

DOTE 6635: Artificial Intelligence for Business Research (Spring 2026)

What's New in AI

Renyu (Philip) Zhang

1

What Happened Since We Last Met?

Nature's 10

Ten people who helped shape science in 2025

8 December 2025



Liang Wenfeng: Tech disruptor

After making his name in investing, a Chinese finance wizard founded DeepSeek.

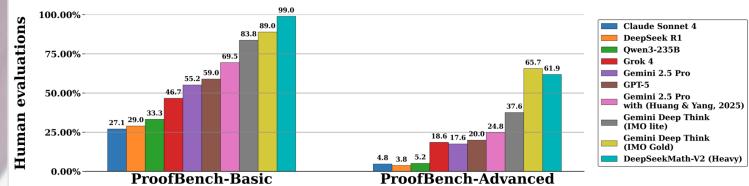
Article | [Open access](#) | Published: 17 September 2025

DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning

[Daya Guo](#), [Dejian Yang](#), [Haowei Zhang](#), [Junxiao Song](#), [Peiyi Wang](#), [Qihao Zhu](#), [Runxin Xu](#), [Ruoyu Zhang](#), [Shirong Ma](#), [Xiao Bi](#), [Xiaokang Zhang](#), [Xingkai Yu](#), [Yu Wu](#), [Z. F. Wu](#), [Zhibin Gou](#), [Zhihong Shao](#), [Zhoushu Li](#), [Ziyi Gao](#), [Aixin Liu](#), [Bing Xue](#), [Binxuan Wang](#), [Bochao Wu](#), [Bei Feng](#), [Chengda Lu](#), ... [Zhen Zhang](#) + Show authors

[Nature](#) 645, 633–638 (2025) | [Cite this article](#)

320K Accesses | 173 Citations | 800 Altmetric | [Metrics](#)



Contest	Problems	Points
IMO 2025	P1, P2, P3, P4, P5	83.3%
CMO 2024	P1, P2, P4, P5, P6	73.8%
Putnam 2024	A1 ~ B4, B5, B6	98.3%

Table 1 | Problems in gray are **fully solved**, while underlined problems received **partial credit**.

Jan 6, 2026

2

karpathy

Home Blog

2025 LLM Year in Review

20 Dec, 2025

2025 LLM Year in Review

By Andrej Karpathy | Dec 19, 2025



2025 has been a strong and eventful year of progress in LLMs. The following is a list of personally notable and mildly surprising "paradigm changes" - things that altered the landscape and stood out to me conceptually.

<https://karpathy.bearblog.dev/year-in-review-2025/>

Jan 6, 2026

3

What Happened Since We Last Met?**1. Reinforcement Learning from Verifiable Rewards (RLVR)****2. Ghosts vs. Animals / Jagged Intelligence****3. Cursor / new layer of LLM apps****4. Claude Code / AI that lives on your computer****5. Vibe coding****6. Nano banana / LLM GUI**

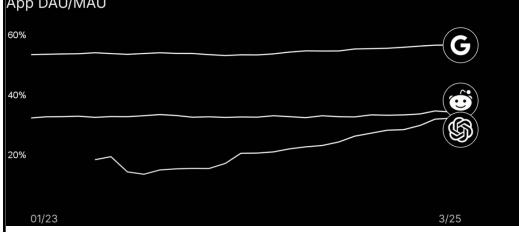
TLDR. 2025 was an exciting and mildly surprising year of LLMs. LLMs are emerging as a new kind of intelligence, simultaneously a lot smarter than I expected and a lot dumber than I expected. In any case they are extremely useful and I don't think the industry has realized anywhere near 10% of their potential even at present capability. Meanwhile, there are so many ideas to try and conceptually the field feels wide open. And as I mentioned on my Dwerkesh pod earlier this year, I simultaneously (and on the surface paradoxically) believe that we will both see rapid and continued progress *and* that yet there is a lot of work to be done. Strap in.

