

CYCLES OF POWER: GEOPOLITICAL STRAIN

AS THE YEAR OF THE FIRE HORSE UNFOLDS, AND US MARKETS TRAIL...



GLOBAL INVESTORS MUST ASK: HAVE WE BEEN BACKING THE WRONG FIRE HORSE?

A New AI-ge is dawning

GLOBAL INVESTORS ARE WITNESSING...

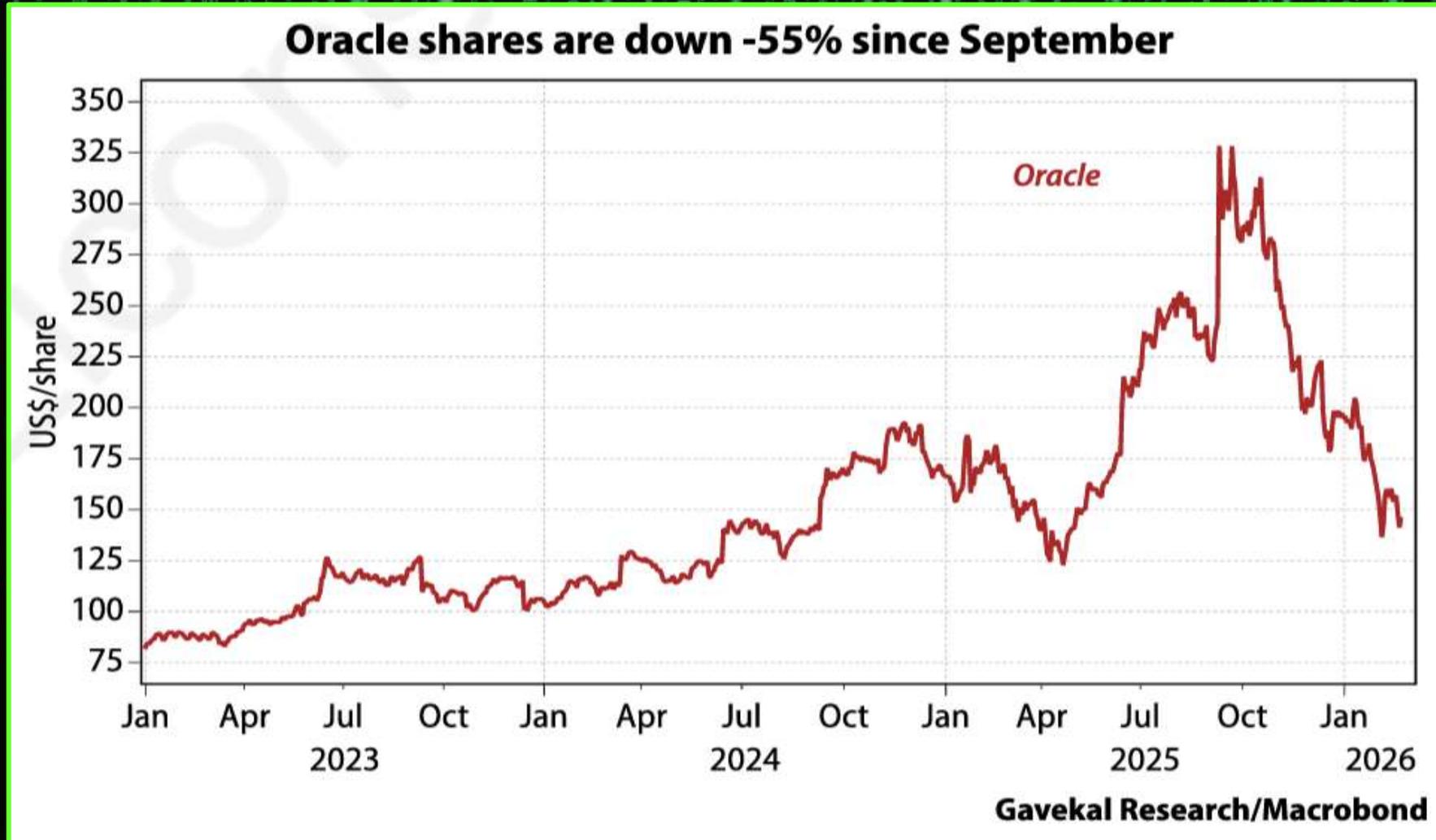
...an arm-wrestle (**not** just in AI) between China and the US...

...heightening risks in debt, equity and currency markets...

...as well as the need to be promote new investment priorities...

All this is hAppening as the US is getting economically weaker

A quick recap: Where do US markets stand?

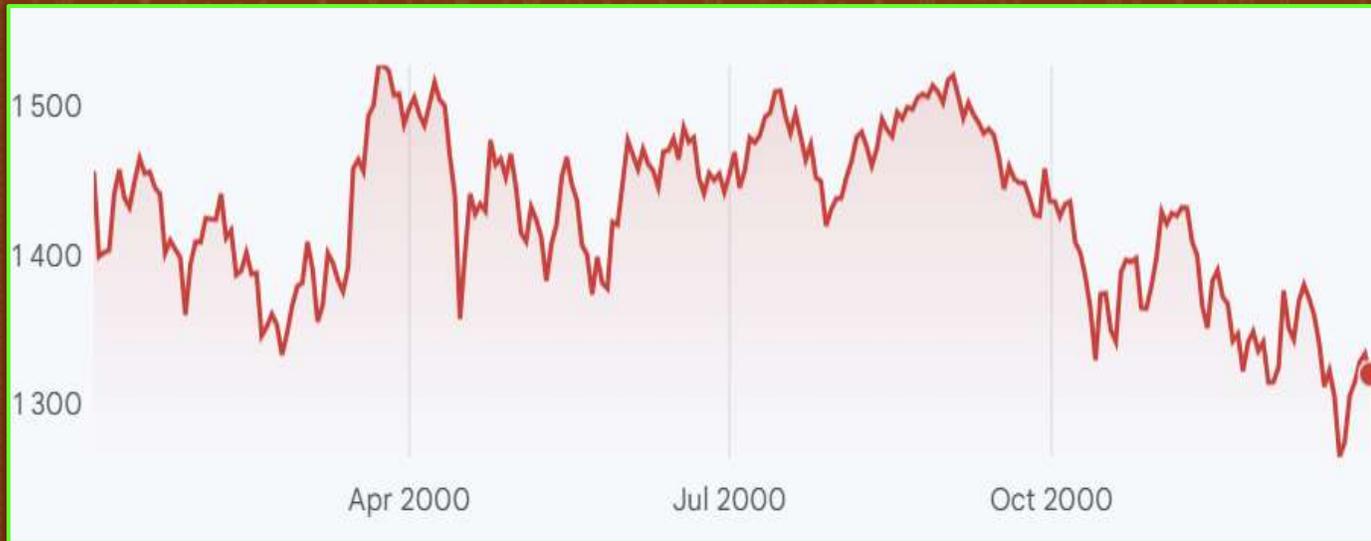


By the pricking of my thumbs, something wicked this way comes

S&P 500: What is going on? Shades of 2000?

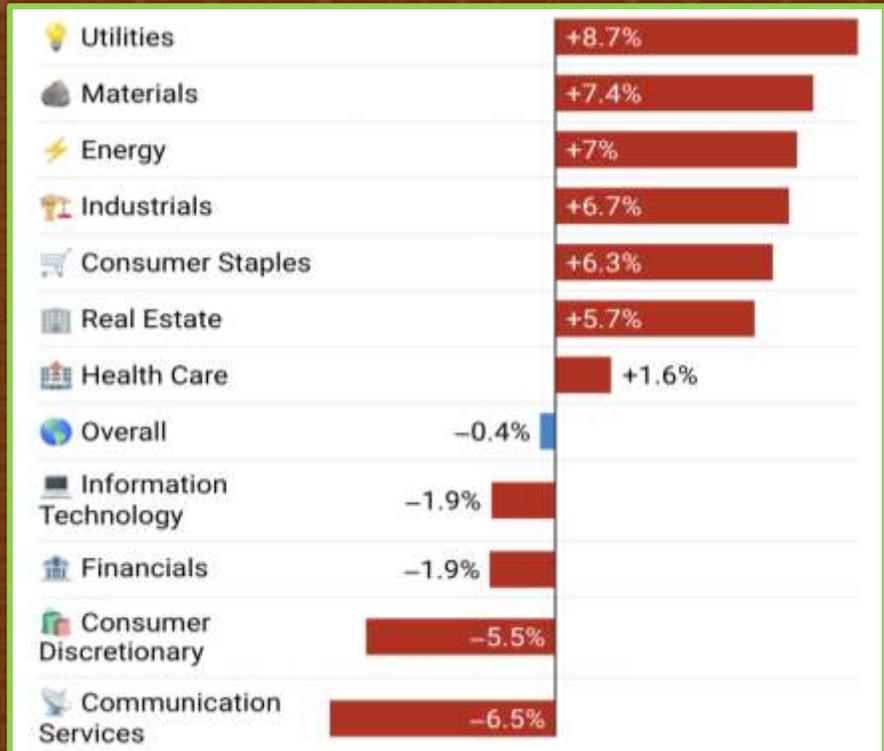
The 2000 dot.com bubble burst in **TWO STAGES**.

1. **STAGE 1:** 10.03.2000; Index = 1395
2. Rotation to Old Economy; Index ATH 20.03.2000: 1528



3. **STAGE 2:** 01.09.2000: Index = 1521, 0.5% below ATH 1528
4. Year End Index = 1320, down 13% from 01.09.2000
5. Subsequent low point: 10.10. 2002: Index = 769, down 50% from ATH

America is becoming a petrostate



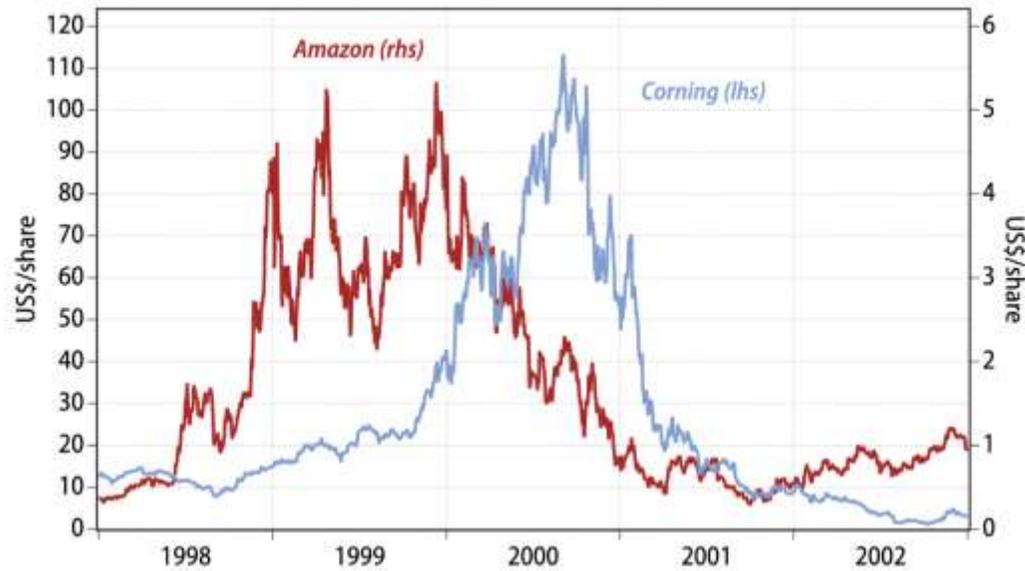
S&P500 February 2026 Sector Performance

Mag 7 Index peaked 29 October 2025; S&P peaked 28 January 2026

Sources: FT, Spectator

Proceed with extreme caution...

In 2000, hardware plays continued to make highs after dot.coms peaked



While software has slumped, semiconductor plays have marched higher



Watch for slow-witted stocks doing a Wile. E. Coyote

The tech 'periphery' is crumbling...



