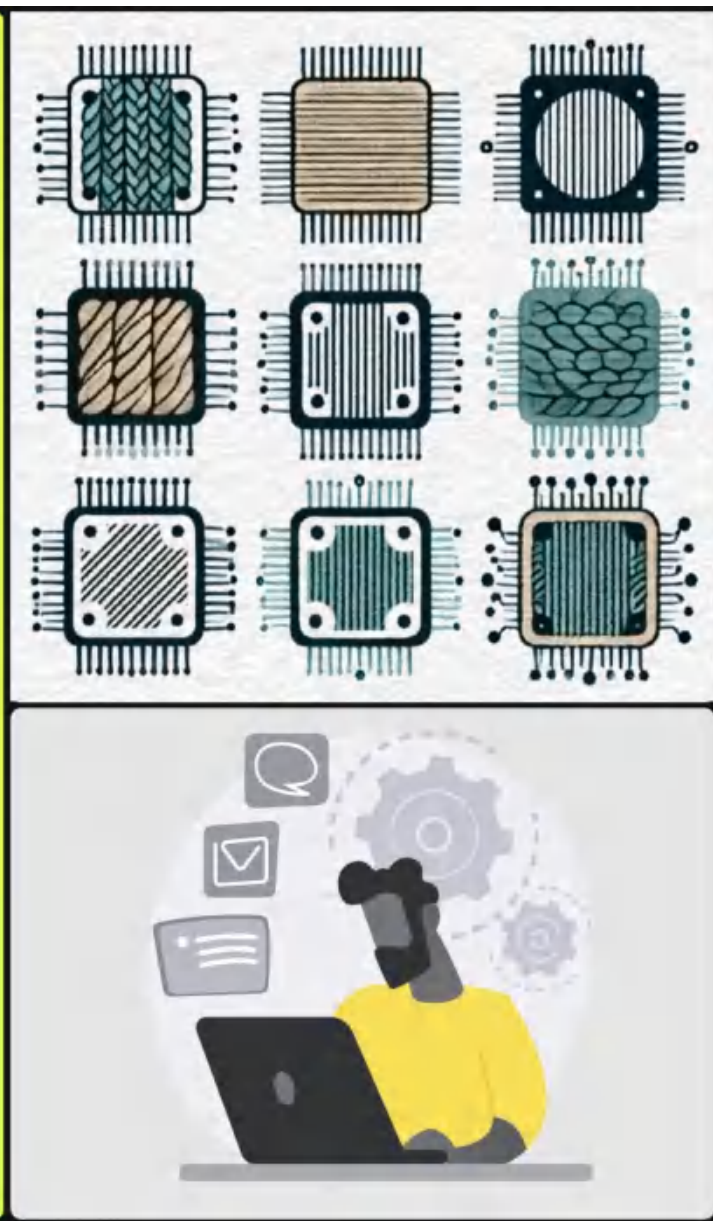




从DeepSeek爆火 看25年AI行业发展

2025-02





人工智能产业链联盟

星主： AI产业链盟主

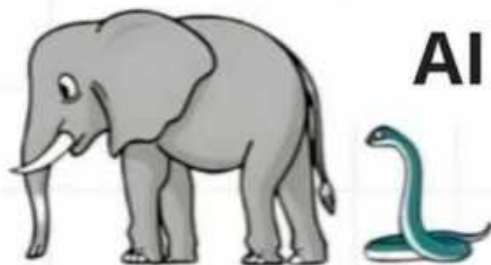
知识星球

微信扫描预览星球详情



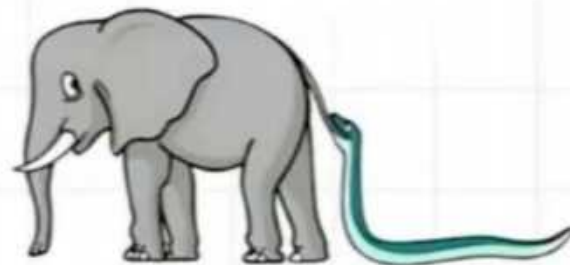
当下AI行业现状

Company

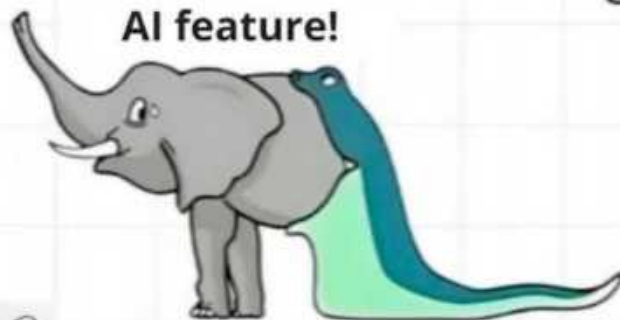


AI

Nobody asked
for this... but



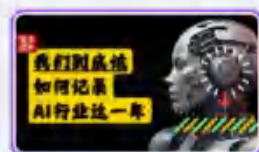
Let's build an
AI feature!