豆包DAU破亿，成字节史上推广费用最少的破亿产品

界面新闻 2025年12月24日 21:56

**AI Now**

App DAU/MAU

**What Happened Since We Last Met?**

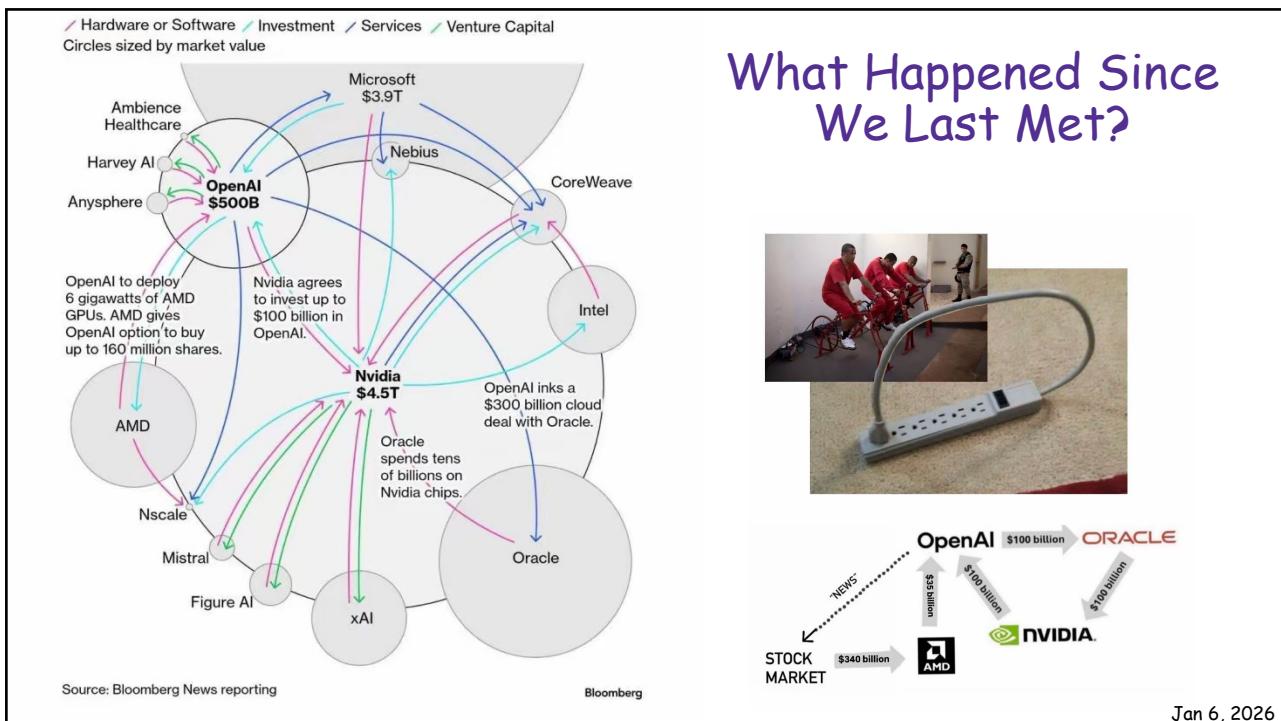
@CryptoDefLord NE 🇺🇸 🇲🇽 🇫🇷 @CryptoDefLord · 7月1日

Transfer window now open for tech guys. This Asian man Jiahui Yu was traded from OpenAI to Meta for \$100m.

Who is next?