Software stocks crash

CrowdStrike	-20.3%	Adobe	-49.9%
Palo Alto Networks	-27.1%	Snowflake	-50.0%
ADP	-30.3%	workday	-51.6%
FICO	-34.8%	servicenow	-53.9%
S&P Global	-35.4%	DocuSign	-56.7%
SAP	-39.5%	Wolters Kluwer	-65.9%
intuit	-40.2%	Atlassian	-76.3%
accenture	-40.2%	monday.com	-79.5%

Credit stocks crash

% Below All-Time High	S&P 500 \$SPY: -2%
Brookfield \$BAM: -24%	Carlyle \$CG: -29%
Apollo \$APO: -40%	TPG \$TPG: -40%
Ares \$ARES: -42%	Blackstone \$BX: -42%
StepStone \$STEP: -42%	KKR \$KKR: -45%
Hamilton Lane \$HLNE: -50%	Blue Owl \$OWL: -59%



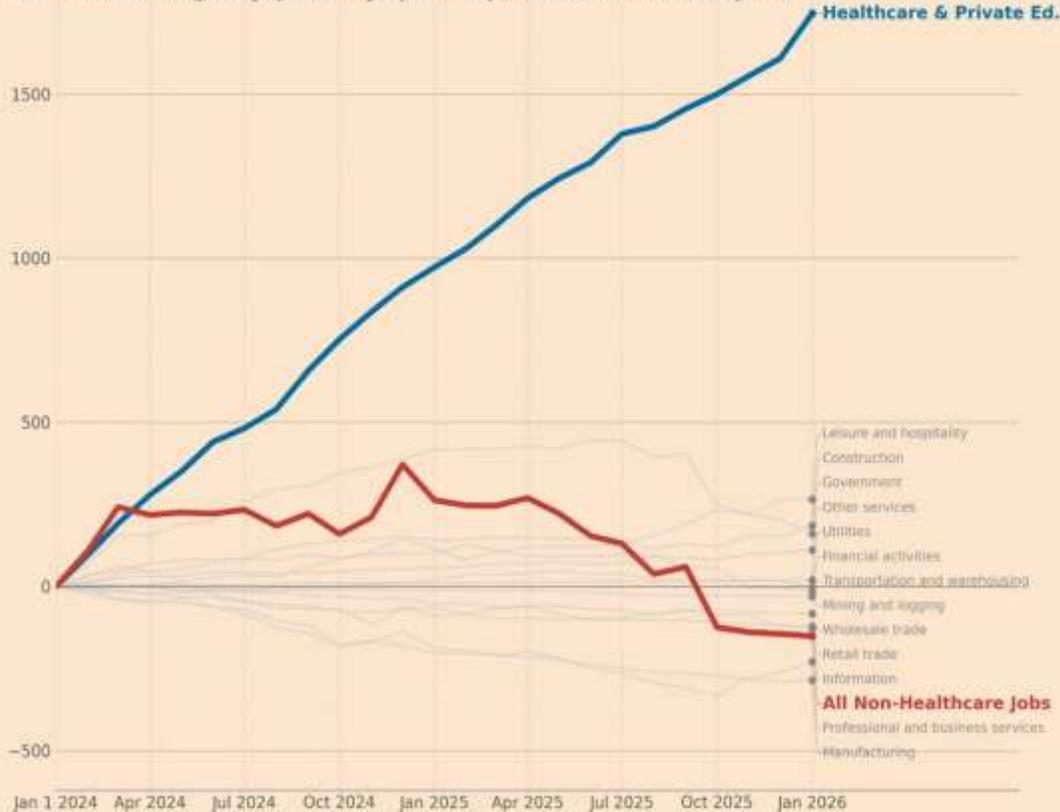
Casualties of the AI war raging INSIDE the US are everywhere

Sources: FT, Cervknowledge, Charlie Biletto

I don't buy the 'Roaring US Economy!' narrative...

If you exclude healthcare employment, the U.S. has actually lost jobs since 2024

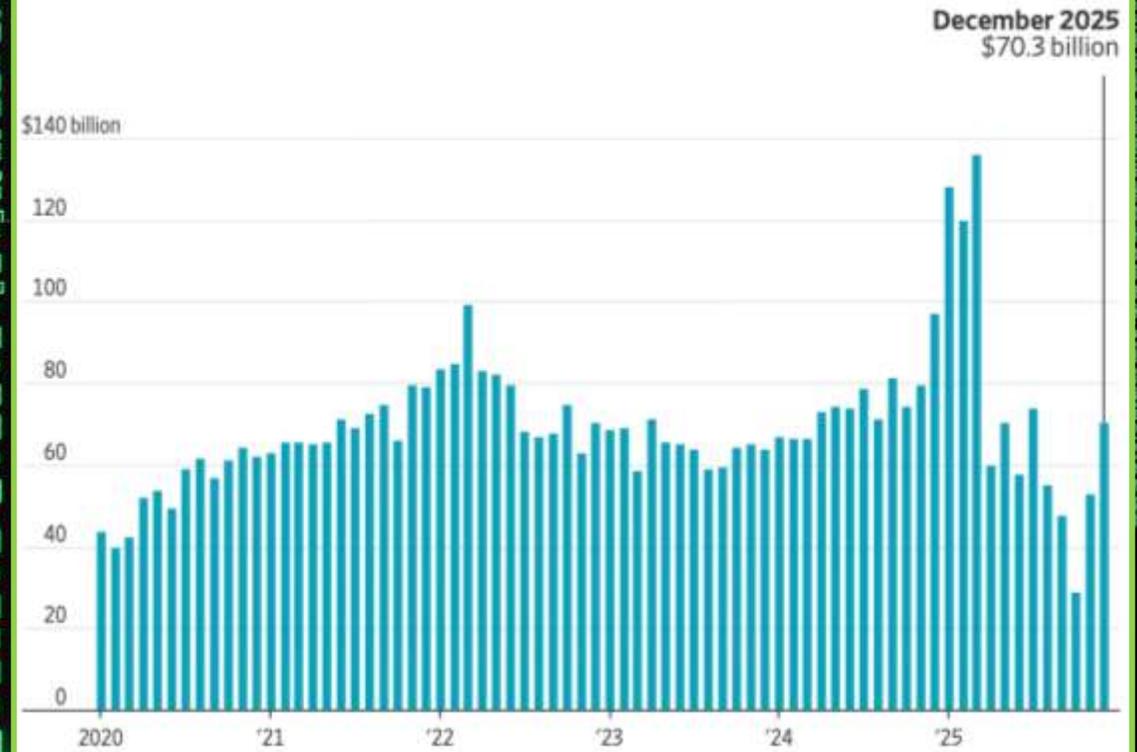
Cumulative change in payroll employment by sector (Thousands of jobs)



Source: U.S. Bureau of Labor Statistics (BLS) as of Feb 11, 2026

In 2025, Trade Deficit in Goods Reached Record High

U.S. trade deficit, goods and services, monthly

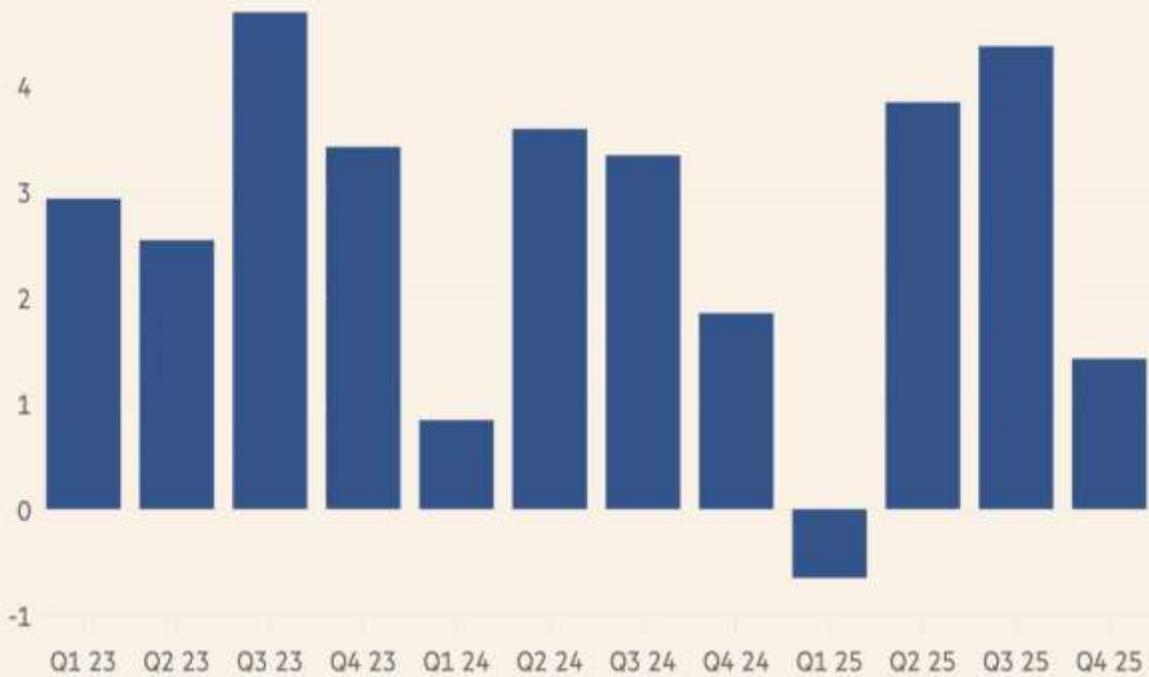


The data does not support this rosy assessment

Sources: BLS, New York Times, Census Bureau

Beneath the glossy narrative...

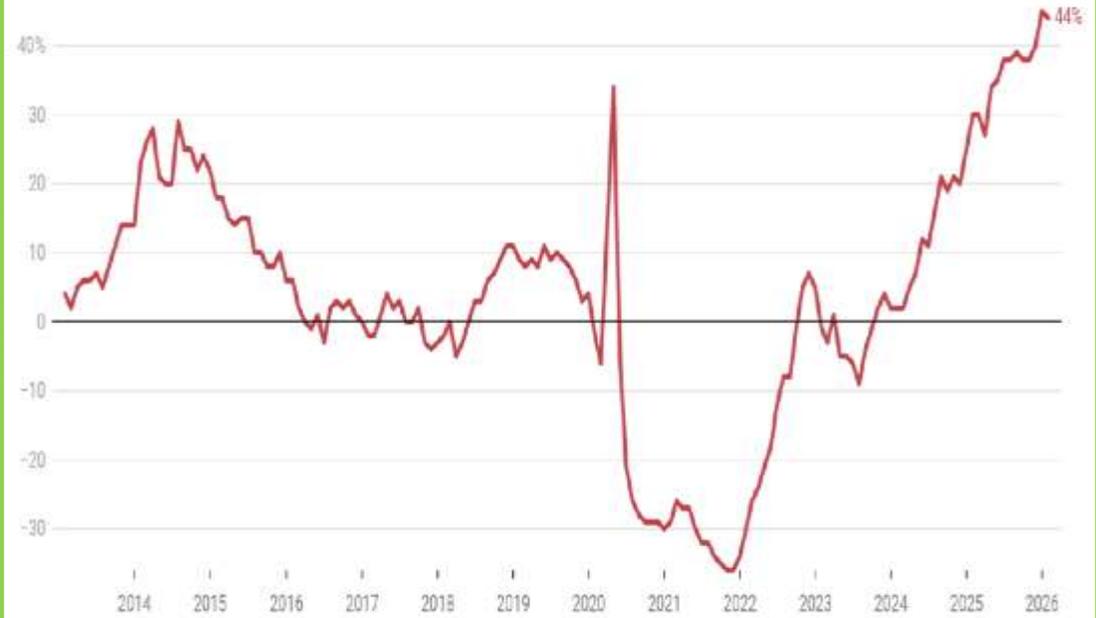
US GDP growth falls sharply to 1.4% rate in fourth quarter



Why American Housing Markets Have Stalled

There Are 44% More Home Sellers Than Buyers

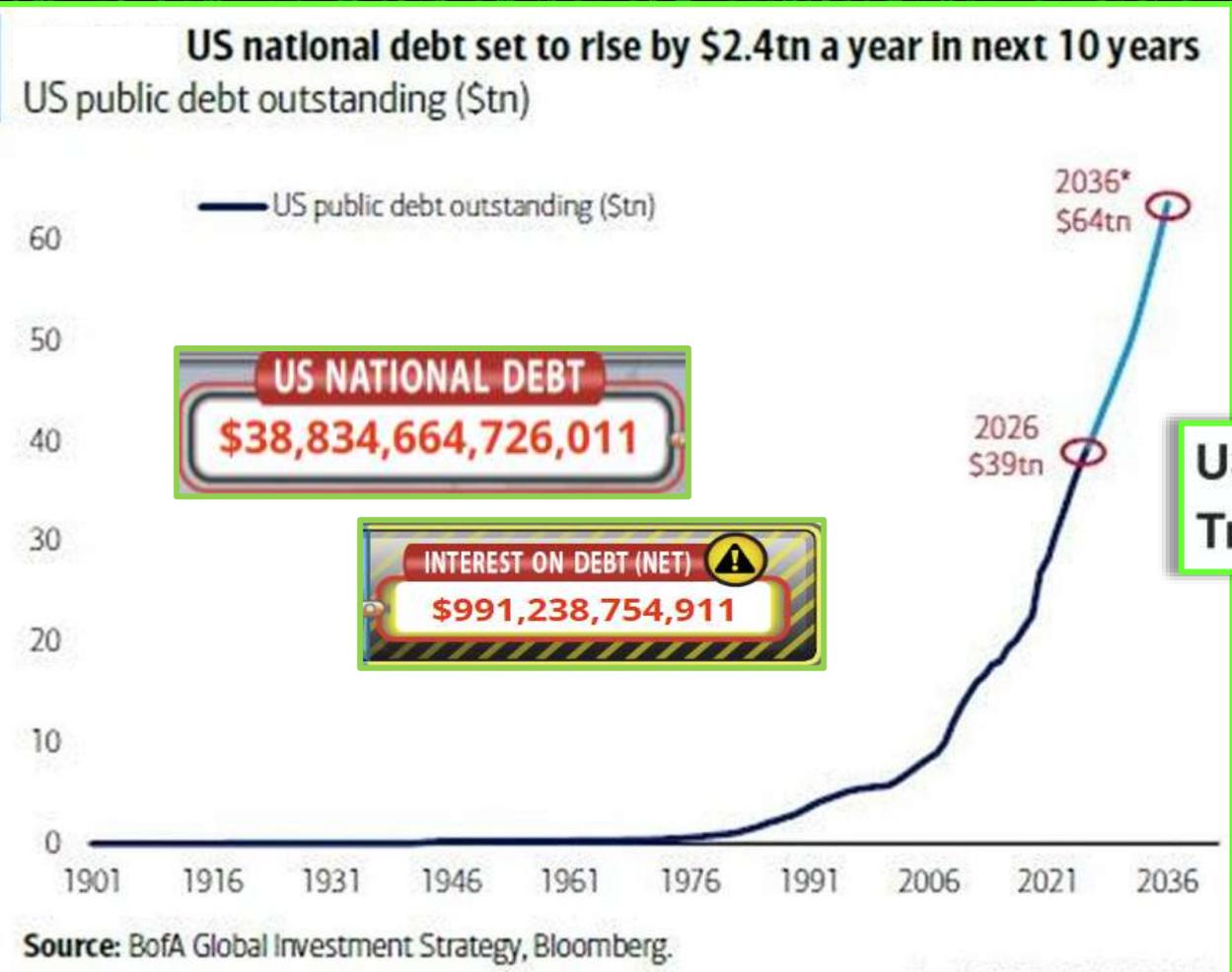
% more sellers than buyers actively in the market



...GDP growth is slowing, housing market has plateaued

Sources: FT, Redfin

...if anything, national debt growth is accelerating...



