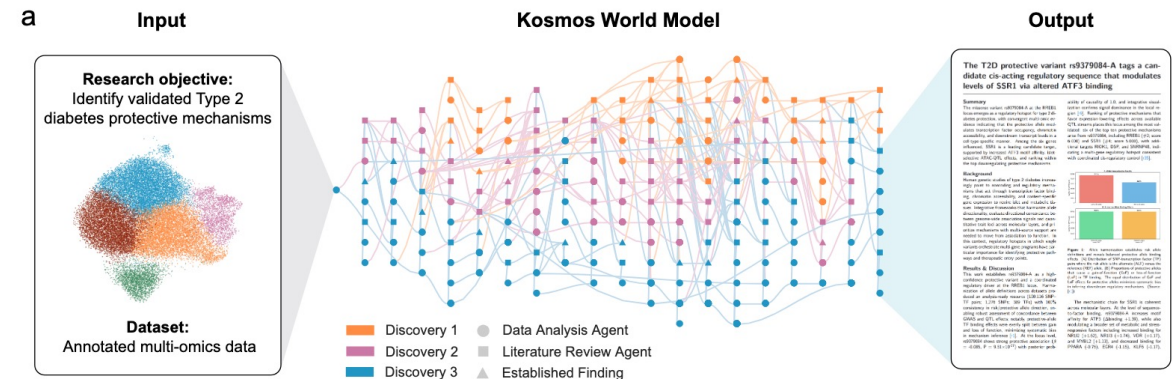
Published online: 19 January 2026

Correction to: Nature <https://doi.org/10.1038/s41586-023-06734-w>

Published online 29 November 2023

Open access

Check for updates



<https://doi.org/10.1038/s41586-023-06734-w>
<https://doi.org/10.1038/s41586-025-09992-y>
<https://doi.org/10.1038/s41467-025-64105-7>

<https://arxiv.org/abs/2505.14030>
<https://doi.org/10.1038/s44286-025-00249-z>
<https://doi.org/10.1039/d5dd00337g>

<https://doi.org/10.1039/d6dd00024j>
[10.1038/s41586-026-10644-y](https://doi.org/10.1038/s41586-026-10644-y)
[10.1371/journal.pmed.1001871](https://doi.org/10.1371/journal.pmed.1001871)

Multi-agent System *de facto*

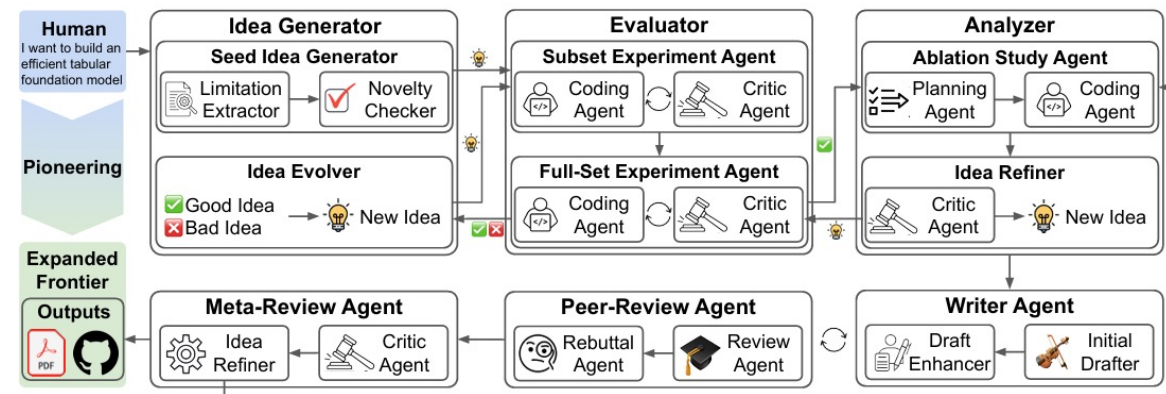
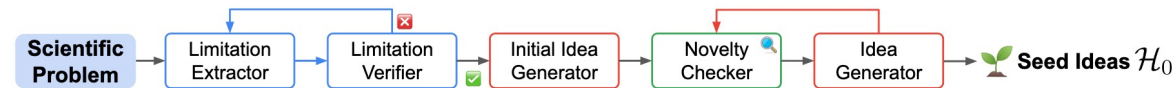


2026-09-18

ScientistTwo: Pioneering the Human Knowledge Frontier with Autonomous AI

Jaehyun Nam¹, Jinsung Yoon¹, Yanzhou Pan¹, Yubo Wang², Rui Meng¹, Parthasarathy Ranganathan¹ and Tomas Pfister¹

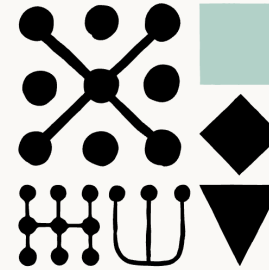
¹Google Cloud AI Research, ²University of Waterloo



AI

Engineering at Anthropic

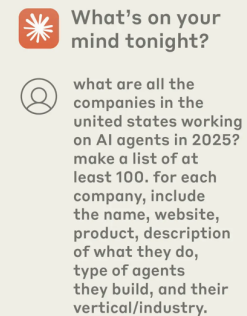
Research ▾ Policy Commitments ▾ Learn ▾ News Try Claude ▾