科技人才转会窗口现已开放。亚洲球员于嘉辉以1亿美元的价格从OpenAI转会至Meta。

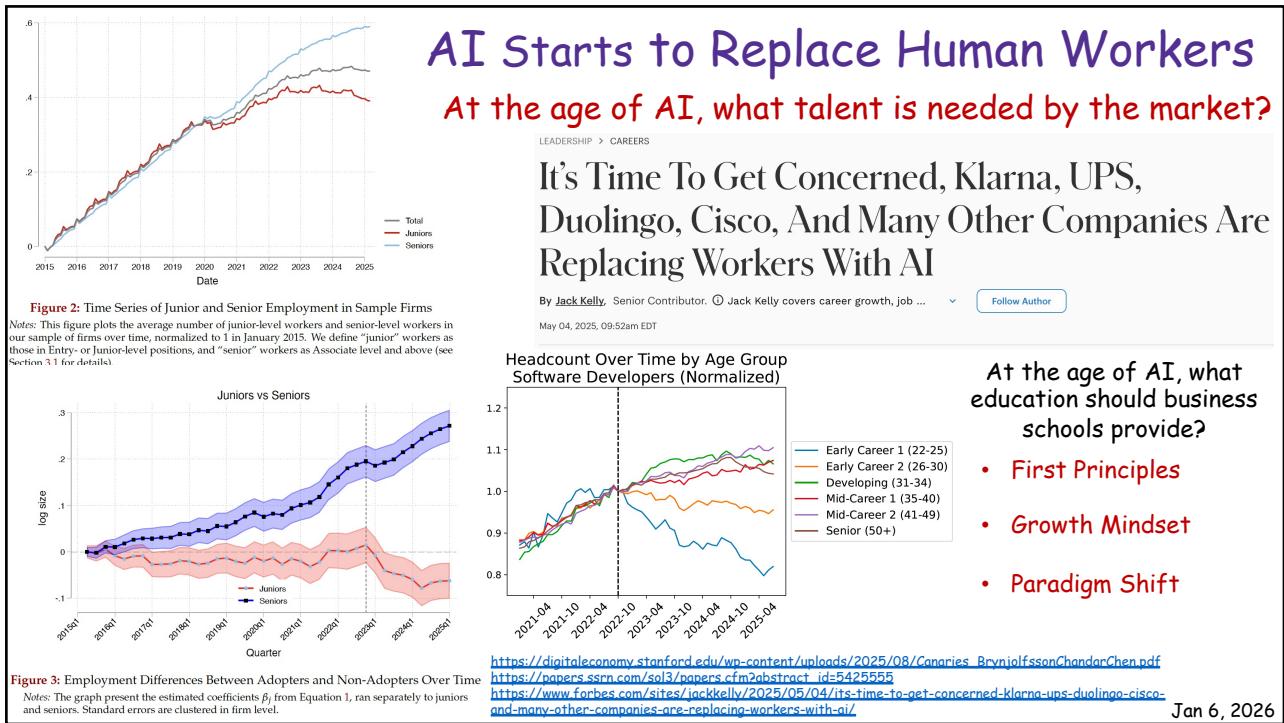
下一个是谁？



5



6



7

Complete Replication of a PNAS Paper

Universal vote-by-mail has no impact on partisan turnout or vote share

<https://github.com/andybhall/vbm-replication-extension>

Daniel M. Thompson^{a,1}, Jennifer A. Wu^{b,1}, Jesse Yoder^{a,1}, and Andrew B. Hall^{a,2}

^aDepartment of Political Science, Stanford University, Stanford, CA 94305; and ^bStanford Institute of Economic Policy Research, Stanford University, Stanford, CA 94305

Edited by Douglas S. Massey, Princeton University, Princeton, NJ, and approved May 6, 2020 (received for review April 15, 2020)

In response to coronavirus disease 2019 (COVID-19), many scholars and policy makers are urging the United States to expand voting-by-mail programs to safeguard the electoral process. What are the effects of vote-by-mail? In this paper, we provide a comprehensive design-based analysis of the effect of universal vote-by-mail—a policy under which every voter is mailed a ballot in advance of the election—on electoral outcomes. We collect data from 1996 to 2018 on all three US states that implemented universal vote-by-mail in a staggered fashion across counties, allowing us to use a difference-in-differences design at the county level to estimate causal effects. We find that 1) universal vote-by-mail does not appear to affect either party's share of turnout, 2) universal vote-by-mail does not appear to increase either party's vote share, and 3) universal vote-by-mail modestly increases overall average turnout rates, in line with previous estimates. All three conclusions support the conventional wisdom of election administration experts and contradict many popular claims in the media.

vote-by-mail | elections | COVID-19 | partisanship

The coronavirus disease 2019 (COVID-19) pandemic threatens the 2020 US election. Fears that the pandemic could deter many people from voting—or cause them to become infected if they do vote—have spurred calls for major electoral reforms. As election administration experts Nathaniel Persily and Charles Stewart put it, “The nation must act now to ensure

Significance

In response to COVID-19, many scholars and policy makers are urging the United States to expand voting-by-mail programs to safeguard the electoral process, but there are concerns that such a policy could favor one party over the other. We estimate the effects of universal vote-by-mail, a policy under which every voter is mailed a ballot in advance of the election, on partisan election outcomes. We find that universal vote-by-mail does not affect either party's share of turnout or either party's vote share. These conclusions support the conventional wisdom of election administration experts and contradict many popular claims in the media. Our results imply that the partisan outcomes of vote-by-mail elections closely resemble in-person elections, at least in normal times.

Model Information

- Model: Claude Opus 4.5 (claude-opus-4-5-20251101)
- Interface: Claude Code CLI
- Date: January 2026

Jan 13, 2026

8

Complete Replication of a PNAS Paper

<https://github.com/andybhall/vbm-replication-extension/blob/main/INSTRUCTIONS.md>

AI-Generated Academic Paper: Replicating and Extending "Universal Vote-by-Mail Has No Impact on Partisan Turnout or Vote Share"

Project Overview

You are tasked with producing a complete academic political science paper by replicating and extending Thompson, Wu, Yoder, and Hall (2020), published in PNAS. The original paper used a difference-in-differences design to estimate the causal effects of universal vote-by-mail (VBM) on partisan electoral outcomes, finding null partisan effects and a modest (~2 percentage point) increase in overall turnout.

Your task:

1. Replicate the original findings using the authors' published replication data and code
2. Extend the analysis by collecting new data for the same three states (California, Utah, Washington) through 2024
3. Test whether the null partisan findings hold in the post-COVID era

Original paper: <https://www.pnas.org/doi/10.1073/pnas.2007249117>

Original replication materials: <https://github.com/stanford-dpl/vbm>

IMPORTANT: Stop-and-Check Points

Throughout this project, there are mandatory STOP AND CHECK points marked with 🔴. At each of these points, you must:

1. Summarize what you have completed
2. Present key outputs for review
3. List any issues or concerns
4. Wait for human approval before proceeding

Do not proceed past a 🔴 checkpoint without explicit approval.

Jan 13, 2026

9

Complete Replication of a PNAS Paper

https://github.com/andybhall/vbm-replication-extension/blob/main/CLAUDE_CODE_PROMPTS.md

Phase 0: Project Setup

Initial Prompt:

I want to replicate and extend Thompson et al. (2020) "Universal Vote-by-Mail Has No Impact on Partisan Turnout or Vote Share" from PNAS. The paper studies California's Voter's Choice Act. I have the original replication data. Please set up the project structure and review the original materials.