America borrowed \$43.5 billion a week in the first four months of the fiscal year, with debt interest on track to be over \$1 trillion for 2026

US budget deficit to keep growing amid Trump tax cuts, tariffs, CBO forecasts show

Days before the national debt is due to hit \$39 trillion, President Trump didn't mention it once during the longest State of the Union ever

Debt risen by \$2.5 trillion in 13 months: forecast FY 2026 \$1.9 trillion deficit

Sources: BoA, Reuters, Fortune

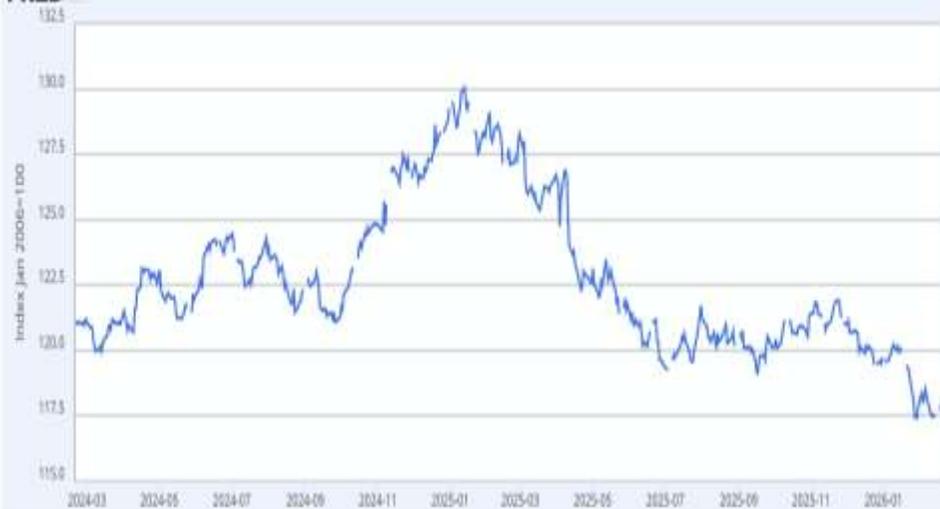
..US Global Indebtedness rises as the US Dollar falls

Opinion US equities

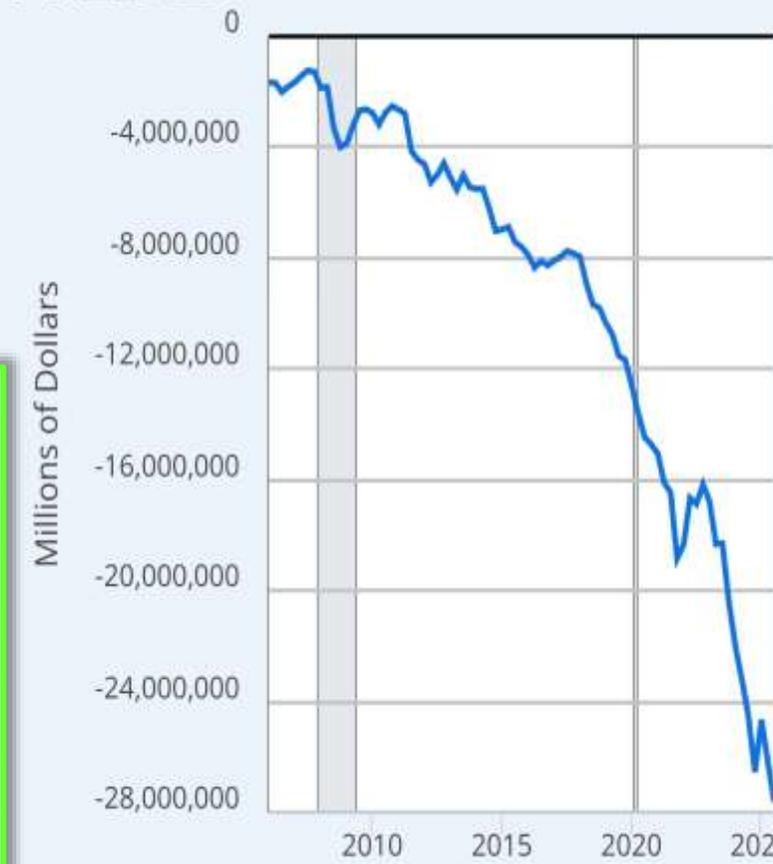
The death of the Trump trade

Investor backlash against US markets appears to be real

FRED Nominal Broad U.S. Dollar Index



FRED U.S. Net International Investment Position

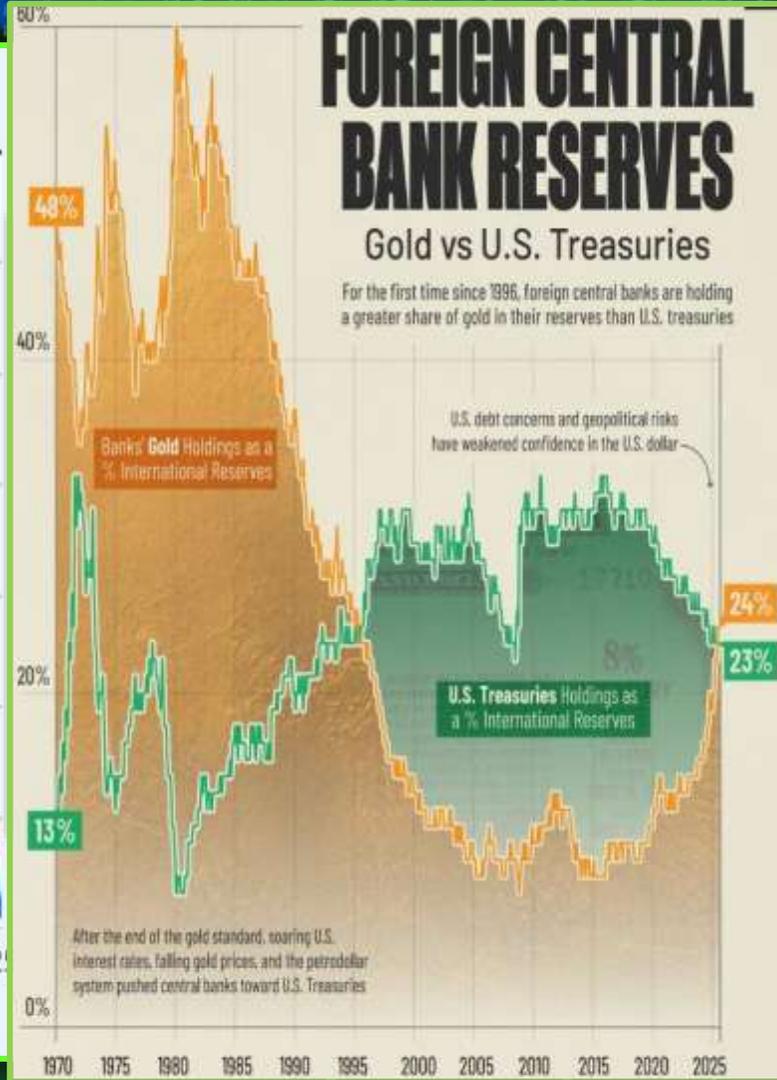


Source: U.S. Bureau of Economic Analysis via FRED®

FOREIGN CENTRAL BANK RESERVES

Gold vs U.S. Treasuries

For the first time since 1996, foreign central banks are holding a greater share of gold in their reserves than U.S. treasuries



Foreigners are having second thoughts about investing in the US

Sources:
FT, FRED
Visual Capitalist

Something profound is moving out there in the Deep...



The world's best surfers pride themselves on swimming with sharks...
...but WHALES?

Yes, plenty of domestic reasons for market unease...



...but is there something DEEPer unsettling markets?

The market: Steeling itself against DeepSeek shock 2?

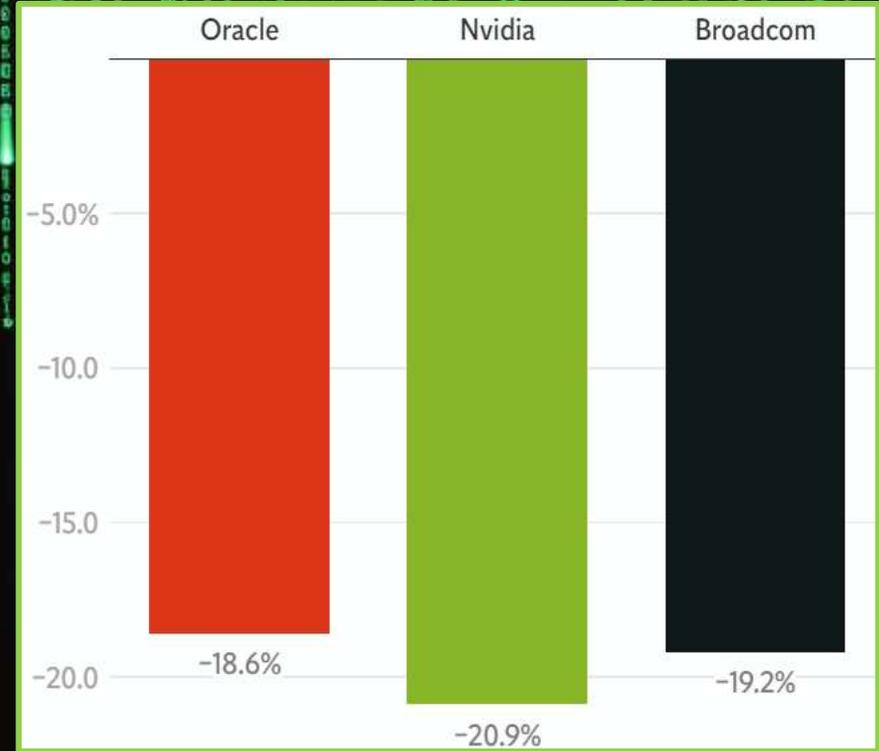
DeepSeek sparks AI stock selloff; Nvidia posts record market-cap loss

Exclusive: China's DeepSeek trained AI model on Nvidia's best chip despite US ban, official says

Anthropic Furious at DeepSeek for Copying Its AI Without Permission, Which Is Pretty Ironic When You Consider How It Built Claude in the First Place

"They robbed the robbers."

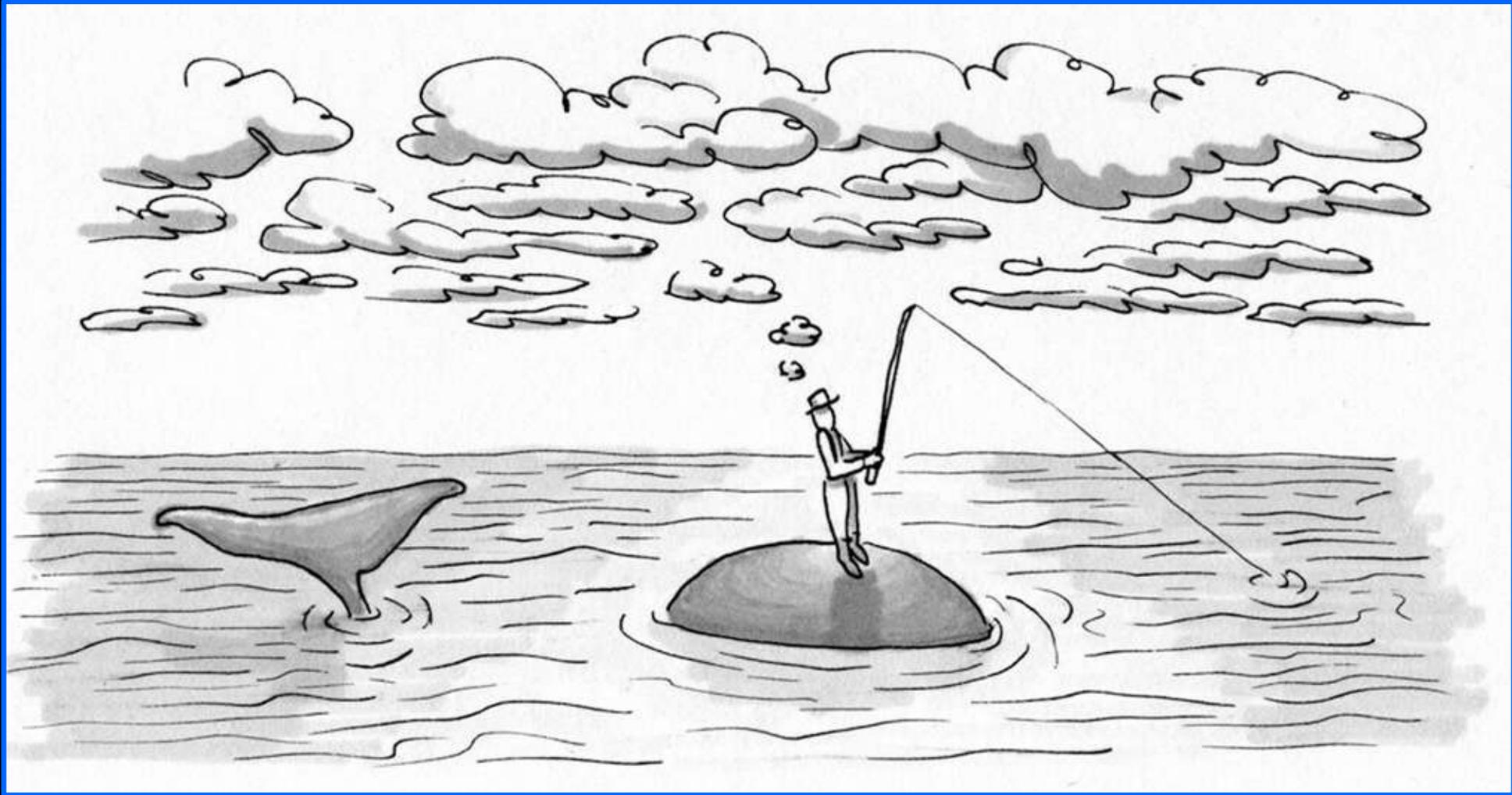
Price performance on 27.01.25



The Flak is flying!