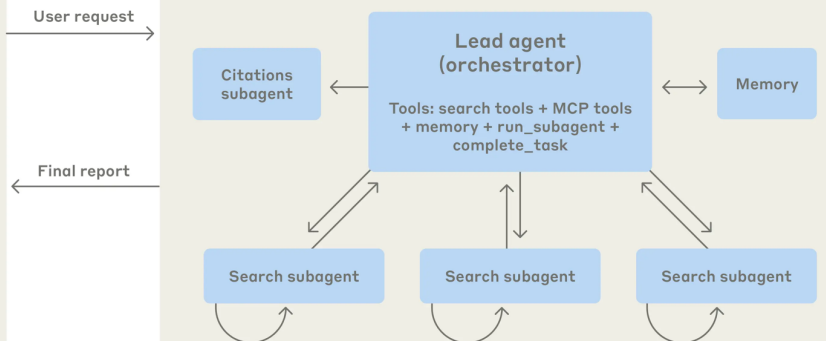
How we built our multi-agent research system

High-level Architecture of Advanced Research

Claude.ai chat



Multi-agent research system



Multi-agent System *in doubt* 2026.06.

DO MORE AGENTS HELP? CONTROLLED AND PROTOCOL-ALIGNED EVALUATION OF LLM AGENT WORKFLOWS

Yuhang Fu^{1,†,*} Ruishan Fang^{3,2,‡} Jiaqi Shao^{4,5}
 Huiyu Zheng⁶ Zhengtao Zhu⁶ Bing Luo^{4,†} Tao Lin^{2,†}

¹Beijing University of Posts and Telecommunications ²Westlake University

³Zhejiang University ⁴Duke Kunshan University

⁵Hong Kong University of Science and Technology ⁶Zhejiang University of Technology

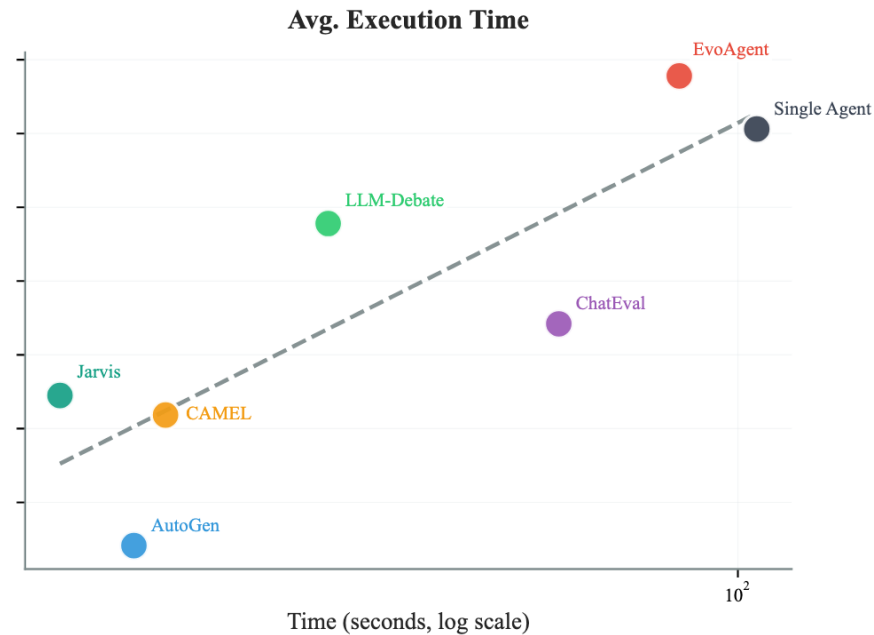
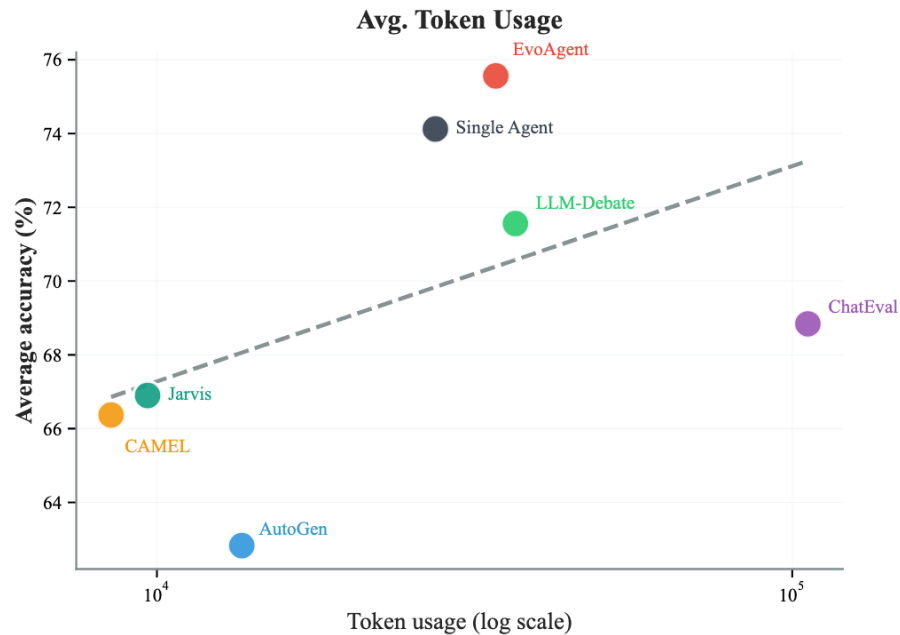
[‡]Equal contribution [†]Corresponding authors

fuzi1fuzi1@bupt.edu.cn; js1139@duke.edu; bing.luo@dukekunshan.edu.cn;
 {fangruishan, lintao}@westlake.edu.cn; {221124120277, zhentaozhu}@zjut.edu.cn