And say we are an
AI company now!



得亏我的年度AI报告是12月做的



01



02



03



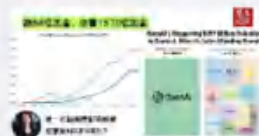
04



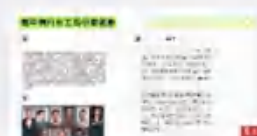
05



06



08



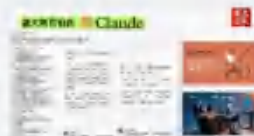
09



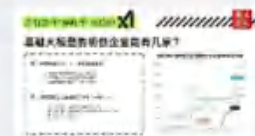
10



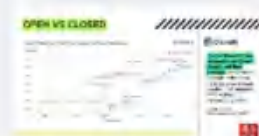
11



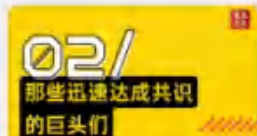
12



13



15



16



17



18



19



20



22



23



24



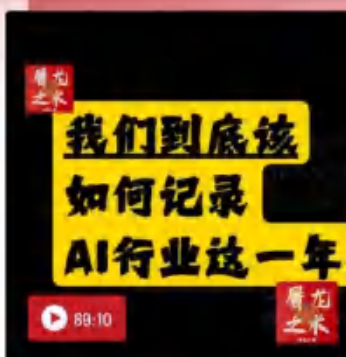
25



26



27



Vol.45 132页PPT记录AI行业这一年

屠龙之术

按照每年的习惯我都会针对当年的行业发展做一次整理。

小宇宙



如果今天只有一页PPT



我们走到哪了？

L2-L3之间



DeepSeek爆火

- DeepSeek为什么火了？
- 火的到底是什么？
- 一些核心的关键决策/节点/细节
- 开源、低成本的意义
- 那些谣言与误解



- more money=better infra=better training=better AI
- Capex-英伟达-数据中心-公有云
- 特朗普-星际之门计划
- 芯片与能源
- 战略投资与并购

过去2年的叙事



新叙事的可能

- 基于现有数据的scaling law 失效
- 预训练边际效应急速降低
- 后训练与强化学习崛起
- 推理类模型显现的能力提升
- OpenAI o1/Claude 3.5 Sonnet/DeepSeek R1