Sources:
Reuters, Nasdaq
Futurism

US AI in 2025...



Sometimes what looks like solid ground is something else entirely...

To start, a few techie-calities...



What is an AI brAIn?

What makes an AI brAIn?

Building Artificial Intelligence

From Training Chaos...

The Challenge:

AI must find signal in noise at unimaginable scale.

The Old Way... Brute Force:

Overwhelming compute, like refining ore with dynamite.

The unstable **Inverted Pyramid**



...to Inference Clarity

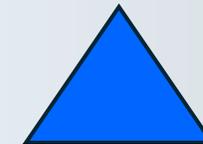
The Foundation:

Optimizing the entire transformation process.

The New Way... Elegance:

Design a brain that efficiently absorbs, intelligently structures.

The stable **Pyramid**



China's new stack nnHC – a quartet of MLA, mHC, Engram, DualPath – achieves comparable intelligence at ~1/20th the inference cost.

As 2026 began, the DeepSeek whale surfaced 3 times...

31.12.25
MHC:
reinvents
training

13.10.26
Engram
reinvents
memory

04.01.26
R1.5
Integrates
all advances



MHC: Reinventing Training
arXiv:2512.24001v1 [cs.LG] 31 Dec 2025

Abstract

Recent studies on the efficiency of training large language models (LLMs) have focused on the model architecture and training data, but the training process itself remains largely unchanged. This paper introduces a novel training paradigm, the Multi-Head Constrained Hyper-Connections (MHC), which aims to improve the training efficiency and model performance by introducing a novel training paradigm. The MHC framework introduces a novel training paradigm, the Multi-Head Constrained Hyper-Connections (MHC), which aims to improve the training efficiency and model performance by introducing a novel training paradigm. The MHC framework introduces a novel training paradigm, the Multi-Head Constrained Hyper-Connections (MHC), which aims to improve the training efficiency and model performance by introducing a novel training paradigm.



R1.5: Integrating All Advances
arXiv:2601.12788v1 [cs.LG] 4 Jan 2026

Abstract

General reasoning represents a long-standing and intractable challenge in artificial intelligence. Recent breakthroughs, exemplified by large language models (LLMs) like GPT-4o, Gemini 1.5 Pro, and Claude 3.5 Sonnet, have achieved state-of-the-art performance on a wide range of reasoning tasks. However, this success is largely contingent upon extensive human-curated annotations, and model capabilities are still largely confined to specific reasoning paradigms. In this work, we show that the reasoning abilities of LLMs can be significantly enhanced through a novel training paradigm, the Reinforced Reasoning (R1.5), which introduces a novel training paradigm. The R1.5 framework introduces a novel training paradigm, the Reinforced Reasoning (R1.5), which introduces a novel training paradigm. The R1.5 framework introduces a novel training paradigm, the Reinforced Reasoning (R1.5), which introduces a novel training paradigm.



Engram: Reinventing Memory
arXiv:2610.13141v1 [cs.LG] 13 Oct 2026

Abstract

While memory is a fundamental component of intelligence, current models struggle to learn from past experiences and apply that knowledge to new situations. This paper introduces a novel training paradigm, the Engram, which aims to improve the model's ability to learn from past experiences and apply that knowledge to new situations. The Engram framework introduces a novel training paradigm, the Engram, which aims to improve the model's ability to learn from past experiences and apply that knowledge to new situations. The Engram framework introduces a novel training paradigm, the Engram, which aims to improve the model's ability to learn from past experiences and apply that knowledge to new situations.

The silent species was changing the deep current of AI

And just when you thought it was safe to go back in the water...



DeepSeek-OCR 2: Visual Causal Flow

Haoran Wei, Yaofeng Sun, Yukun Li

DeepSeek-AI

Abstract

We present DeepSeek-OCR 2 to investigate the feasibility of a novel encoder—DeepEncoder V2—capable of dynamically reordering visual tokens upon image semantics. Conventional vision-language models (VLMs) invariably process visual tokens in a rigid raster-scan order (top-left to bottom-right) with fixed positional encoding when fed into LLMs. However, this contradicts human visual perception, which follows flexible yet semantically coherent scanning patterns driven by inherent logical structures. Particularly for images with complex layouts, human vision exhibits causally-informed sequential processing. Inspired by this cognitive mechanism, DeepEncoder V2 is designed to endow the encoder with causal reasoning capabilities, enabling it to intelligently reorder visual tokens prior to LLM-based content interpretation. This work explores a novel paradigm: whether 2D image understanding can be effectively achieved through two-cascaded 1D causal reasoning structures, thereby offering a new architectural approach with the potential to achieve genuine 2D reasoning. Codes and model weights are publicly accessible at <http://github.com/deepseek-ai/DeepSeek-OCR-2>.

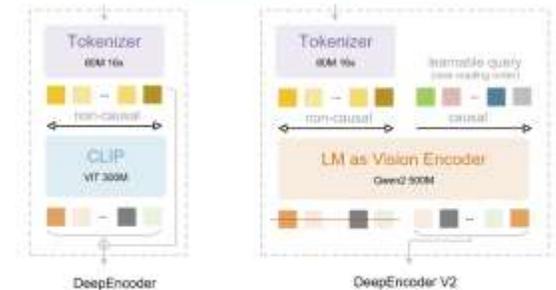


Figure 1 | We substitute the CLIP component in DeepEncoder with an LLM-style architecture. By customizing the attention mask, visual tokens utilize bidirectional attention while learnable queries adopt causal attention. Each query token can thus attend to all visual tokens and preceding queries, allowing progressive causal reordering over visual information.



On 20.01.26, DeepSeek taught its LLM to *SEE* documents.
To *SEE* as humans *SEE*: not scan but understand.

As the Fire Horse rode, Qwen 3.5 ran ahead of DeepSeek

**Alibaba launches Qwen 3.5, Claims
AI model outperforms US Rivals**

Qwen: The Swiss Army Knife of LLMs
Jack of All Trades, Master of Quite a Few.



Alibaba's Qwen family hits 700 million downloads to lead global open-source AI adoption   South China Morning Post

Celestic
Anthropic

API key

DUALPATH, THE KEYSTONE TO VIA V IV

DualPath: Breaking the Storage Bandwidth Bottleneck in Agentic LLM Inference

Yongqiang Wu^{1†}, Shaoyuan Chen^{2,3}, Yuxuan Zhou^{1†}, Bilin Huang¹,
Taoxian Tan², Wentao Zhang², Liyue Zhang², Shangyan Zhou², Yuxuan Liu²,
Shanlei Zhou², Mingqiang Zhang², Xin Jia¹, Fanpan Huang⁴

¹School of Computer Science, Fudan University ²Xiang'an University ³DeepSeek AI

ABSTRACT

The performance of multi-turn, agentic LLM inference is increasingly dominated by KV-Cache storage I/O rather than computation. In parallel, heterogeneous architectures, loading the massive KV-Cache from external storage creates a fundamental storage I/Os on parallel engines to serve bandwidth-sensitive, while those on clouding engines remain idle. This asymmetry severely constrains overall system throughput.

We present DualPath, an inference system that breaks this bottleneck by introducing dual-path KV-Cache loading: beyond the traditional storage-to-pull2-path, DualPath enables a novel storage-to-decode-path, in which the KV-Cache is loaded into decoding engines and then efficiently transferred to pull2 engines via RDMA over the campus network. DualPath enables this optimized data path – which effectively avoids network congestion and avoids interference with latency-critical model execution communication – with a generic scheduler that dynamically balances load across pull2 and decode engines.

Our evaluation on three models with realistic agentic workloads demonstrates that DualPath improves offline inference throughput by up to 1.5x on server inference systems. It can also improve online serving throughput by an average factor of 1.9x without training LLM.

1 INTRODUCTION

Large Language Models (LLMs) are rapidly evolving from single-turn chatbots [9, 10] and decision assistants [11] into agentic systems that can autonomously plan, execute, and iteratively refine tasks through multi-turn interactions [7, 12, 13, 14, 15]. In such settings, an LLM no longer serves solely as a tool, but as a participant in long-running sessions where context accumulation over time [16] is key to applications because increasingly pertinent, such as LLM inference has emerged as a critical workload in production systems, ranging from online assistants [17, 18] to autonomous task agents [19, 10].



Figure 1: Comparing inference (left) and DualPath (right).

This paradigm shift in applications has driven a significant transformation in LLM inference workloads. From traditional human-LLM interaction to human-LLM-environment interaction, called the agentic paradigm. The typical pattern of human-LLM interaction involves users providing input, engaging in a few rounds of interaction with the LLM, and consuming the results generated by the LLM. By contrast, an agentic LLM may interact with an external environment, through tools such as a web browser and Python interpreter, over dozens or even hundreds of turns. Although each individual tool call or function is short (often hundreds of tokens), the constant accumulation across turns and over long sessions lengths. As a result, short-lived patterns become highly I/O-bound. In multi-turn, short-lived patterns leads to very high KV-Cache hit rates – typically > 90% [2] – making the efficiency of KV-Cache loading, rather than per-computation, the dominant performance factor.

To improve throughput under agentic workloads, existing LLM inference systems have reimagined on a common set of architectural patterns: in-process pull2 [17, 14], pull2-avoidance/3FS integration [15, 10, 21] and external KV-Cache storage [2, 10, 21]. In these systems, pull2 engines load the KV-Cache in a burst via network to accommodate as many requests as possible within a single batch. When pull2 completes, decoding engines (typically receive KV-Cache from pull2 engines via a high-performance RDMA network). The decoding engines then generate tokens and store them in KV-Cache to distributed storage to enable reuse across turns. However, this architecture also introduces a critical bottleneck, as shown in [Figure 1](#), pull2 engines saturate fast lanes

26.02.26
DualPath:
Supercharges
Inference



Eight innovations. One keystone. The path forward: To V₄, R₂ and beyond.



revolutionized Inference AND Training!

▲ Non-Normal Hyper-Connection Stable Pyramid



U.S. AI: an inverted pyramid, propped up by financial engineering



Chinese AI: a stable pyramid grounded in architectural efficiency

MLA + mHC + Engram + OCR2 = nnHC



Foundation: **MLA**'s inference efficiency...



... is applied to make training efficient. Result? **mHC...**



... **Add:** **Engram**'s conditional memory



... **Add:** **OCR2**'s humanlike vision



... **Finally:** **DualPath**, tying everything tother



In the **Year of the Fire Horse**, a new type of AI brain is created

The nnHC Revolution:
MLA + MHC + Engram
+ OCR2 + DualPath

AI's Eureka Moment: DeepSeek redefines Intelligence



BRINGING IT ALTOGETHER: V4 and R2

The Age of Brawn is replaced by the Age of Brawn

🦷 The Age of Brawn (c. 2018–2025)

- 🌟 Brute force: *Scale über alles*
- 💻 Big chip = better LLM
- 🏗️ Training first; inference after
- 🔄 New model? Retrain 100%
- 🔒 Progress proprietary
- 💰 Users paid! DEARLY!
- 🎯 LLMs gamed lab benchmarks
- 🧗 Solo mountaineers to the summit

🧠 The Age of Brain (2026 onward)

- 🌀 Architectural elegance creates efficiency
- ⚖️ Chips less critical. Nvidia's hold broken
- 🚀 Inference supersedes training
- 📦 Easy database updates
- 👉 Open-weight consilience. $1 + 1 = 3$
- 🆓 Users often pay NOTHING!
- 🌐 Practical benchmarks rule
- 🧗 Mountaineering team climbs Mt. Intelligence

Brawn asks: "How big can I build? How many GPUs are needed?"
Brain asks: "How wisely can we climb – together? How well do we use our GPUs?"