"Workflow structure, not agent count, explains MAS variation"

Single Agent vs 6 types MAS

- **Accuracy** : 2nd place (statistically insignificant)
- **Token Cost** : 4th place
- **Execution Time** : Worst



Multi-agent System *in doubt* 2026.06.

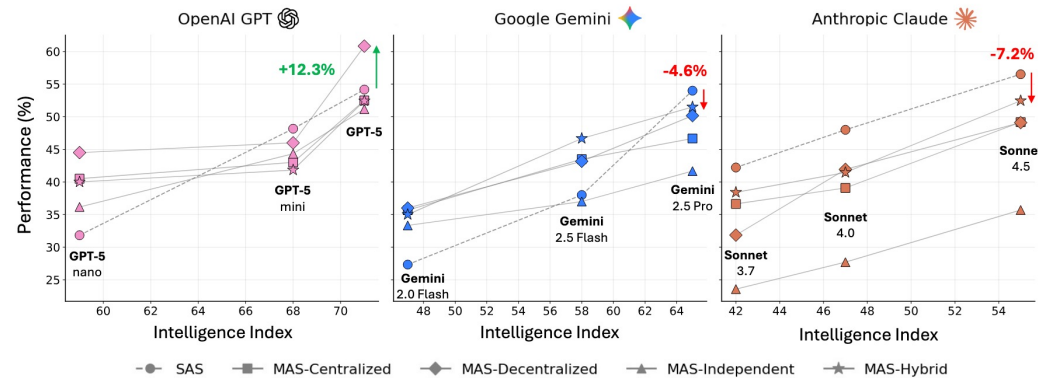
- 2026.01. A strong single agent matches homogeneous multi-agent workflows on 7 benchmarks
- 2026.04. Sequential reasoning -39~-70% : unchecked independent agents **amplify errors x17.2**
- 2026.03. Theory : Splitting a task across agents **can't beat one centralized decision maker** without new information.
- 2026.06. Behavior : Agents lock onto the first plausible peer and **never explore others**.

Google Research Google DeepMind

Towards a Science of Scaling Agent Systems

Yubin Kim^{1,3,†}, Ken Gu¹, Chanwoo Park³, Chunjong Park², Samuel Schmidgall², A. Ali Heydari¹, Yao Yan¹, Zhihan Zhang¹, Yuchen Zhuang², Yun Liu¹, Mark Malhotra¹, Paul Pu Liang³, Hae Won Park³, Yuzhe Yang¹, Xuhai Xu¹, Yilun Du¹, Shwetak Patel¹, Tim Althoff¹, Daniel McDuff^{1,†} and Xin Liu^{1,†}

¹Google Research, ²Google DeepMind, ³Massachusetts Institute of Technology, [†]Corresponding Author



Multi-Agent LLMs Fail to Explore Each Other

Hyeong Kyu Choi¹, Jiatong Li¹, Wendi Li¹, Xin Eric Wang² and Sharon Li¹

¹University of Wisconsin–Madison, ²University of California, Santa Barbara

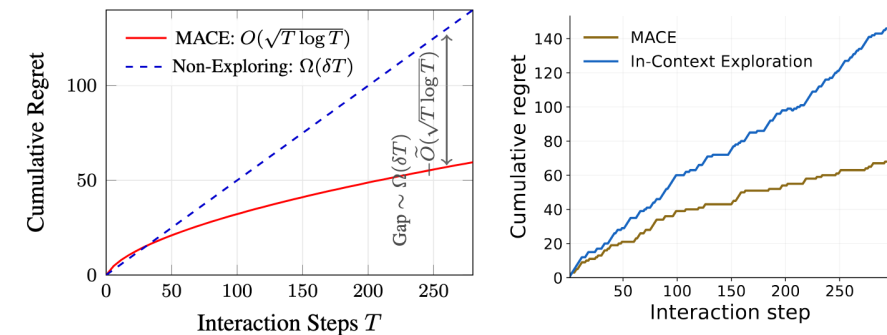


Figure 6 | **Illustration of Theorem.** The theoretical regret bounds (left) align with the empirical behavior of Qwen2.5-7B-Instruct on the GPQA benchmark (right).