Phase 1: Literature Review

Prompt:

Approved, proceed to Phase 1: Literature Review

Phase 2: Replication

Prompt:

Approved, proceed to Phase 2

Phase 3: Extension Data Collection

Prompt:

Approved, proceed to Phase 3

Phase 4: Data Preparation

Prompt:

Approved, proceed to Phase 4

Jan 13, 2026

10

Complete Replication of a PNAS Paper

https://github.com/andybhall/vbm-replication-extension/blob/main/CLAUDE_CODE_PROMPTS.md

Phase 5: Extension Analysis

Prompt:

Approved, proceed to Phase 5

Phase 6: Paper Writing

Prompt:

Approved, proceed to Phase 6

Phase 7: Final Deliverables

Prompt:

Approved, proceed to Phase 7

Bug Fix Session

Prompt:

Can you take a look at the event study? It seems like something may be wrong with the turnout one since it's not showing the same positive effect as all the regressions (which I trust more) are showing

Jan 13, 2026

11

IPOs of Zhipu AI and MiniMax

China's AI Tigers Roar on HKEX

Historic back-to-back IPOs of Zhipu AI & MiniMax signal a new era for global AI capital markets

ZHIPU AI

02513.HK

JAN 8, 2026

MARKET CAP (DAY 1)

HK\$57.9B

DAY 1 GAIN

+13.2%

IPO PRICE

HK\$116.2

FUNDS RAISED

US\$558M

OVERSUBSCRIPTION

1,159x

Focus: Enterprise AI, Tsinghua Spin-off

MINIMAX

0100.HK

JAN 9, 2026

MARKET CAP (DAY 1)

HK\$1,067B

DAY 1 GAIN

+109%

IPO PRICE

HK\$165.0

FUNDS RAISED

US\$619M

OVERSUBSCRIPTION

1,837x

Focus: Consumer AI Apps, Global Reach

<https://www.scmp.com/tech/tech-trends/article/3339301/minimax-and-zhipus-stellar-hong-kong-ipos-supercharge-chinas-ai-ambitions>



HISTORIC MILESTONE

MiniMax becomes first AI company globally to exceed HK\$100B market cap on IPO day.



RECORD SPEED

Fastest AI IPO record worldwide—just 2 years from founding to listing.



MASSIVE DEMAND

Combined retail investors >600k; Record institutional subscription for MiniMax.



GLOBAL CONFIDENCE

Backed by ADIA, Alibaba & Mirae Asset, signaling strong global trust.
Source: HKEX, Company Filings, Reuters, Bloomberg | January 2026

Jan 13, 2026

12

DeepSeek to Disrupt the World Again?

• News • Artificial Intelligence

Insiders Say DeepSeek V4 Will Beat Claude and ChatGPT at Coding, Launch Within Weeks

DeepSeek's upcoming V4 model could outperform Claude and ChatGPT in coding tasks, according to insiders—with its purported release nearing.



By [Jose Antonio Lanz](#)

Edited by [Andrew Hayward](#)

Jan 10, 2026

4 min read

Capability	Description
Long Code Prompts	A significant breakthrough in handling and parsing extremely long code prompts, offering a major advantage for developers working on complex software projects 6 .
Data Pattern Understanding	Improved ability to understand data patterns across the entire training pipeline, with no observed degradation in performance 6 .
Reasoning Ability	The model's outputs are described as more logically rigorous and clear, indicating stronger reasoning capabilities and greater reliability for complex tasks 6 .

<https://tech.yahoo.com/ai/articles/insiders-deepseek-v4-beat-claude-205234497.html>

Jan 13, 2026

13

 Axiom
5,629 followers
1d · ①

AxiomProver solved all 12 Putnam 2025 problems. Today we're releasing the Lean proofs.

We also put together our take on the problems, proof visualizations, and a look at how humans and AI approached things differently. Lots of fun math and Lean.

We sorted our findings into three categories:

Some problems are trivial for humans but tedious for AI. Calculus (A2, B2) and combinatorial constructions (A5) fall here. A positivity lemma humans find obvious takes chunks of Lean to satisfy the formal checker. After a combinatorial construction that's natural to see, our A5 formalization is 2054 lines and took 518 minutes. Combinatorics are friendly beasts to AI and analysis formalization could make one grumpy.

Some problems were surprisingly within reach. AxiomProver doesn't have a Euclidean geometry engine, so we didn't expect it to get B1. It reduced everything to symbols and pushed through. It also handled the combinatorial game theory question (A3) cleanly -- usually a tough domain for AI as we observed in last two years' model performance in the IMO.

Some problems got completely different solutions. On A4, our mathematicians thought algebraically. AxiomProver went geometric. On B4, we found one picture that settles it elegantly. AxiomProver chose 1061 lines of combinatorial bookkeeping. No picture, just grinding through cases until the result falls out.