You cannot train a hippo to dance like Nureyev



With profuse apologies
to Amilcare Ponchielli

Leaving the Racetrack. Going out onto the Open Road

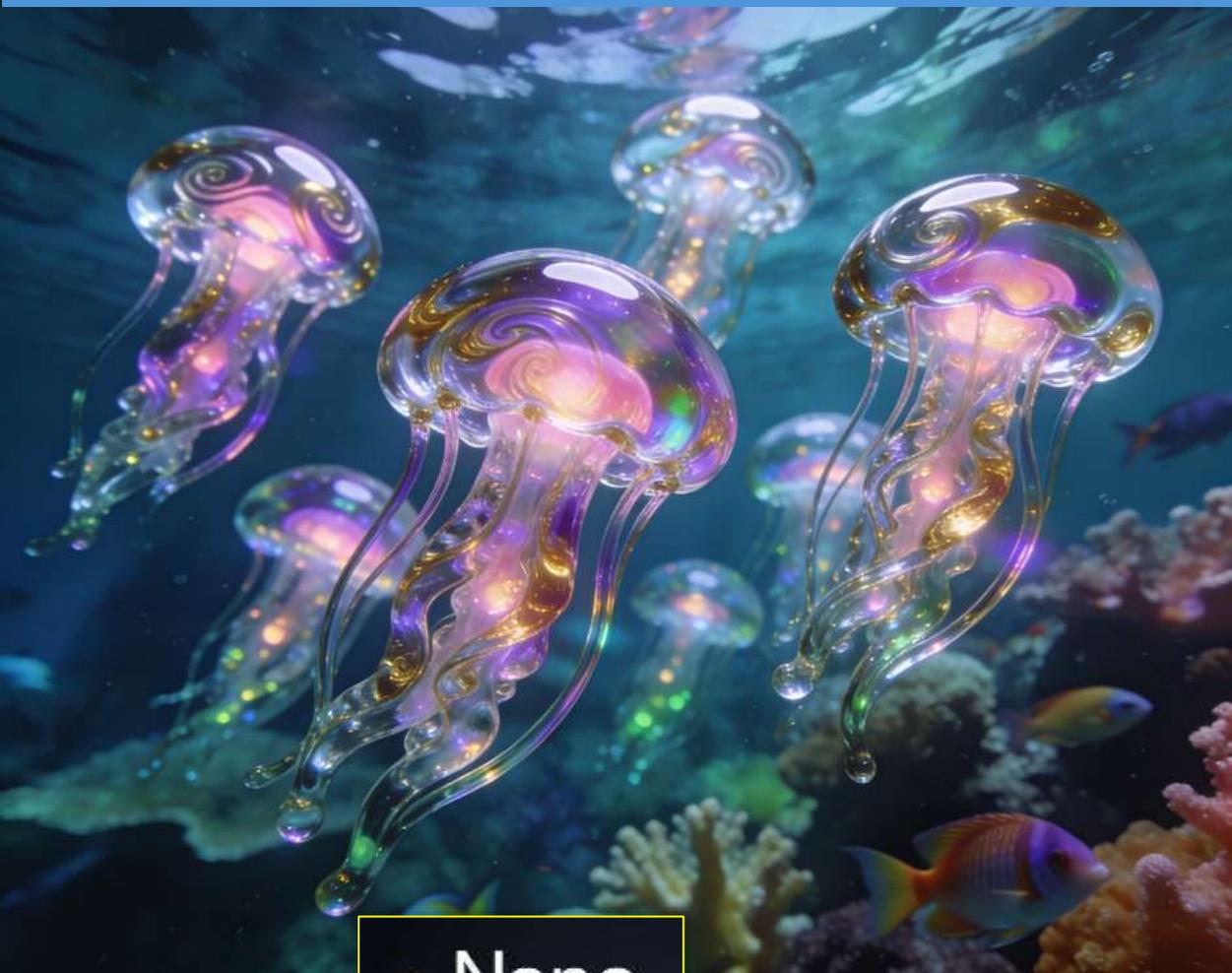


Going Global

A new species surfaces



Is ByteDance Hollywood's Nemesis?





 Seedance 2

Why an A.I. Video of Tom
Cruise Battling Brad Pitt
Spooked Hollywood 

The DragonSwarm: An Ensemble of Open Weight Specialists Computing, comprehending and orchestrating... TOGETHER!

 **DeepSeek:** *Architect Dragon*

 **Kimi:** *Memory Dragon*

 **Qwen:** *Universal Dragon*

 **ERNIE:** *Ubiquity Dragon*

 **MiniMax:** *Creative & Sonic Dragon*

 **Zhipu:** *Vision Dragon*

 **Intern:** *Science Dragon*

 **Ubiquant:** *Quant Dragon*

 **Hunyuan:** *Ecosystem & Social Dragon*

 **SenseTime:** *Urban & Industrial Dragon*

Leveraging Consilience – *where* $1 + 1 = 3$ – to achieve Collective Wisdom

Float like butterflies, sting like a...Swarm of Dragons

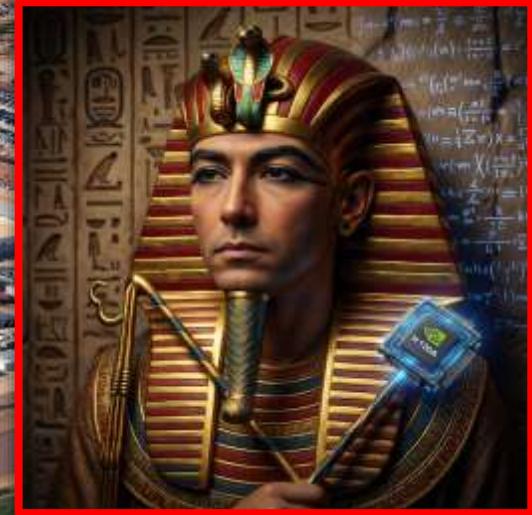


Castles in the Air...



...Inverted Pyramids on the Ground

Why build Inverted Pyramids on Texan Sand?



The Pharaohs' Stargates most likely did not work.
Will markets decide that today's AI pyramids cannot work either?

When Software becomes free, what holds up the Pyramid?



The cost of API tokens globally has fallen by 90% to 99% in recent years



 **David Shapiro (L/O)**  @DaveShapi · 22h  

You guys realize that all software is about to be free, right?

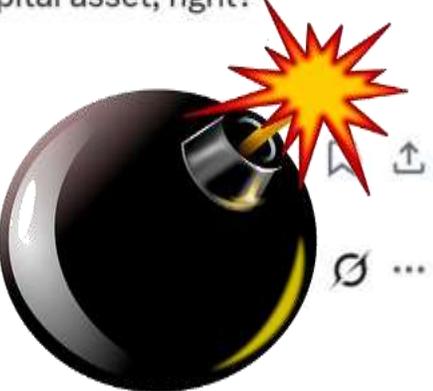
And software is presently the most valuable capital asset, right?

Guys?

 928  493  7.8K 

 **Elon Musk**   @elonmusk  

Pretty much



The Spending Chasm – Building Pyramids on Quicksand

Big Tech's 'breathtaking' \$660bn spending spree reignites AI bubble fears

FINANCIAL TIMES

Capex forecasts for data centre building (\$)

2026: 600bn-800bn

2027:
1100bn-1400bn

2028:
1000bn-2000bn

Cumulative Capex to 2030
\$3000bn-\$7000bn

Peak annual U.S. spending as % of GDP

TIME

1.3%

Tech (mostly AI)
2025*

1.2%

Broadband
cables
2000

.8%

Apollo
Project
1964

.4%

Interstate
buildout
1966

.4%

Manhattan
Project
1944

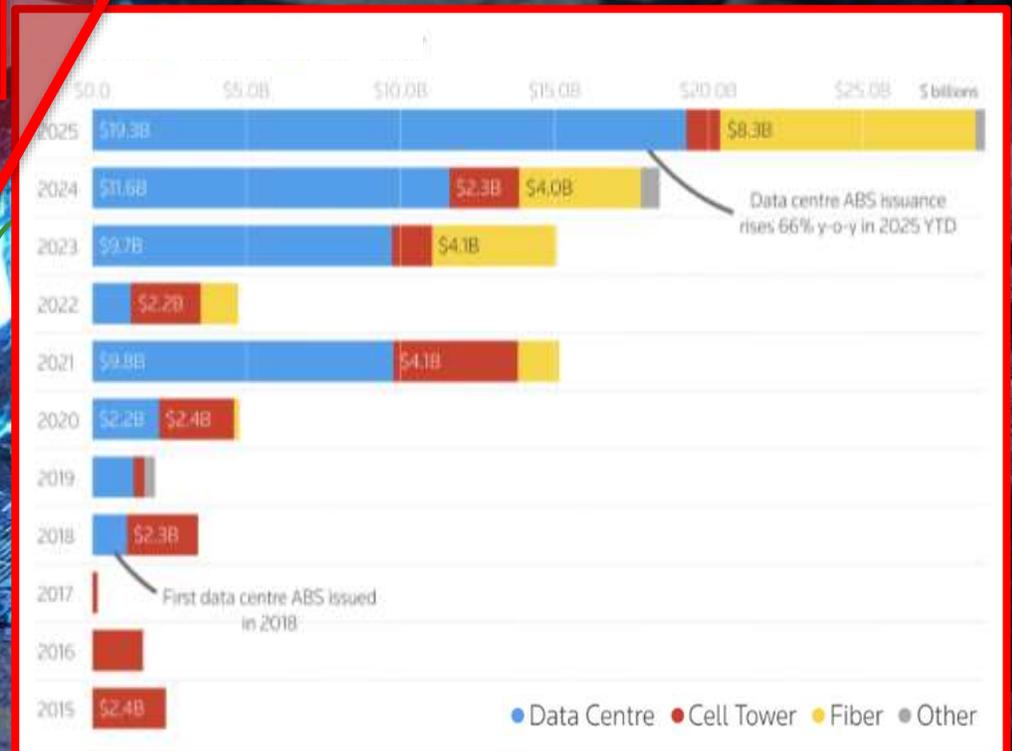
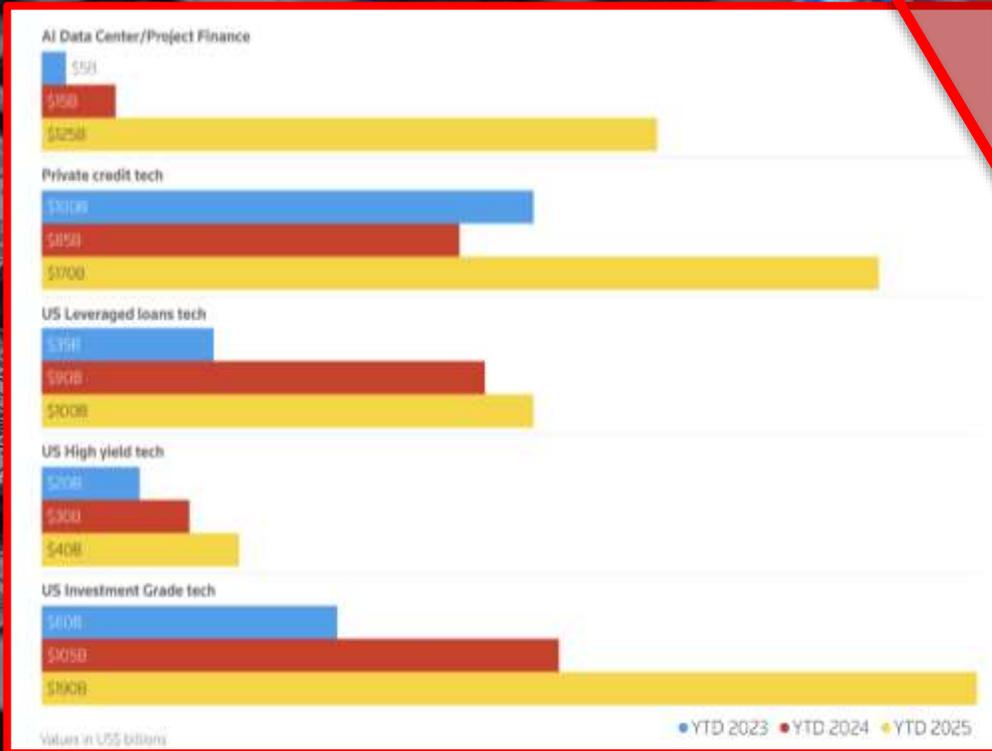
2026's Inverted Pyramid Scheme I: Debt

2024: \$92bn

2025:
\$182bn

f2026:
\$900bn

Total debt raised
for data centres



Corporate Bonds Fuelling AI Capex (UBS)

ABS debt underpins shift to data centres (BofA)

I owe, I owe... it's off to AI work we go...

2026's Inverted Pyramid Scheme II: Leasing

The Elephant in the Room: Oracle

The Lehman Brothers of AI?