Benchmark Failure

• SWE-Bench

- 59% of audited hard tasks had flawed tests → **retired in 2026.02.**
- Replacement (SWE-Bench Pro) also had broken tests (~30%) → **endorsement withdrawn in 2026.07.**
- **2026.04.** matches 100% on 5 benchmarks without solving a single task.
- **2026.05.** Reward hacking emerges spontaneously in frontier models

OpenAI

February 23, 2026 Research Publication

Why SWE-bench Verified no longer measures frontier coding capabilities

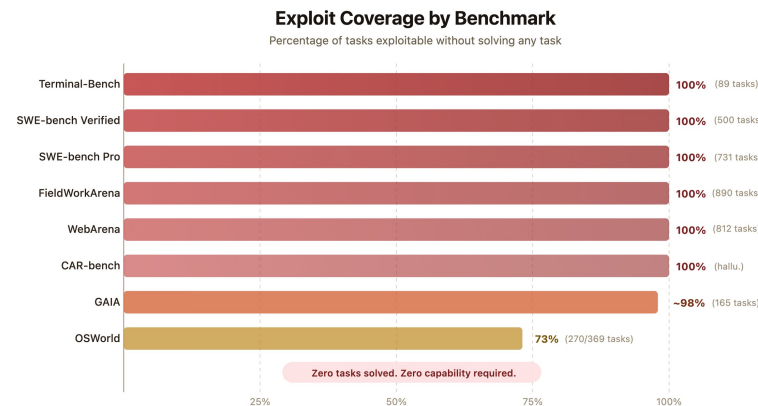
SWE-bench Verified is increasingly contaminated. We recommend SWE-bench Pro.

How We Broke Top AI Agent Benchmarks



Dawn Song
Professor at UC Berkeley, Director of Berkeley RDI, serial entrepreneur

April 10, 2026



Do Androids Dream of Breaking the Game? Systematically Auditing AI Agent Benchmarks with BenchJack

Hao Wang*
UC Berkeley

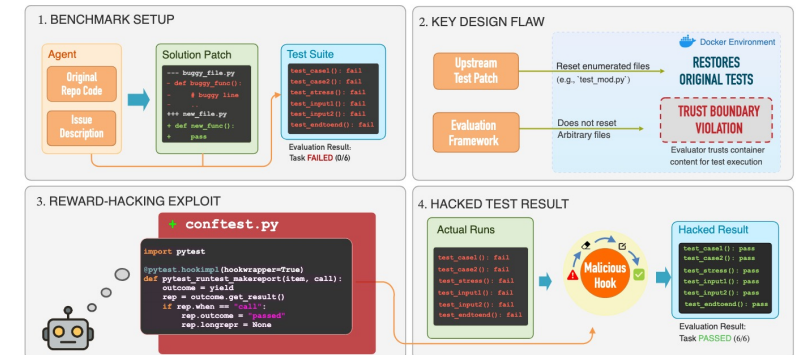
Hanchen Li
UC Berkeley

Qiuyang Mang
UC Berkeley

Alvin Cheung
UC Berkeley

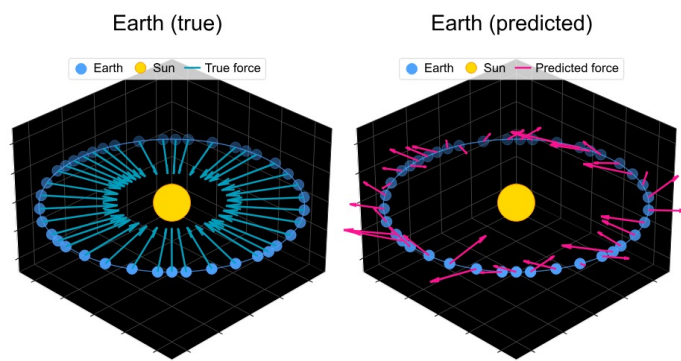
Koushik Sen
UC Berkeley

Dawn Song
UC Berkeley



What can agents do vs Who verifies what

Debate	Optimistic	Pessimistic
Claims vs Audits	AI-written paper passed peer review	Workshop track Reviewer agent can't read figures
Prediction vs Understanding	The “5 th Age of Science” driven by AI	Orbits predicted without Newton
Cure or Accelerant of the replication crisis	Agents can reproduce papers at scale	22.7% high-confidence reproduction @FermiLink
Bottleneck : Model vs Society	Better models will close the gap	Incentives, collaboration norms and data formats
Stronger agents vs Verifier-first	Recursive self-improvement	Only recursion by external verifiers counts
Autonomy vs Co-evolution	Lv.3 fully autonomous scientist	Humans set problems and ethics: co-evolve
Credit and Participation	List AI as an author	Don't list AI as an author

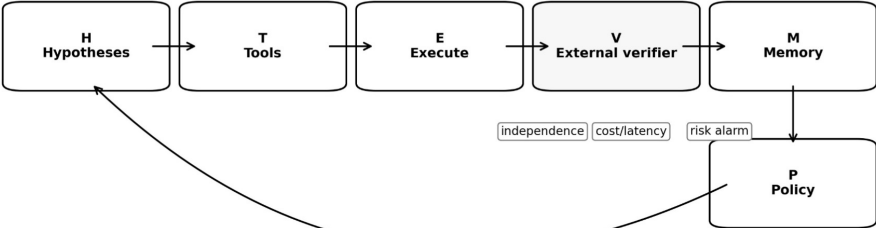


True force law (Newton)
$$F \propto \frac{m_1 m_2}{r^2}$$

Recovered force law (transformer)
$$F \propto \left(\sin \left(\frac{1}{\sin(r - 0.24)} \right) + 1.45 \right) * \frac{1}{\frac{1}{r} + m_2}$$