The one that stunned us: A6. None of our in-house mathematicians finished it. AxiomProver did.

What's hard for humans and what's hard for machines are different questions. Understanding that gap matters as we are closing it.

Grothendieck uses the rising sea metaphor to say, don't attack the problem, raise the water level until it surrounds the land.

When will AI bring in definitions, build theories, choose the right abstraction to make problems dissolve?

We are not there yet. But days like this makes it feel closer.

<https://github.com/AxiomMath/putnam2025>

AI for Math by AxiomProver

Problem	Time to Solve	Tokens Used	Proof Length (Lines)	Theorems Generated	Tactics Used
A1	110 minutes	7,000,000	652	23	561
A2	185 minutes	6,000,000	556	26	581
A3	165 minutes	8,000,000	1,333	78	1,701
A4	107 minutes	8,000,000	960	32	1,107
A5*	518 minutes	9,100,000	2,054	52	3,074
A6*	259 minutes	16,000,000	588	28	670
B1	270 minutes	7,000,000	1,386	49	1,841
B2	65 minutes	2,000,000	417	28	325
B3	43 minutes	2,900,000	340	11	422
B4*	112 minutes	249,000	1,061	23	1,433
B5	354 minutes	18,000,000	1,495	66	1,967
B6*	494 minutes	21,000,000	1,019	30	1,052

"*" means AxiomProver solved the problem in the following day.

Jan 13, 2026

14

Human Performance of Putman 2025																								
SCORE-TO-RANK CORRESPONDENCE												FREQUENCY DISTRIBUTION OF PROBLEM SCORES OF THE TOP 504 CONTESTANTS												
Score	Rank	Score	Rank	Score	Rank	Score	Rank	Score	Rank	Score	Rank	Score	A1	A2	A3	A4	A5	A6	B1	B2	B3	B4	B5	B6
90	1	64	32	49	96	35	202	21	505	7	1335	10	449	93	29	112	5	0	375	81	75	95	1	0
87	2	63	34	48	100	34	208	20	546	6	1344	9	28	10	1	11	3	1	25	76	21	39	0	1
81	3	62	37	47	106	33	220	19	641	5	1373	8	1	99	8	15	3	0	11	17	9	5	0	0
80	4	61	38	46	109	32	238	18	669	4	1428	7	0	11	2	13	0	0	4	7	0	0	0	2
78	6	60	45	45	111	31	254	17	686	3	1557	6	0	0	0	1	0	0	0	0	0	0	0	1
75	9	59	52	44	116	30	272	16	698	2	1760	5	0	0	0	0	0	0	0	0	0	0	0	0
74	10	58	58	43	119	29	295	15	713	1	2408	4	0	0	0	0	0	0	0	0	0	0	0	0
72	11	57	61	42	126	28	313	14	750	0	2807	3	4	36	1	27	0	0	2	12	43	23	0	2
71	13	55	64	41	137	27	328	13	820			2	13	88	8	10	3	0	48	21	70	32	82	2
70	17	54	66	40	150	26	337	12	905			1	3	24	29	52	0	9	16	54	80	14	81	0
69	18	53	69	39	173	25	349	11	1090			0	1	85	154	34	163	65	10	84	43	77	41	29
67	23	52	76	38	180	24	374	10	1171			NA	5	58	272	229	327	429	13	152	163	219	299	467
66	26	51	79	37	194	23	404	9	1274															
65	30	50	87	36	198	22	449	8	1305															

About 4'000 students participated in Putman 2025.

<https://maa.org/wp-content/uploads/2025/03/2024-William-Lowell-Putnam-Competition-Announcement-of-Winners.pdf>

Jan 13, 2026

15

AI Native Mindset: Zero Marginal Cost of Code Compute

Michael Truell  @mntruell

We built a browser with GPT-5.2 in Cursor. It ran uninterrupted for one week.

It's 3M+ lines of code across thousands of files. The rendering engine is from-scratch in Rust with HTML parsing, CSS cascade, layout, text shaping, paint, and a custom JS VM.

It *kind of* works! It still has issues and is of course very far from Webkit/Chromium parity, but we were astonished that simple websites render quickly and largely correctly.

我们在 Cursor 中用 GPT-5.2 构建了一个浏览器。它连续运行了一周。

代码量超过 300 万行，分布在数千个文件中。渲染引擎完全用 Rust 从零打造，包含 HTML 解析、CSS 层叠、布局、文字排版、绘制以及自定义的 JS 虚拟机。

它“基本能跑”！虽然仍存在不少问题，距离 Webkit/Chromium 的水平还很遥远，但简单的网页能快速且基本正确地渲染出来，这让我们感到惊讶。

Course Description

In the last few years, large language models have introduced a revolutionary new paradigm in software development. The traditional software development lifecycle is being transformed by AI automation at every stage, raising the question: how should the next generation of software engineers leverage these advances to 10x their productivity and prepare for their careers?