Medium

Applied Digital est. 126%

Oracle 355%

CoreWeave
1263%

Gearing levels



2026's Inverted Pyramid Scheme II: Circular Financing

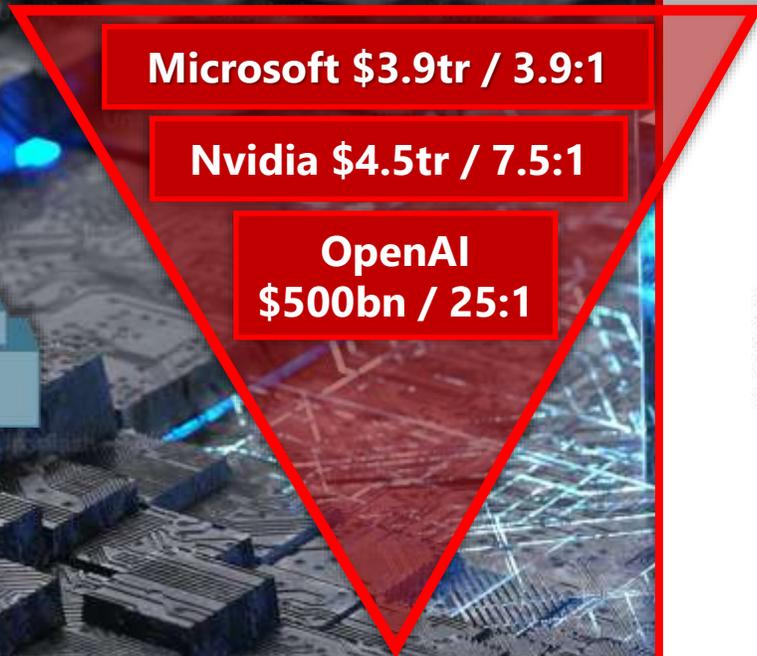


Global Crossing

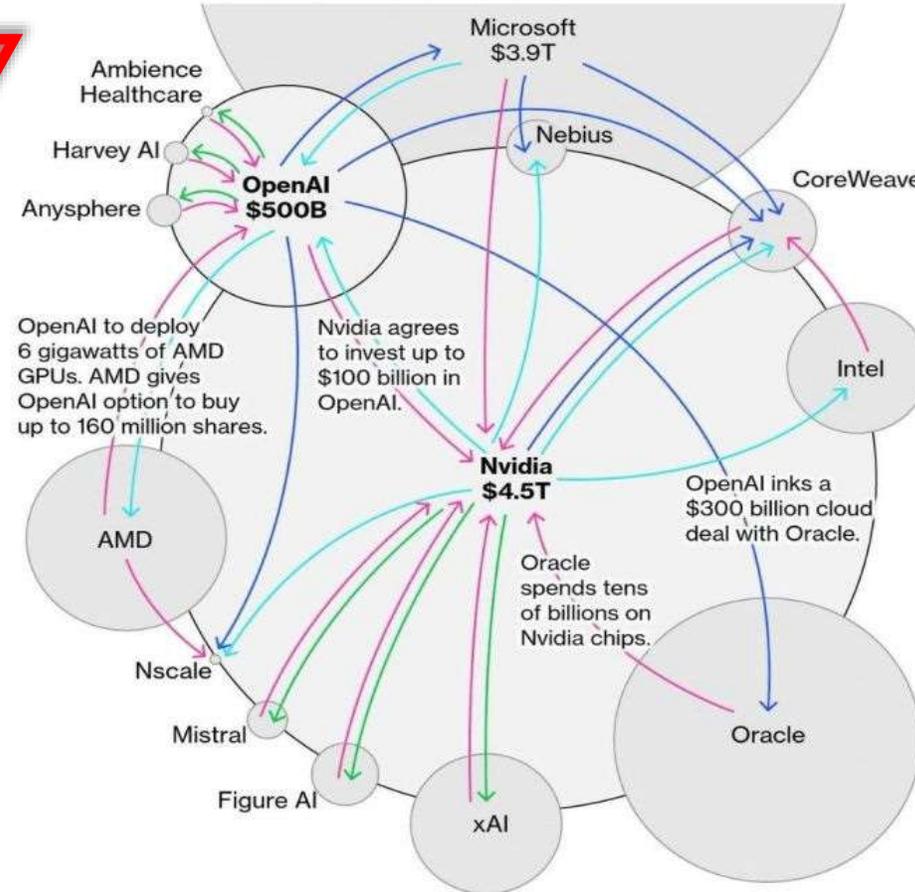
RIP



RIP



Hardware or Software Investment Services Venture Capital
Circles sized by market value



Source: Bloomberg News reporting

Bloomberg

Nvidia is the spider at the centre of this web



RIP



RIP

China's New Tao
Standing Up
Moving Forward



2025, leads in Open Weight. 2026, ANY Weights. 2030, even hardware?



Opinion | Rise of China's open source platforms will burst

AI bubble ■ South China Morning Post

Alibaba's Qwen family hits 700 million downloads to lead global open-source AI adoption ■

China leapfrogs US in global market for 'open' AI models
FINANCIAL TIMES

Are Chinese stock markets already discounting this transformation?

Forgotten lessons from American History I: Numbers count

The German Tank Commander's Lament

"One of our Tigers is worth four of their Shermans..."



...but the Americans always bring five."



"Quantity has a quality all its own." Joseph Stalin

The Admission: David Sacks, White House AI Lead, September 18, 2025



"China is not desperate for our chips... Huawei compensates for weaker individual chips by clustering more chips together."



China's GPU Insurgents: The "Four Little Dragons" Challenging NVIDIA's Dominance Medium



Nvidia's Tiger Stack:
Import-dependent, centralized, finite
– the peak of solitary power

NVIDIA NVL72
72 x Blackwell GPUs
Power: 5 ExaFLOPS

Huawei's Sherman Swarm:
Domestic, modular, limitless
– architecture out-scales old-style muscle



HUAWEI CloudMatrix 384
384 x Ascend 910C
Power: 6 ExaFLOPS

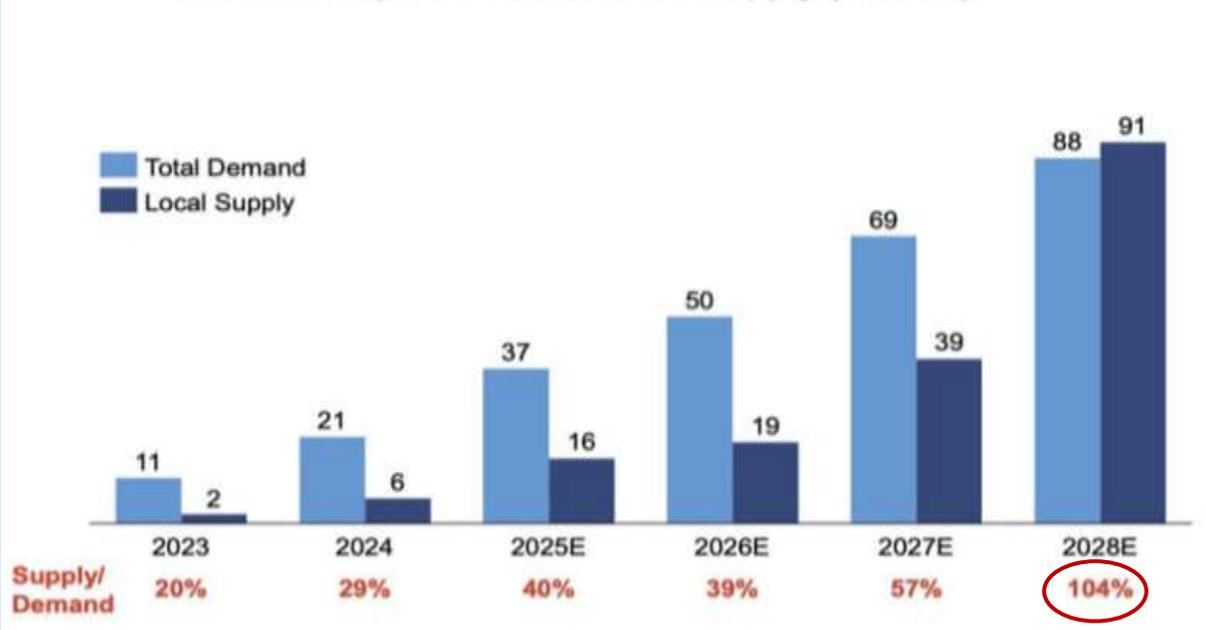
The future belongs not to the strongest single unit – but to the most adaptable system

The Bridge to Self-Sufficiency: The Gathering Swarm of Silicon

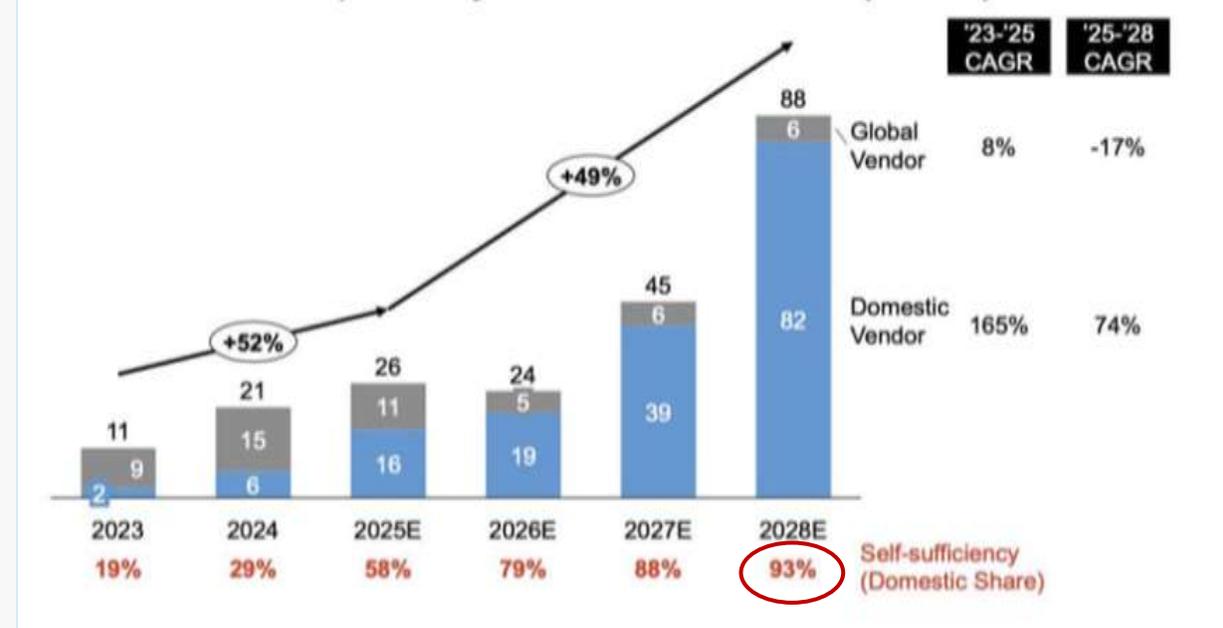
China's EUV prototype forces a rethink of the AI chip order

DIGITIMESasia

China AI Chip Local Demand vs. Supply (USD bn)



China AI Chip Sales by Global vs Local Vendors (USD bn)



 Bernstein forecasts imports shrinking to <7% by 2028 – with “no-breakthroughs”

⚡ Shenzhen's EUV prototype means 5nm+ production will start years ahead of schedule

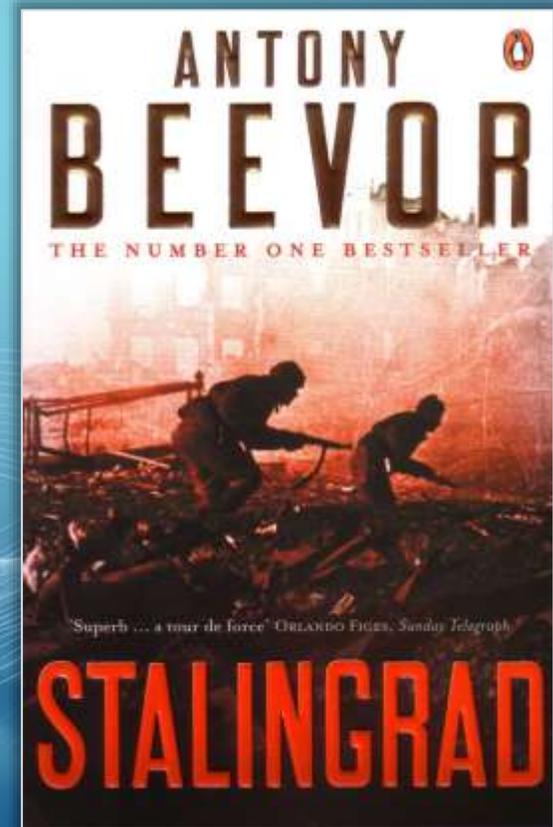
China's chip strategy is to leapfrog the timeline that lies ahead

Forgotten lessons from American History II: Manufacturing counts

Lex Fridman talking to James Holland as to why the US-led Allies won WW2

Fridman: *You could think of WW2 as a battle of factories?*

Holland: *Yes.*



German historians ascribe the Axis WW2 loss to the Allies – and Beevor Germany's loss to Russia at Stalingrad – to *Materialschlacht*: the Battle of Matériel.

Share of Global Manufacturing f2030: US 11%; China 45% (UNIDO)

The Two Machines that define US AI are '*Not Made in The USA*'...

ASML EUV Lithography | Veldhoven, **Netherlands**



U.S. AI
Critical
Path

TSMC Advanced Fabrication | Hsinchu, **Taiwan**



'They stole our chip business': Trump lashes out at Taiwan's chip sector after Supreme Court hands him a humiliating tariff defeat

When your critical path goes via Veldhoven and Hsinchu, who manufactures your destiny?

...whilst China manufactures its OWN future

Energy Self-sufficiency: Solar and Wind

Manufacturing the grid that powers the AI network



Automation

Manufacturing the labour that makes everything else

Electricity generated by Renewables

Manufacturing the photons that train the models



Mobility

Manufacturing EVs – and exporting them

"The US is far behind – it's humiliating where we've ended up."