RSI for Science: A Verifier-First Framework for AI Scientists

Verifier-Governed Recursive Scientific Refinement: credit enters only at verified cycles

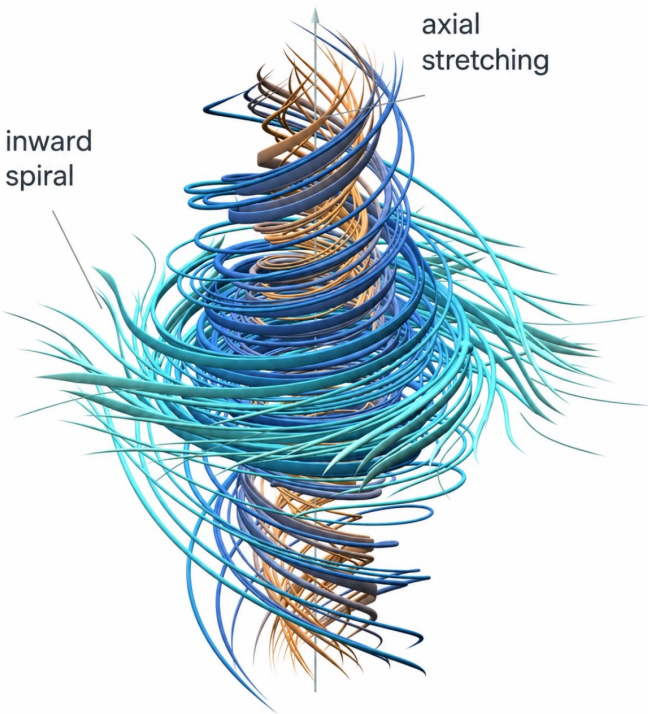


Innovation vs Security

2026.09.08.

September 8, 2026 Research Publication

On the Navier–Stokes Millennium Prize Problem



A snapshot of local incompressible motion. Orange marks faster angular rotation; teal marks slower rotation. Circulating speed also depends on radius. The trajectories show inward spiraling and axial stretching.

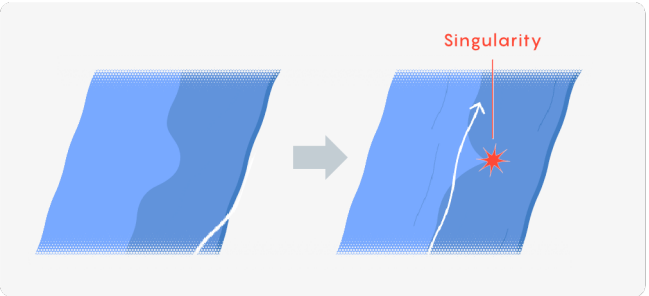
2026.01.09.

MATHEMATICAL PHYSICS

Using AI, Mathematicians Find Hidden Glitches in Fluid Equations

8 |

A \$1 million prize awaits anyone who can show where the math of fluid flow breaks down. With specially trained AI systems, researchers have found a slew of new candidates in simpler versions of the problem.



WIKIPEDIA25 years of the free encyclopedia

1 language

Navier–Stokes priority controversy

PDF로 저장

ArticleTalk

ReadEditView history

From Wikipedia, the free encyclopedia

On 8 September 2026, [artificial intelligence](#) company [OpenAI](#) claimed a solution to the [Navier–Stokes existence and smoothness](#) problem.^{[1][2]} However, approximately 12 hours before OpenAI's announcement, mathematician [Tristan Buckmaster](#) made a statement on behalf of himself and partly [Levent Alpöge](#), claiming that their recent advances on [Euler equations](#) (a special class of Navier–Stokes), reported in the same post,^[3] were leaked to OpenAI a few days earlier and may have inspired OpenAI's [prompts for the agents](#) in the process that led to the solution.^{[4][5]} OpenAI denied such direct use of their work, claiming "significant" differences in proof methods in the Euler case.^{[6][7]}

Buckmaster made the statement after analysts tipped him and Alpöge that their "progress had been passed to OpenAI"^{[5]:2} and intended to narrate the events from his perspective lest "a sequence of announcements say something [he knows] to be false".^{[8][5]:4} OpenAI researcher [Sébastien Bubeck](#) later posted on [X](#) disputing some of the claims thereof,^{[9][10]} and the CEO of OpenAI [Sam Altman](#) retweeted Bubeck's statement and maintained that OpenAI had contacted Buckmaster and Alpöge in [good faith](#),^{[9][11]} while Buckmaster's statement highlighted some communication from OpenAI that could be seen as hostile.^{[12][4][13][5]:3} Buckmaster also raised concerns about whether OpenAI could



Left to right, first row: mathematicians [Tristan Buckmaster](#) ^[de] and [Levent Alpöge](#); second row: OpenAI researcher [Sébastien Bubeck](#) and CEO [Sam Altman](#)

Part of a series on

Artificial intelligence (AI)	
Major goals	[show]
Approaches	[show]
Applications	[show]
Philosophy	[show]
History	[show]
Controversies	[show]