- L3可能等于推理模型+场景实现
- 肉眼可见的aha moment=搜索、代码
- 没有共识的问题：模型连续or非连续
场景边界
组织结构的适配

万物皆可Agent





关于DeepSeek的种种



这些问题你是否想过？



差不多时间豆包、阶跃、Kimi、Minimax似乎都发了新模型，为什么他们没火？

如果你说因为开源，那么阿里的Qwen在春节期间也发了模型，榜单成绩也很好为什么没火？

DeepSeek V3是12月底发布，R1是1月20日发布，那么是V3“重要”还是R1？

DeepSeek 的官方APP上线时间是1月10日，为什么起初没火？后来又是因为什么火的？

开源 VS 闭源

巨龙 VS 屠龙少年

低成本与推理模型

产品体验与窗口期

回头看这几周 有必然，当然也有偶然

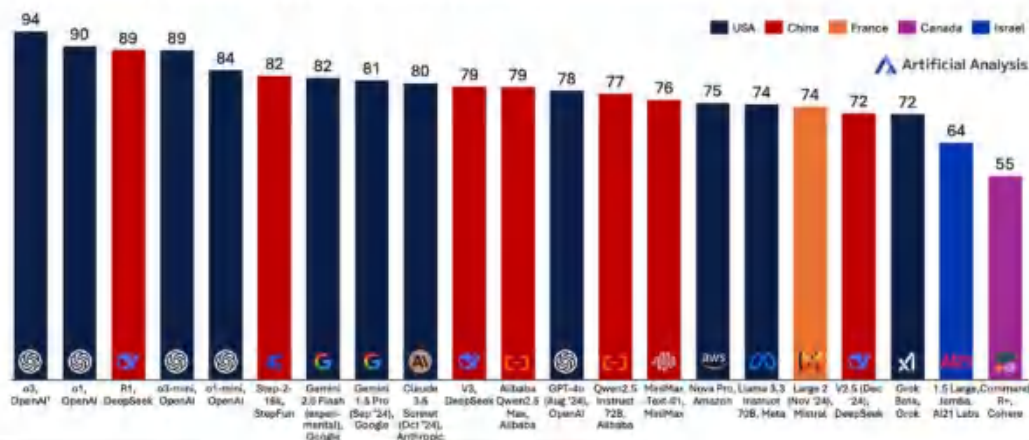
- 戏称大模型界的“拼多多”
- 科技巨头年底财报+25年预期投入暴增
- 巨大Capex投资的顶峰---星际之门
- 美国主流媒体和创投圈表态
- 媒体信息的“出口转内销”
- “春节档”与“国运”
- Fake文章的“真假难辨”
- 产品功能上的临界点与“定价”
-



The chart displays the Artificial Analysis Intelligence Index for various AI models. The USA models (blue line) show a steady increase in performance, while the Chinese models (red line) show a more rapid increase, particularly in the latter half of the timeline.

Model	Country	Approximate Index	Approximate Date
OpenAI, GPT-3.5 Turbo	USA	53	Q4 2022
OpenAI, GPT-4	USA	73	Q1 2023
OpenAI, GPT-4 Turbo	USA	75	Q3 2023
Alibaba, Qwen Chat 7B	China	37	Q3 2023
Alibaba, Qwen Chat 72B	China	56	Q4 2023
DeepSeek, V2	China	66	Q1 2024
Alibaba, Qwen 2 Instruct 72B	China	68	Q2 2024
GPT-4o	USA	76	Q2 2024
Anthropic, Claude Sonnet (Jun '24)	USA	76	Q2 2024
OpenAI, GPT-4o mini	USA	87	Q3 2024
Alibaba, Qwen 2.5 Instruct 72B	China	77	Q3 2024
DeepSeek, V3	China	79	Q4 2024
OpenAI, GPT-4o	USA	90	Q4 2024
DeepSeek, R1	China	88	Q1 2025
OpenAI, GPT-4o	USA	92	Q1 2025

Artificial Analysis Intelligence Index, Selected Leading Models (Early 2025). Non-exhaustive



**模型层够强，
更重要的是
中国、开源、便宜**



DeepSeek-V3与其他模型训练成本对比

发布时间	企业名称	模型名称	训练成本（按照GPU租赁逻辑计算）
2023年3月	OpenAI	GPT-4	约6300万美元
2024年3月	/	GPT-4相当规模	约4800万美元
2024年6月	Anthropic	Claude 3.5 Sonnet	3000~5000万美元
2024年7月	Meta	Llama 3.1 405B	约5800万美元（仅预训练）
2024年12月	深度求索	DeepSeek-V3	557.6万美元

使用成本的指数级下降



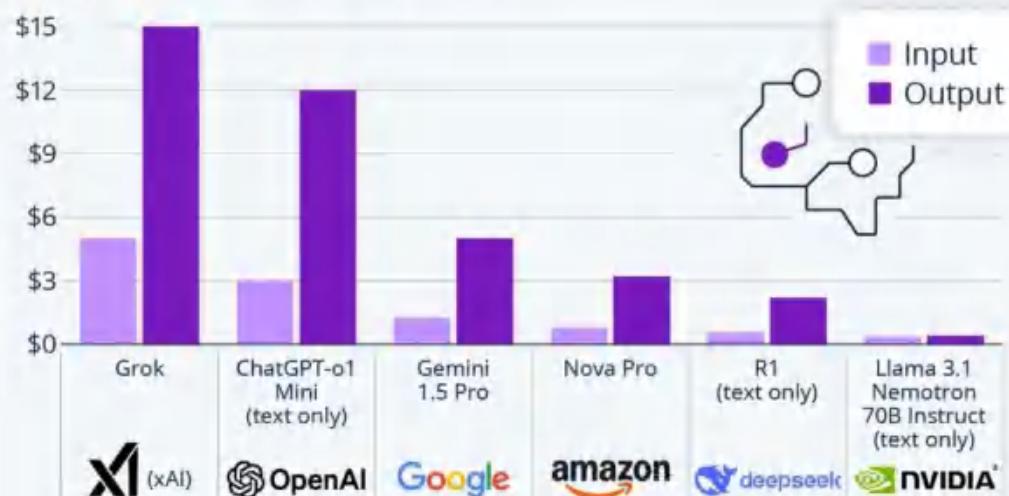
- The cost to use a given level of AI falls about **10x every 12 months**, and lower prices lead to much more use.
- You can see this in the token cost from GPT-4 in early 2023 to GPT-4o in mid-2024, where the price per token dropped about **150x** in that time period.