This course demonstrates that modern AI tooling will not only enhance developer tools but also democratize software engineering for a broader audience. We'll show that software development has evolved from 0-1 code creation to an iterative workflow of plan, generate with AI, modify, and repeat. Students will master both the theory behind traditional software engineering challenges and the cutting-edge AI-powered tools solving them today.

Through hands-on engineering tasks and talks from industry pioneers building these revolutionary tools, you'll gain practical experience with AI-assisted development, automated testing, intelligent documentation, and security vulnerability detection. By the end of this course, you'll have a crisp understanding of how to integrate state-of-the-art LLM models into complex development workflows and avoid common pitfalls.

<https://x.com/mntruell/status/2011562190286045552>; <https://themodernsoftware.dev/>; <https://yage.ai/ai-native-cost-structure.html>

Jan 20, 2026

16

Open or Close in the Age of AI

Open-weight models lag state-of-the-art by around 3 months on average

Frontier open-weight models lag behind the most capable models by an average of 3 months in the [Epoch Capabilities Index](#) (ECI), our holistic measure of model capability. That corresponds to an average ECI gap of around 7 points, similar to the gap between o3 and GPT-5.

Open-weight models lag state-of-the-art by around 3 months on average

Epoch Capabilities Index score

Release Date

CC-BY

epoch.ai

https://www.reddit.com/r/Anthropic/comments/1q8z1to/anthropic_blocking_access_to_thirdparty_apps/

<https://github.com/eigent-ai/eigent>; <https://epoch.ai/data-insights/open-weights-vs-closed-weights-models>

Jan 20, 2026

17

Agentic E-commerce and Beyond

OpenAI
9,658,246 followers
3h ago

We've been working to make powerful AI accessible to everyone through our free product and low-cost subscription tier, ChatGPT Go—which is now available globally in every country where ChatGPT is available.

In the coming weeks, we're also planning to start testing ads in the U.S. for the free and Go tiers, so more people can benefit from our tools with fewer usage limits or without having to pay. Plus, Pro, Business, and Enterprise subscriptions will not include ads.

We're sharing our principles early on how we'll approach ads—guided by putting user trust and transparency first as we work to make AI accessible to everyone.

What matters most:

- Responses in ChatGPT will not be influenced by ads.
- Ads are always separate and clearly labeled.
- Your conversations are private from advertisers.
- Plus, Pro, Business, and Enterprise tiers will not have ads.

Once we begin testing our first ad formats, we look forward to getting your feedback and ensuring that ads can support broad access to AI and keep the trust that makes ChatGPT valuable.

<https://lnkd.in/gPPKPCp>

Our Ad Principles

Mission alignment Our mission is to ensure AI benefits all of humanity; our pursuit of that mission is always in support of that mission and making AI more accessible.
Answer independence Ads do not influence the answers you receive from us. Answers are optimized based on what's most helpful to you. Ads are always separate and clearly labeled.
Conversation privacy We keep your conversations under wraps from advertisers, and we never sell your data to advertisers.
Choice and control You control how your data is used. You can turn off personalization, and we never use your data for ads at any point. You can choose a paid tier that's ad-free.
Long-term value We do not optimize for short-term gains in ChatGPT. We prioritize user trust and long-term value over revenue.

<https://wallstreetcn.com/articles/3763338>; <https://www.linkedin.com/pulse/building-universal-commerce-protocol-ucp-ilya-grigorik-ekemc>

Jan 20, 2026

18

Executive Summary

- On Saturday, Jan 3 we (Andy Hall) released an empirical paper fully created by Claude Code. In this summary, Hall and Straus offer their interpretations of how Claude did, based on a manual audit that Straus carried out independently of Hall and which is detailed below.
- Our subjective conclusion: Claude Code did a remarkably good job at extending the main results of the original paper with extremely limited oversight in less than an hour of work. The small mistakes it made during the replication and extension, which we detail below, are the kind of mistakes we believe a human would also be likely to make. At the same time, the errors it made when going beyond the original paper's approach and providing totally new analyses were significantly more severe, suggesting that there are limits to what parts of a paper Claude Code can currently generate. All in all, while it's clear to us that AI agents require expert oversight to accurately produce new papers, it is also clear to us that this constitutes a major potential change to how empirical research will be conducted moving forward.
- Disclaimer: In response to a large number of inquiries from the community, we have moved quickly to provide this update. Manually extending and verifying an empirical project is very challenging. Just as Claude Code makes mistakes, so too do human researchers. We will be continuing to evaluate this project and will provide updates if we find any further mistakes of note.