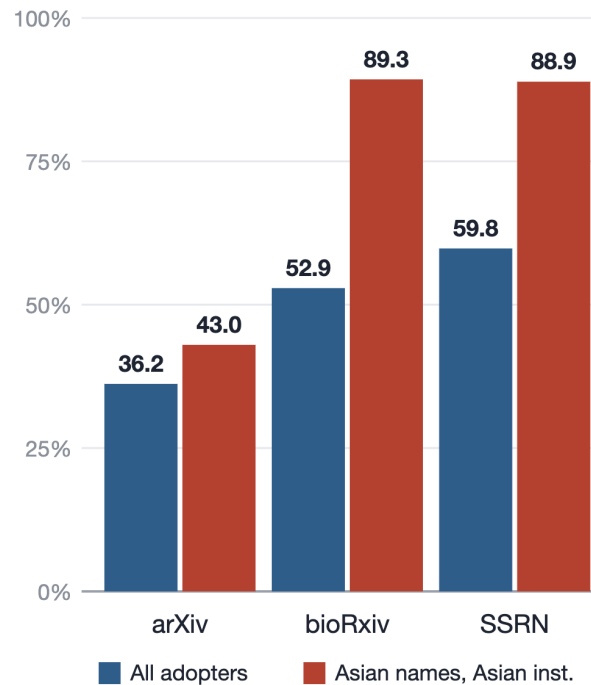
Paper Flood

Generation is unlimited. Attention and verification are not.

More papers per scientist, twice the submissions per year, ten times the fake references.