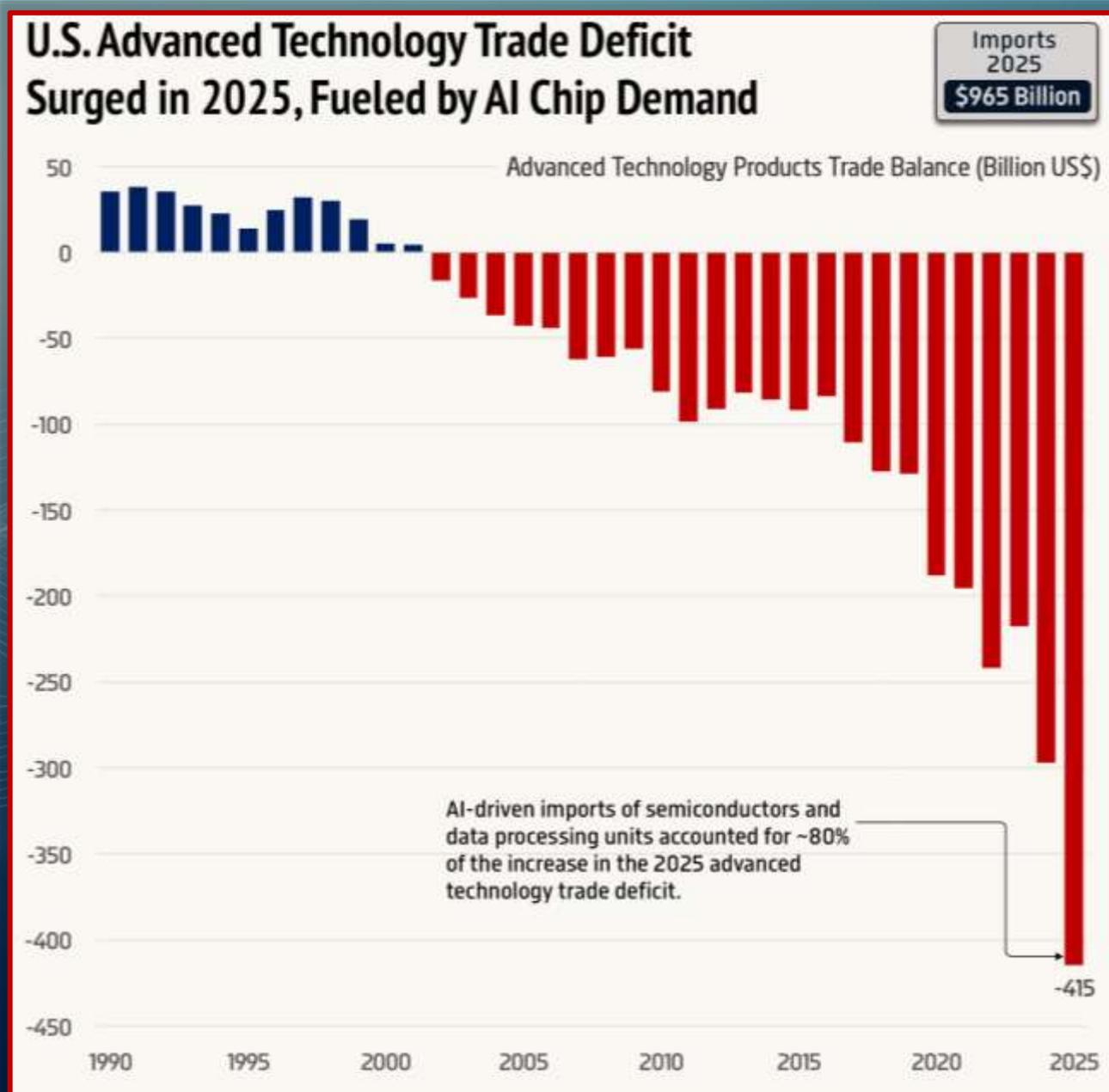
Jim Farley, CEO of Ford, after visiting China



Ford held talks with China's Xiaomi over EV partnership FINANCIAL TIMES

When you remotely control factories via AI – and those factories manufacture what defines tomorrow's world – do you control the future?

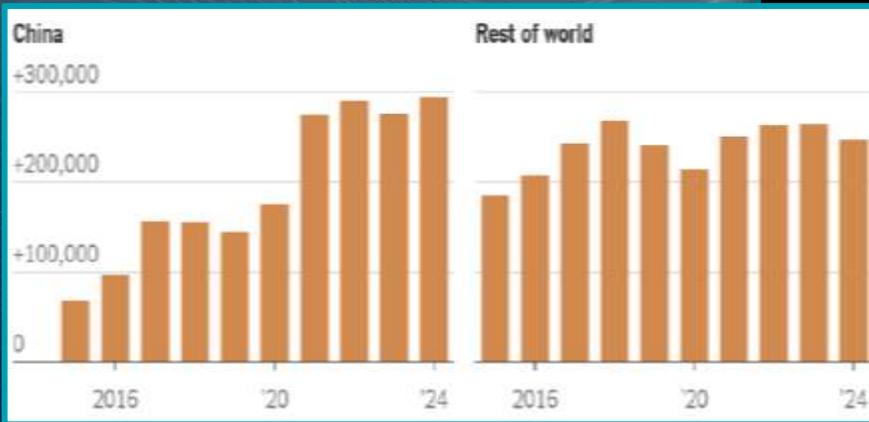
Making things counts



The Proof: Xiaomi's Dark Factory powered by SenseTime Dragon

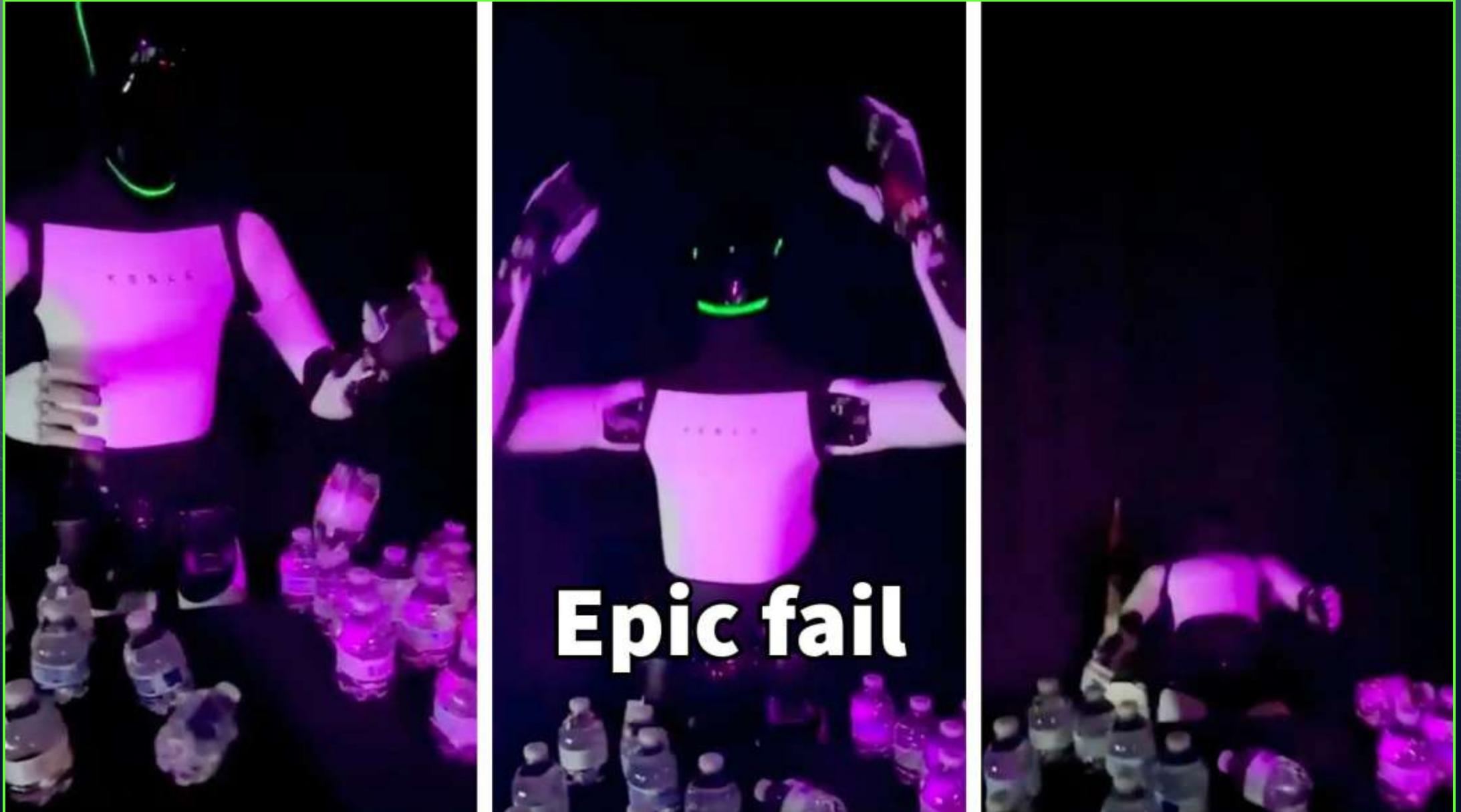
-  **Location:** *Beijing, China*
-  **Human Staff on Floor:** *0*
-  **Products:** *Smartphones, IoT devices, robotics components*
-  **Annual Output:** *12+ million units*
-  **Uptime:** *165/168 hours per week*

Annual Installations of Industrial Robots



This is not just a factory. It is a blueprint for sovereign, scalable, AI-driven production.

And Musk wants to switch Tesla from making EVs to Robots?!



Less than two months later, at the Spring Festival Concert...



Shi po tian jing

Stone shattering the sky

U.S. AI Bubble



 DeepSeek

OpenAI

Google

Google

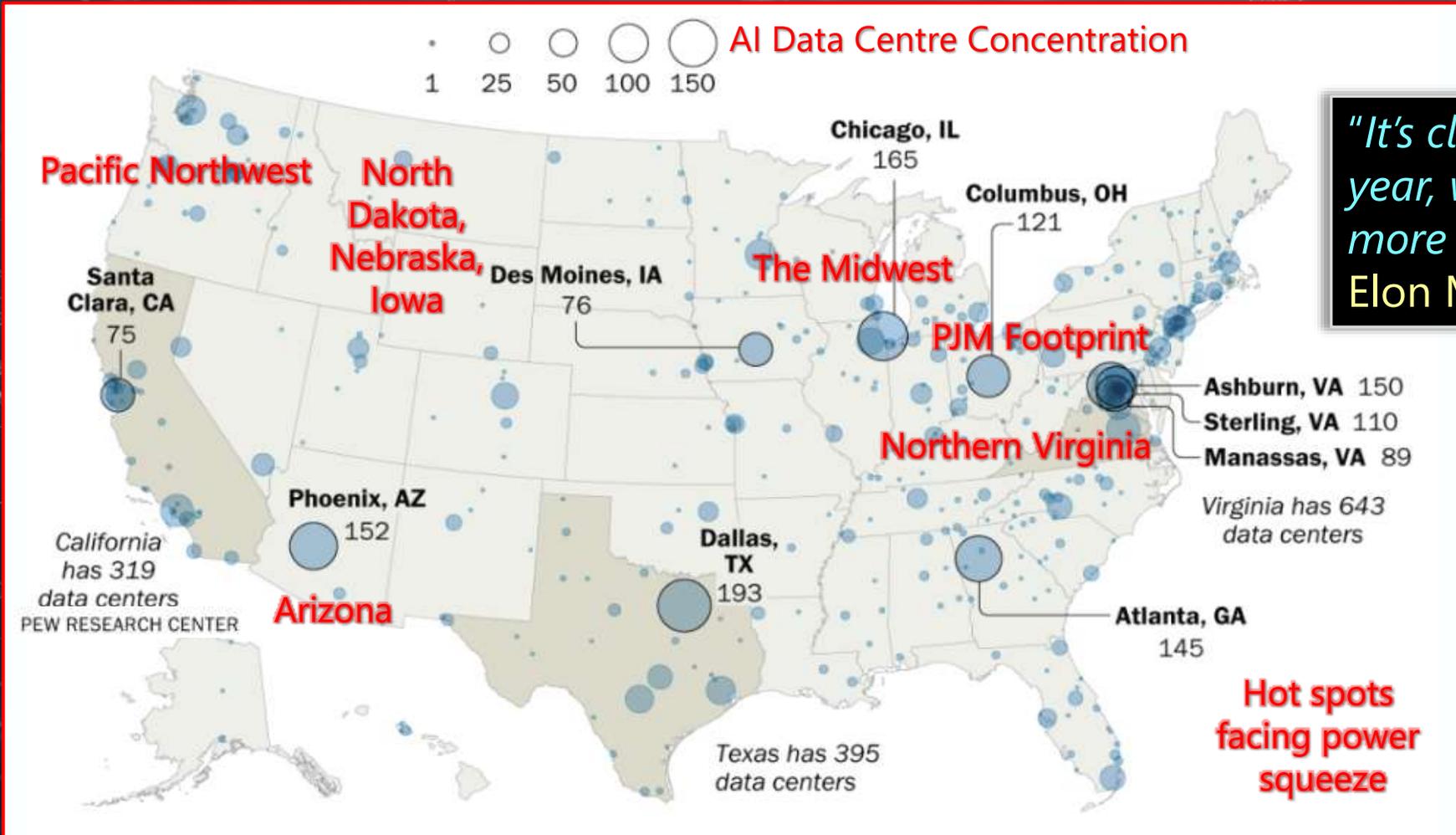
Meta

 Meta

/AI

AI's Energy Wall, when compute is held hostage by the Grid

"AI's natural limit is electricity, not chips." Eric Schmidt, former Google CEO



"It's clear that, maybe later this year, we (the US) will be producing more chips than we can turn on."
Elon Musk at Davos