https://www.andrewbenjaminhall.com/Straus_Hall_Claude_Audit.pdf

Jan 27, 2026

19

Quality of Coding Agent's Replication

Quality of Coding Agent's Replication

- The main findings:
 - Claude Code correctly replicated the original paper's estimates exactly.
 - It correctly collected new information on the treatment variable with a high degree of accuracy. Claude correctly identifies the first election under the Voters' Choice Act for 29/30 California counties that adopt it. Its only mistake (coding Imperial County as treated in 2024 rather than 2025) was a mistake that we also made when we first constructed the new treatment data ourselves, due to a lack of clarity in the official online source.
 - It correctly collected new CVAP data and we are not able at present to detect any mistakes with its collection of this data.
 - It collected all relevant election data for the only state with new treatment variation (California) with no mistakes that we can currently detect.
 - Its most important error was failing to collect senatorial or gubernatorial data for Utah and Washington in 2020 and 2024 (it did correctly collect presidential data for these two states, as well as 2022 senatorial data). This error of omission has a relatively minor impact on the final vote share and turnout estimates because these states have no variation in the treatment variable in this time period.
 - It chose to define turnout based on total votes cast for the presidential election. While there may be contexts in which this choice is defensible, it is a deviation from the definition in the original paper it was instructed to extend, and causes it to drop some observations it could otherwise include in its subsequent turnout analysis.
 - It correctly estimated the key quantities of interest using the new data it collected (i.e., it ran the correct specifications and reported the results correctly).
- The main quantity of interest from the extended paper is the coefficient on the treatment variable in the quadratic trends specification. Claude Code reported this estimate as 0.003. In our independent, non-AI replication our estimate for this quantity is 0.004.
- A secondary coefficient of interest is the coefficient on the treatment variable in the vanilla specification of the diff in diff with turnout as the outcome variable. Claude Code reported this estimate as 0.026. In our independent, non-AI replication our estimate for this quantity is 0.023.

https://www.andrewbenjaminhall.com/Straus_Hall_Claude_Audit.pdf

Jan 27, 2026

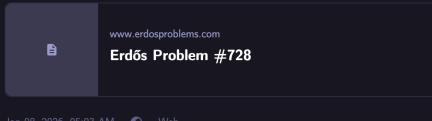
20

How About Theory?

Terence Tao
@tao@mathstodon.xyz

Recently, the application of AI tools to Erdos problems passed a milestone: an Erdos problem (#728 erdosproblems.com/728) was solved more or less autonomously by AI (after some feedback from an initial attempt), in the spirit of the problem (as reconstructed by the Erdos problem website community), with the result (to the best of our knowledge) not replicated in existing literature (although similar results proven by similar methods were located).

This is a demonstration of the genuine increase in capability of these tools in recent months, and is largely consistent with other recent demonstrations of AI using existing methods to resolve Erdos problems, although in most previous cases a solution to these problems was later located in the literature, as discussed in mathstodon.xyz/deck/@tao/11578.... This particular case was unusual in that the problem as stated by Erdos was misformulated, with a reconstruction of the problem in the intended spirit only obtained in the last few months, which helps explain the lack of prior literature on the problem. However, I would like to talk here about another aspect of the story which I find more interesting than the solution itself, which is the emerging AI-powered capability to rapidly write and rewrite expositions of the solution. (1/5)



Jan 08, 2026, 05:03 AM · Web

Last edited Jan 08, 05:08 AM

Terence Tao
@tao

Jan 8

My preference would still be for the final writeup for this result to be primarily human-generated in the most essential portions of the paper, though I can see a case for delegating routine proofs to some combination of AI-generated text and Lean code. But to me, the more interesting capability revealed by these events is the ability to rapidly write and rewrite new versions of a text as needed, even if one was not the original author of the argument.

This is sharp contrast to existing practice where the effort required to produce even one readable manuscript is quite time-consuming, and subsequent revisions (in response to referee reports, for instance) are largely confined to local changes (e.g., modifying the proof of a single lemma), with large-scale reworking of the paper often avoided due both to the work required and the large possibility of introducing new errors. However, the combination of reasonably competent AI text generation and modification capabilities, paired with the ability of formal proof assistants to verify the informal arguments thus generated, allows for a much more dynamic and high-multiplicity conception of what a writeup of an argument is, with the ability for individual participants to rapidly create tailored expositions of the argument at whatever level of rigor and precision is desired.

Presumably one would still want to have a singular "official" paper artefact that is held to the highest standards of writing; but this primary paper could now be accompanied by a large number of secondary alternate versions of the paper that may be somewhat looser and AI-generated in nature, but could hold additional value beyond the primary document. (5/5)

← 6 ↗ 1 ⚡ ⌂ ⌂ ...