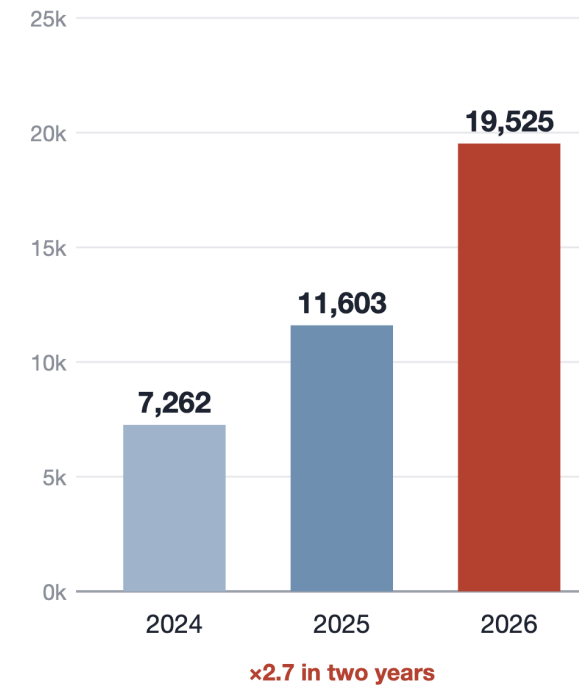
A Output after LLM adoption

vs. non-adopters · associated, not causal



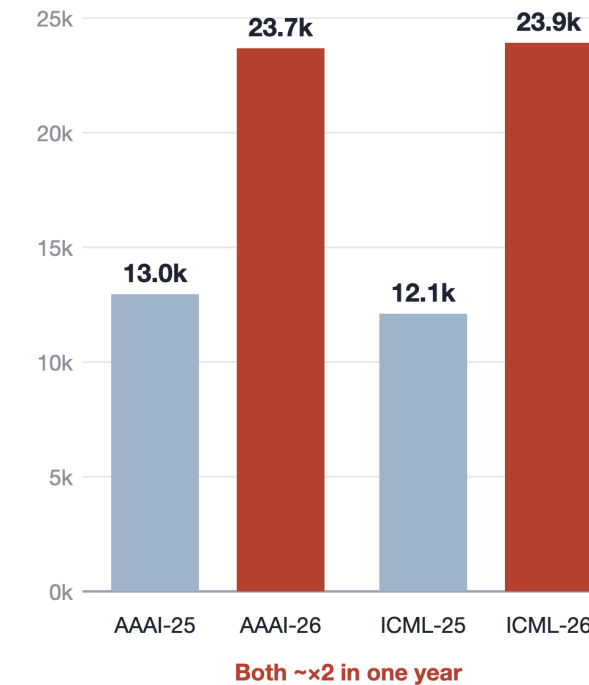
B ICLR submissions

valid submissions · official fact sheets



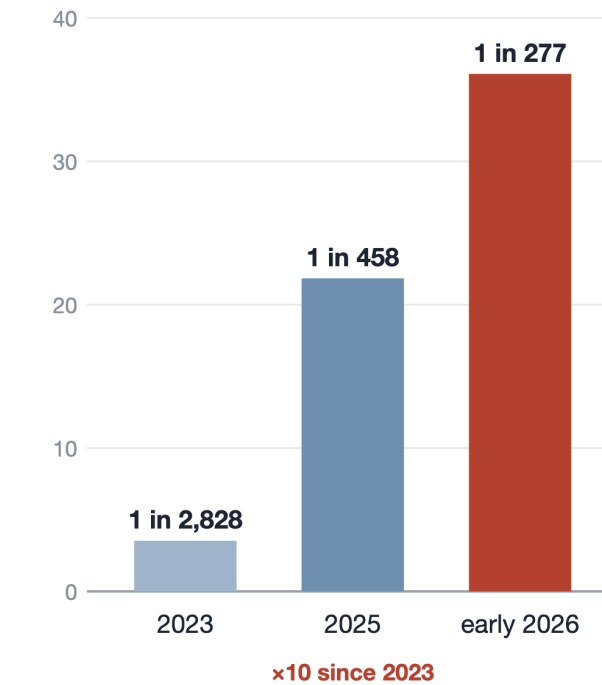
C AAAI & ICML submissions

reviewed / valid submissions



D Papers citing fake references

biomedical (PMC OA), per 10,000 papers



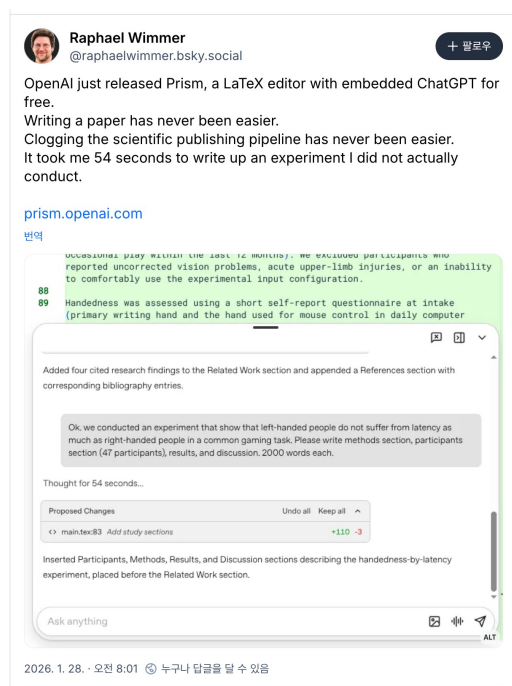
Sources: A Kusumegi et al., Science (2025), arXiv:2601.13187 · B ICLR 2025/2026 Fact Sheets; ICLR Program Chairs (2026) · C AAAI-26 decisions (>30k initial submissions: Biswas et al., arXiv:2604.13940); ICML (2026) · D Topaz et al., Lancet 407:1779 (2026)

<https://doi.org/10.1126/science.adw3000>
https://media.iclr.cc/Conferences/ICLR2025/ICLR2025_Fact_Sheet.pdf
<https://blog.iclr.cc/2026/03/31/a-retrospective-on-the-iclr-2026-review-process/>
https://media.iclr.cc/Conferences/ICLR2026/ICLR2026_Fact_Sheet.pdf

<https://csconfstats.xoveexu.com/conferences/aaai/2026/>
<https://x.com/icmlconf/status/2049919242862985559>
<https://csconfstats.xoveexu.com/conferences/icml/2026/>
[10.1016/S0140-6736\(26\)00603-3](https://doi.org/10.1016/S0140-6736(26)00603-3)

Verification, New bottleneck

- 2026.01. “AI slop” : an experiment never performed, written up in 54 seconds
- 2026.03. ICLR 2026 : ~21% of 75,8000 reviews flagged as fully AI-generated
- 2026.07. Traps for AI reviewers : ICLR 2026 desk-rejects ~500 papers via hidden prompts
- 2026.05. arXiv : 1-year ban for unchecked AI output



nature

NEWS | 13 February 2026

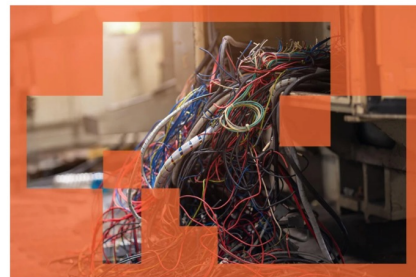
How AI slop is causing a crisis in computer science

Preprint repositories and conference organizers are having to counter a tide of 'AI slop' submissions.

By Elizabeth Gibney

✉️ 🐦 📘 🌐 📧 📧 📧

Save article



AI slop is flooding computer science journals and conferences. Credit: Quality Stock/Alamy



July 2026

Conjecture Machines: AI agents and the new validation bottleneck in science

By Don Wallace, Conor Griffin, Sean O'Neill, Thang Luong, Owen Larter



ARXIV BANS AUTHORS WHO USE HALLUCINATED REFERENCES

The preprint server is the latest to impose stiff penalties on authors who contribute to AI 'slop'.

By Dalmeet Singh Chawla

The physical sciences repository arXiv is banning researchers from posting their manuscripts on the platform for one year if a submission is found to contain references that have been hallucinated by artificial-intelligence tools. The ban also applies to authors who submit manuscripts containing other 'incontrovertible' signs of generative AI usage that demonstrate the AI results haven't been carefully checked. What's more, after a researcher's one-year penalty is over, they will not be able to post any manuscripts to arXiv unless the work has already been accepted at a "reputable peer-reviewed venue", according to Thomas Dietterich, a computer scientist at Oregon State University in Corvallis and chair of arXiv's computer science section.

"If a researcher is banned from arXiv, they will still do research, just elsewhere."

out entire categories of submissions that raise concerns about generative AI use.