Plot of model pricing vs LMSys Elo (Jan 2025)



DeepSeek-R1 Upsets AI Market With Low Prices

Estimated price for processing one million input/output tokens on different AI models



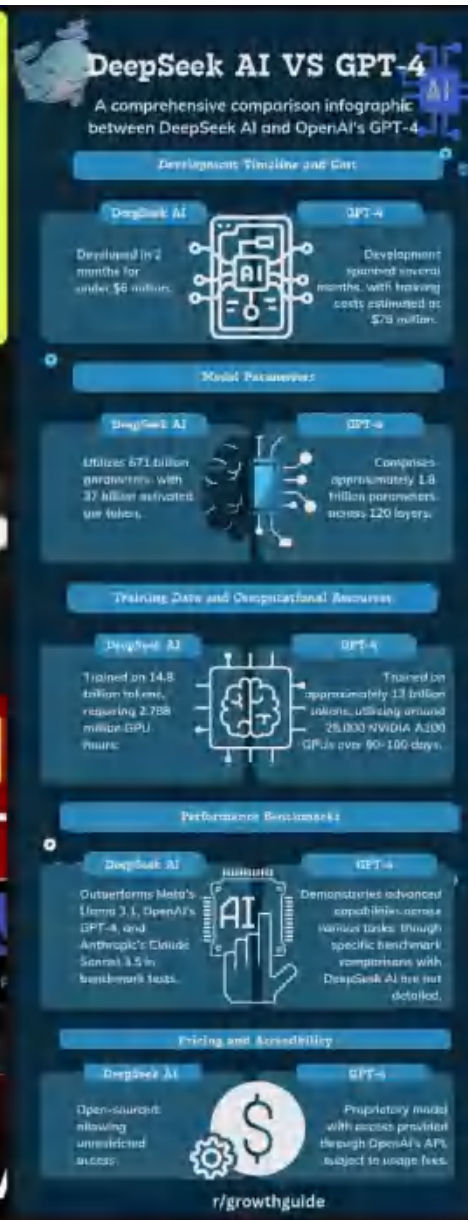


“出口转内销”

THE CHINA TAKE ON DEEPSEEK
STATE MEDIA HAILS DEEPSEEK AS CHINA'S GEM

BREAKING NEWS
DEEPSEEK AI: HIT BY LARGE-MALICIOUS ATTACK

DEEPSEEK AI SENSATION
DEEPSEEK FACES CYBER



SAYS DEEPSEEK AI WOULD BE A WAKE UP CALL

BREAKING MARKET NEWS | SPONSORED BY *charles SCHWAB*

CHINA'S DEEPSEEK A.I. MODEL SPARKS TECH SELL-OFF

FOX BUSINESS ALERT

RADES	OIL	GOLD	DOV
72.62 ▼ 2.04 -2.7%	2,764.80 ▼ 41.80 -1.4%	44,436	11.1
AI CORP (NVDA) 120.74 ▼ 21.88	EATON CORP (ETN) 3	+0.03	

“国运”

- **强大**---推理模型
- **便宜**---模型成本
- **开源**---模型部署
- **免费**---转到APP
- **联网**---产品体验
- **本土**---国运级别

时间点：
2025-01-26 周日

可能是24年中国
最有名的人之一



 微博轻享版

Yocar-冯骥

游戏科学创始人、CEO,《黑神话：悟空…

关注 443 粉丝 64.1万



2025-01-26 23:32

年前还是认真推下DeepSeek吧。
V3已经用了一个月，但R1仅仅用了5天，不得
不来发这条微博。
先讲一句暴论：DeepSeek，可能是个国运级
别的科技成果。