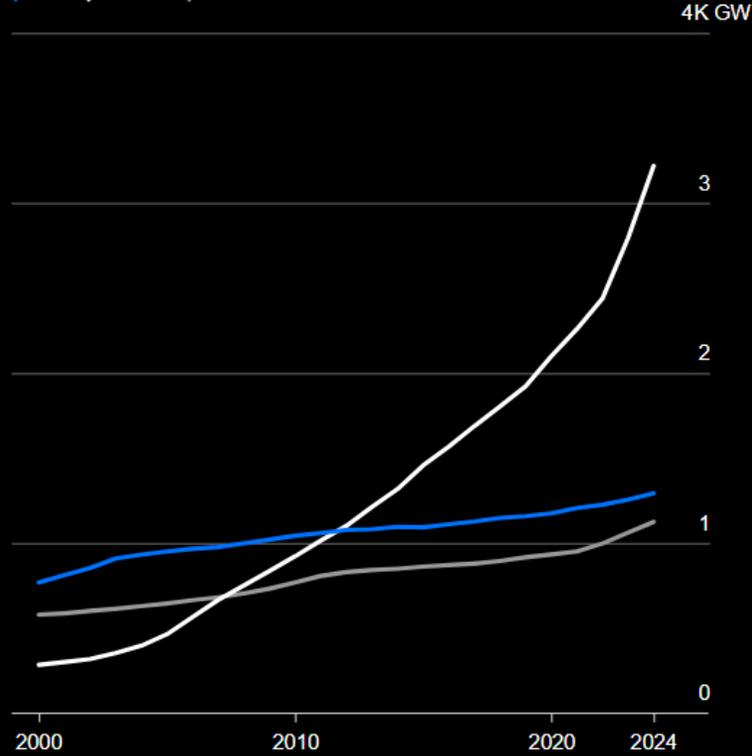
You cannot scale intelligence without energy. And the US is running out of power

The Contrast: As China powers ahead, the US crawls

China Has Twice as Much Power as the US

Installed electric generation capacity has skyrocketed since the 2000s

US China EU

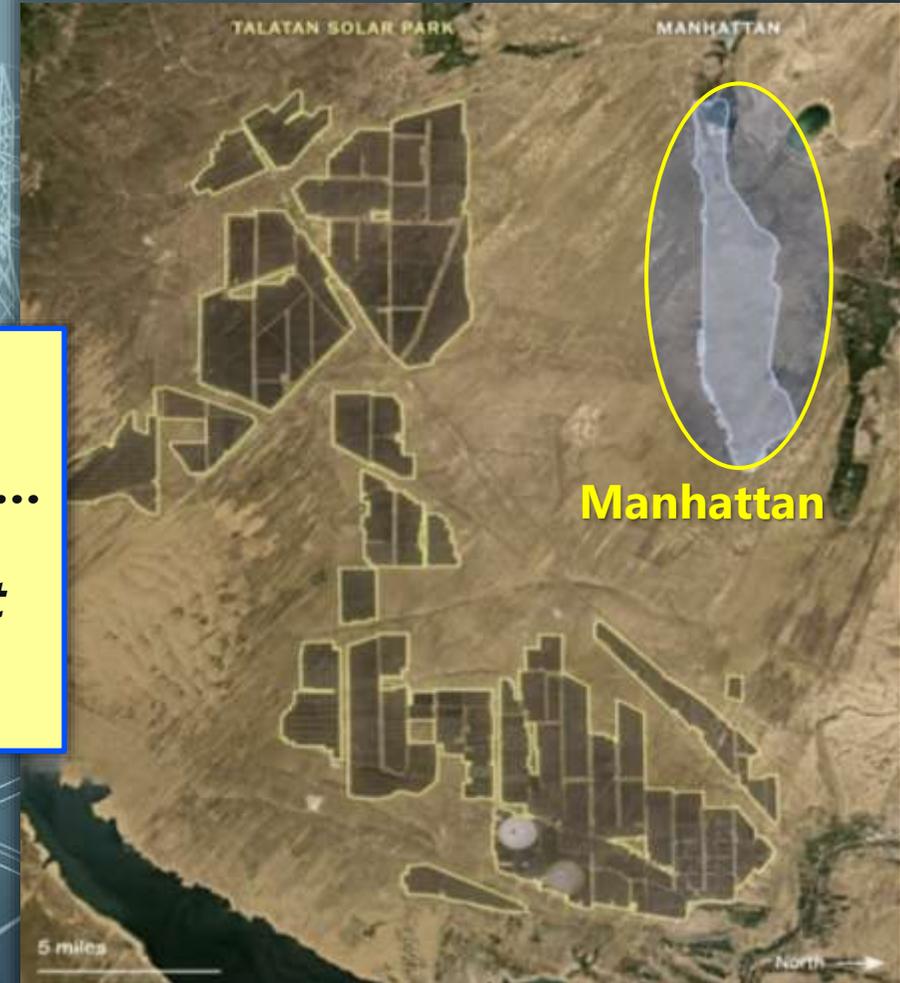


Source: International Energy Agency

Bloomberg Opinion

"China's going to have more power than anyone else and probably will have more chips... Based on current trends, China will far exceed the rest of the world in AI compute."
Elon Musk at Davos

In 2025, China added 470GW of power, 7 times the US's 64GW



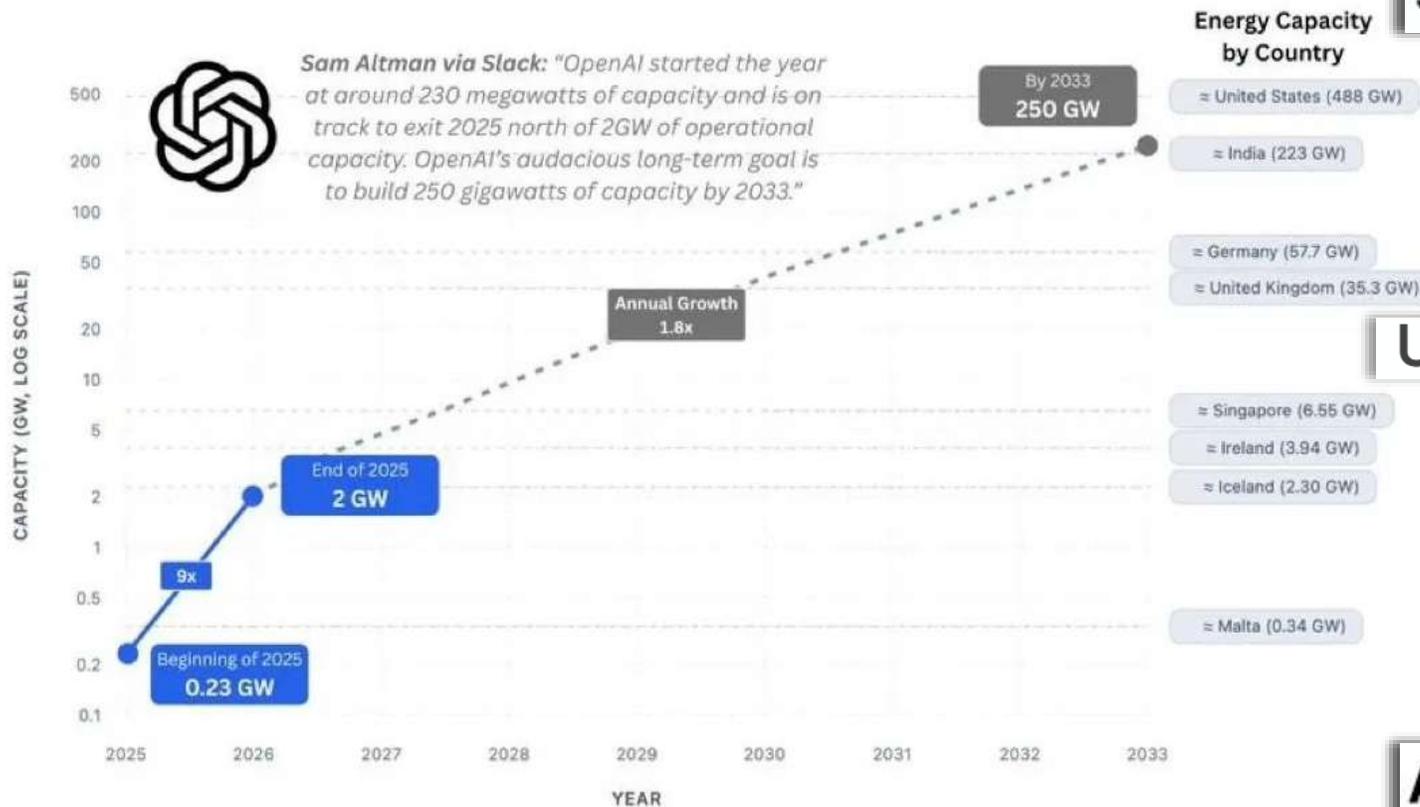
Talatan Solar Park: 4x size of Manhattan; powering 1.5m homes; built in 3 years

China is building the grid. America is searching for the plug

US power supply is already "Grid-locked"...and getting worse...

OpenAI planning to 125x energy capacity in 8 years

This would mean using more than India's energy capacity today



Source: Alex Heath, Sources.News

Peter Gostev (<https://x.com/petergostev>); (<https://www.linkedin.com/in/peter-gostev/>)

AI Chip Inventory Problem: Power Shortages Leave Hardware Idle

US AI boom faces electric shock

Amazon Says Berkshire Utility Failing to Power Data Centers

From Moore's Law to the Jevons Paradox:
"The more efficient we get, the more energy we use."

Where elegance shapes immensity – even in space.

Where structure outlives extravagance – even in cyberspace.

Where intelligence sings – even in code.

Where simple genius, not brute force, balances the most profound equations:

$$I = \infty \times \frac{S}{C} \quad \nabla \cdot \mathbf{E} = \frac{\rho}{\epsilon_0} \quad \sum_n (x_i \cdot w_i) + b$$

$$E=mc^2$$

"Pure mathematics is, in its way, the poetry of logical ideas." Albert Einstein

"Whom the gods would destroy, they first put on the cover of TIME"



TODAY'S POET LAUREATES?

1. Jensen Huang (Nvidia)
2. Sam Altman (OpenAI)
3. Mark Zuckerberg (Meta)
4. Elon Musk (xAI)
5. Lisa Su (AMD)
6. Demis Hassabis (Google)
7. Dario Amodei (Anthropic)
8. Fei-Fei Li (Stanford)



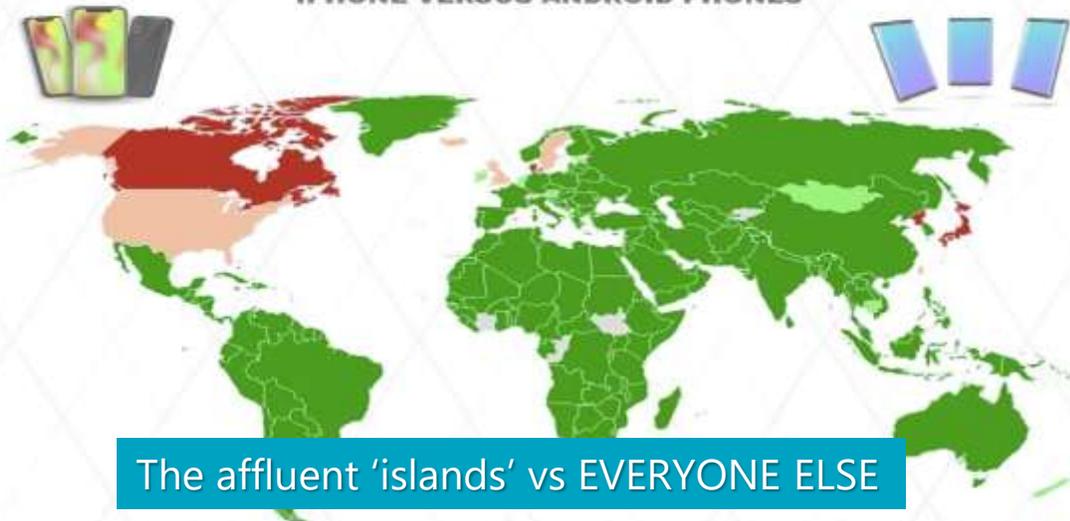
Will Chinese LLMs Become the World's Open Weight Lingua Franca?

Microsoft warns that China is winning AI race outside the west



ANDROID OR APPLE?

WHICH IS MORE POPULAR, APPLE OR ANDROID?
IPHONE VERSUS ANDROID PHONES



The affluent 'islands' vs EVERYONE ELSE

● APPLE USERS (60-100%) ● ANDROID USERS (60-100%)
● APPLE USERS (50-60%) ● ANDROID USERS (50-60%)

LLM Leaderboard

Compare the most popular models on OpenRouter ⓘ

1.	MiniMax M2.5 by minimax	2.09T tokens ↓15%
2.	Gemini 3 Flash Preview by google	912B tokens ↑14%
3.	Kimi K2.5 by moonshotai	878B tokens ↓27%
4.	GLM 5 by z-ai	789B tokens ↓10%
5.	DeepSeek V3.2 by deepseek	774B tokens ↑4%

80% of US AI startups rely on Chinese open-source models for survival. Investors from Andreessen Horowitz are shocked. The top 16 on the global open-source list are all occupied by Chinese entries.

36Kr

Open source won mobile. Open Weight will win AI.
And China now owns Open Weight AI.

What's in DeepSeek V4...

THE SPECS

Scale & Architecture:

- 1.4 trillion parameters
- Sparse MoE
- Engram-enabled
- 1 M+ token context

Openness & Efficiency:

- Open Weight
- 95–99% cheaper training cost:
- 98% cheaper token inference cost

R1 was a quake...



...and what will it mean?

THE IMPACT

Market Shockwaves:

- Chips commoditize
- Models Radically Cheaper
- Ecosystem DragonSwarm ascends
- Closed weight models fall

Structural Shifts:

- Market bubble pressures intensify
- Chinese self-sufficiency by 2028
- AI Centre of Gravity moves to China

...V4, R2 are tsunamis!

In short, DeepSeek is a porcupine at a balloon party!



OpenAI expects another 'seismic shock' from China amid speculation of new DeepSeek release  South China Morning Post



27.01.26 & 16.02.26



The first tremors? Warnings.
The rupture? Now.
The aftershocks? You're living in them.

Thank you

“Panics do not destroy capital; they merely reveal the extent to which it has been previously destroyed by its betrayal into hopelessly unproductive works.”

JOHN STUART MILL