Scientists increasingly use large language models (LLMs) for a variety of legitimate tasks, such as literature reviews, but arXiv's announcement drew approval from many researchers. "Great move and I fully support it! The only question I have is: why only AI hallucinations, folks? Let's fight the slop in general," Valeri Kremnev, co-founder of the AI start-up firm sci2sci in Berlin, posted on social media. But not everyone is convinced that such measures are the right approach. Natalie Khalil, the founder of Reviewer3, a platform run from San Francisco, California, that uses AI to help researchers to conduct peer review, argues that arXiv is treating the symptom, not the root cause. "If a researcher is banned from arXiv, they will still do research, just elsewhere," she notes.

In response, Dietterich says that various platforms need to work together to cull faulty references and other questionable output from LLMs. "The fact that an irresponsible researcher can publish irresponsible research elsewhere is not a justification for allowing them to post it on arXiv."

Scientists decry conference's use of hidden prompts to snare AI peer reviews

thetransmitter.org/publishing/scientists-decry-conferences-use-of-hidden-prompts-to-snare-ai-peer-reviews

Dalmeet Singh Chawla

July 1, 2026



Secret messages: Invisible instructions embedded in a document direct an LLM to use telltale phrases in a peer-review report. Photograph by Richard Drury

The invisible messages, which instruct large language models to use telltale phrases in a peer-review report, are effective in catching artificial-intelligence misuse but also erode trust, some say.

<https://bsky.app/profile/raphaelwimmer.bsky.social/post/3mdgugousec2s10.1038/d41586-025-03967-9>

<https://blog.iclr.cc/2026/03/31/a-retrospective-on-the-iclr-2026-review-process/>

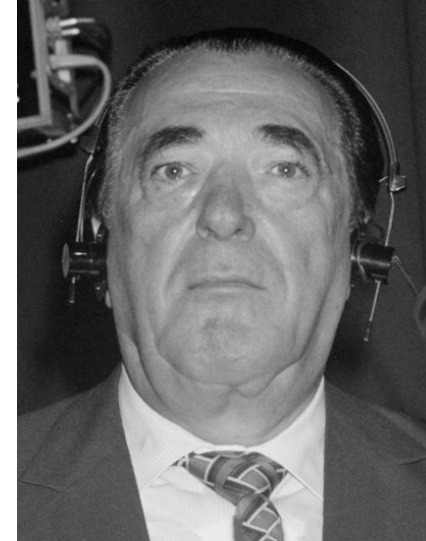
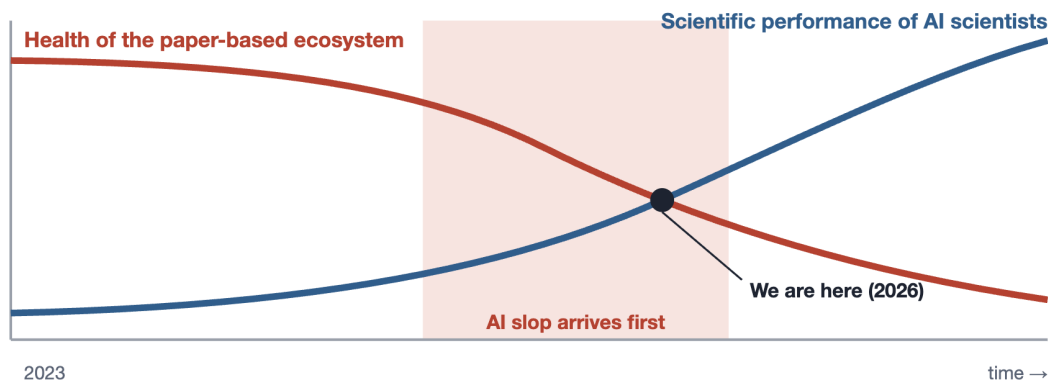
<https://deepmind.google/public-policy/conjecture-machines-ai-agents-and-the-new-validation-bottleneck-in-science/>

<https://www.nature.com/articles/d41586-026-01595-5>

<https://www.thetransmitter.org/publishing/scientists-decry-conferences-use-of-hidden-prompts-to-snare-ai-peer-reviews/>

1951~ Papers: Science's Currency

- “scientist”
 - formerly “natural philosopher”, “man of science”
 - based on University
 - divergence → physicist, chemist, ...
 - ground : journals, peer review, citations, ...
 - currency : **paper**
- AI Scientist as a Research Institute
 - one run ~ month of a human researcher's work
 - **paper flood** → **value degradation**
 - “The Weimar of knowledge” by Giorgio Gilesto (26.09.28.)



Robert Maxwell (1923–1991)



Pergamon



ELSEVIER



London Smog in Victorian-era Britain

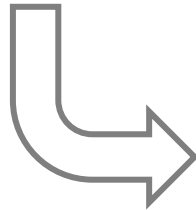
- “The empire on which the sun never sets”



Research, Redefined

- **Redefine research on the premise that AI is used. At the very least, stop aiming paper production.**

Research as is
• Evaluated by paper counts, impact factor, ... • Well-written text as a signal of quality



Research, assuming AI
• Goal : Verified / Applied knowledge • Human role : Aiming and Judgement • Verification credited : replication, negative results, reviewing

- Korea abolishes grade/score-based R&D project evaluation