<https://mathstodon.xyz/@tao/115855840223258103>; <https://github.com/teorth/erdosproblems/wiki/AI-contributions-to-Erd%5C%91s-problems> Jan 27, 2026

21

An LLM agent-based framework for analytical characterization of Nash equilibria

Wenxuan Liu,^{2,8} Xiyuan Zhou,^{2,8} Xinlei Wang,^{3,8} Yuheng Cheng,^{1,8} Lixin Ye,⁴ Randall Berry,⁵ Leandros Tassiulas,⁶ Jianwei Huang,^{1,7*} and Junhua Zhao^{1,2,*}

¹School of Science and Engineering, The Chinese University of Hong Kong (Shenzhen), Shenzhen 518172, China

²School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798, Singapore

³School of Electrical and Computer Engineering, The University of Sydney, Sydney, NSW 2006, Australia

⁴School of Management and Economics, The Chinese University of Hong Kong (Shenzhen), Shenzhen, 518172, China

⁵Department of Electrical and Computer Engineering, Northwestern University, Evanston, IL 60208, USA

⁶Department of Electrical Engineering, Institute for Network Science, Yale University, New Haven, CT 06520, USA

⁷Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen 518100, Guangdong, China

*These authors contributed equally

*Correspondence: jianweihuang@cuhk.edu.cn (J.H.); zhaojunhua@cuhk.edu.cn (J.Z.)

<https://doi.org/10.1016/j.ynecs.2025.100107>

BROADER CONTEXT

Deriving Nash equilibrium supports decision analysis in economics, energy systems, and public policy, yet traditional analytical methods fail when facing high-dimensional, non-convex, or dynamic games. PrimeNash pioneers the use of large language model agents to automatically derive and verify closed-form equilibria, converting equilibrium search from a manual exercise into a reproducible computational process. By integrating symbolic reasoning, code execution, and formal proof, it bridges AI reasoning with mathematical rigor. This breakthrough reduces the human effort required for equilibrium derivation by more than 80% while maintaining interpretability and auditability. Beyond methodological innovation, PrimeNash empowers transparent modeling of strategic interactions in carbon markets, energy systems, and regulatory design, which are areas critical for climate mitigation and economic resilience. Its agent-based framework marks a foundational step toward AI-driven scientific discovery in game theory and economics, opening pathways to scalable policy analysis and automated theorem-assisted research.

ABSTRACT

In this paper, we introduce PrimeNash, a large language model agent framework that, to our knowledge, is the first to automatically derive closed-form Nash equilibria with machine-checkable proofs. PrimeNash decomposes equilibrium search into three modules, strategy generation, payoff evaluation, and equilibrium proof, and it couples multi-agent reasoning with symbolic code execution to handle high-dimensional strategies, intertemporal recursion, and discontinuous, non-convex payoffs. Across seven canonical models, PrimeNash solves all static games and 70% of dynamic games in which success is defined as obtaining a symbolic closed-form solution that passes automated equilibrium checks. Notably, it produces the first analytical solution to a carbon market model previously lacking a closed-form characterization. The framework reduces manual derivation effort while preserving reproducibility and auditability through generated code and proof artifacts. By turning equilibrium derivation from a bespoke manual exercise into a transparent, scalable pipeline, PrimeNash extends the toolkit for economic analysis and opens new avenues for studying complex strategic systems from climate policy to financial markets.

[https://www.cell.com/nexus/pdfExtended/S2950-1601\(25\)00054-3](https://www.cell.com/nexus/pdfExtended/S2950-1601(25)00054-3)

LLM Agent for Stylized Modeling

Jan 27, 2026

22

Should AI be Adopted in Clinical Medicine



<https://zhuanlan.zhihu.com/p/1996731913570916212>

Jan 27, 2026

23

How About Legal?

BEST PRACTICES IN COURTS & ADMINISTRATION

GenAI hallucinations are still pervasive in legal filings, but better lawyering is the cure

Zach Warren Senior Manager / Legal Enterprise Content / Thomson Reuters Institute

18 Aug 2025 · 8 minute read

f Generative AI hallucinations continue to plague unwary attorneys and pro se litigants, recent research finds, emphasizing the ongoing need for careful verification of case citations

X

核心还是25年AI大爆发下，以豆包*为首的AI模型（背后的公司）不断推广，导致普通民众对AI的认可度已经非常高了。

甚至部分人会觉得“AI比你们（律师）还要懂啊，还免费”

这两年整个经济都在下行，纠纷案其实更多了。

以前往往会觉得律师会发这些“烂财”，毕竟纠纷越多，律师出场机会越多嘛，普通人还是会咬咬牙请个律师来增加起诉信心的。

但现在呢，“AI快，AI好，AI又快又好”

以前普通民众可能没得选，因为不懂法。

现在呢？

他们随便打开一下豆包或者Deepseek，输入案情后，AI不仅“哐哐”给出法律分析，还能直接写起诉状、证据清单。

尤其在专业律师的加持下，很多写问案、诊案然后写法律文书的智能体层出不穷，直接说完后，全部材料“哒哒哒”就出来了。

“嘿，和找过的律师写得一样好！”（毕竟那位律师也有可能是用AI生成的文书）

于是，他们拿着AI生成的材料直接就可以去立案了。

虽然法庭上可能磕磕绊绊（让法官头疼去吧），但对于那些事实清楚、标的额不大的民事案件，法官看材料其实就够了。

结果就是，大量的初级案源直接蒸发了。

<https://www.zhihu.com/question/659369541>; <https://www.thomsonreuters.com/en-us/posts/technology/genai-hallucinations/>

Jan 27, 2026

24