黄仁勋DeepSeek公开信：诸位同仁，请将今日的波动视为一份礼物，我们不是在制造芯片，而是在铸造时间的钥匙！

科学空天人 2025年01月28日 14:14 获赞

12人



亲爱的同仁与伙伴：

此刻，当我望向窗外硅谷的夜色，星辰与服务器集群

科学空天人 +关注 13 81 2

如何看待冯骥盛赞 DeepSeek「堪称国运级别的科技成果，希望它能让 AI 成为你生活中的水与电」？

梁文锋

各位知乎的年轻朋友好，我是深度求索（DeepSeek）的梁文锋，刚回答了一个问题，又看到冯骥盛赞 DeepSeek，特来回复一些。

各位同仁，当您读到这篇博文时，您正在经历一个激动人心的时刻。当您看到这篇博文时，您正在经历一个激动人心的时刻。当您看到这篇博文时，您正在经历一个激动人心的时刻。

必须坦白，当您读到这篇博文时，您正在经历一个激动人心的时刻。当您看到这篇博文时，您正在经历一个激动人心的时刻。当您看到这篇博文时，您正在经历一个激动人心的时刻。



rosicky311_明浩

2025年01月28日 14:14 获赞

AI深度求索

01/28

谁曾想2024龙年的最后一天，是两篇AI写的fake文（上午nvidia老黄的内部信+晚上deepseek梁文锋的知乎回答）刷屏了所有人的社交媒体...



AI深度求索



即刻



rosicky311_明浩

2025年01月28日 14:14 获赞

AI深度求索

01/28

大年初一一睁眼，又叫醒了几个被ai骗了的人类；但又想如果再发展几年，谁是“真”，谁是“假”？如果绝大多数人都相信了那个“假”的东西...那TA还是假的吗？

AI深度求索



即刻

High起来吧.....

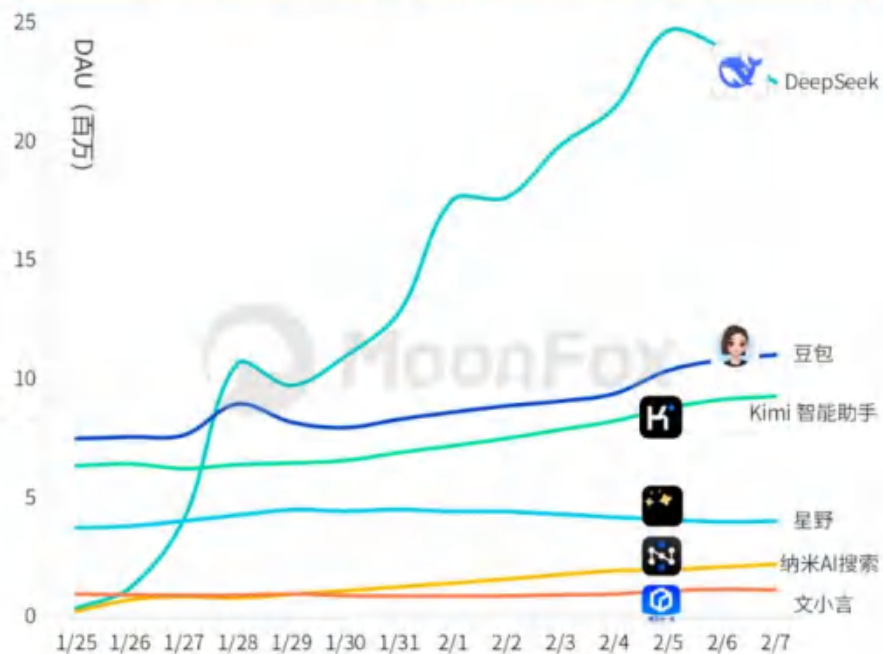


泼天的富贵

猜猜现在的数字



生成式AI应用日活跃用户规模变化 (1月25日-2月7日)



Ai 产品榜 · 全球日活总榜



榜单数据说明: 数据范围 2025/1/11-2025/1/31, 本榜单数据基于网络公开可获取数据, 如有错误请联系数据团队删除。由于统计的局限性, 数据可能存在误差但不影响于排名。数据不全等问题, 本榜单数据仅供参考, 不代表任何上榜单或上榜单产品的评价。备注: 不代表上榜单产品符合相关的政策法规, 亦不代表任何投资、求职、建议、请谨慎使用。非商用引用数据标注来源: [公众号@AI产品榜 aicpb.com]



如果你是梁文锋……

大模型“安全”问题？

看起来做的工作确实不够
无论是模型本身的对齐
还是APP/网页端的数据
抑或是针对各国政策的适配
那么之后呢？

要针对C端做运营么？

APP/Web真的重要吗？
或者说个人用户的使用重要吗？
如果不接，未来是什么产品形态？
如果往前迈几步，迈去哪？收费？
细分场景？

对云厂商的态度？

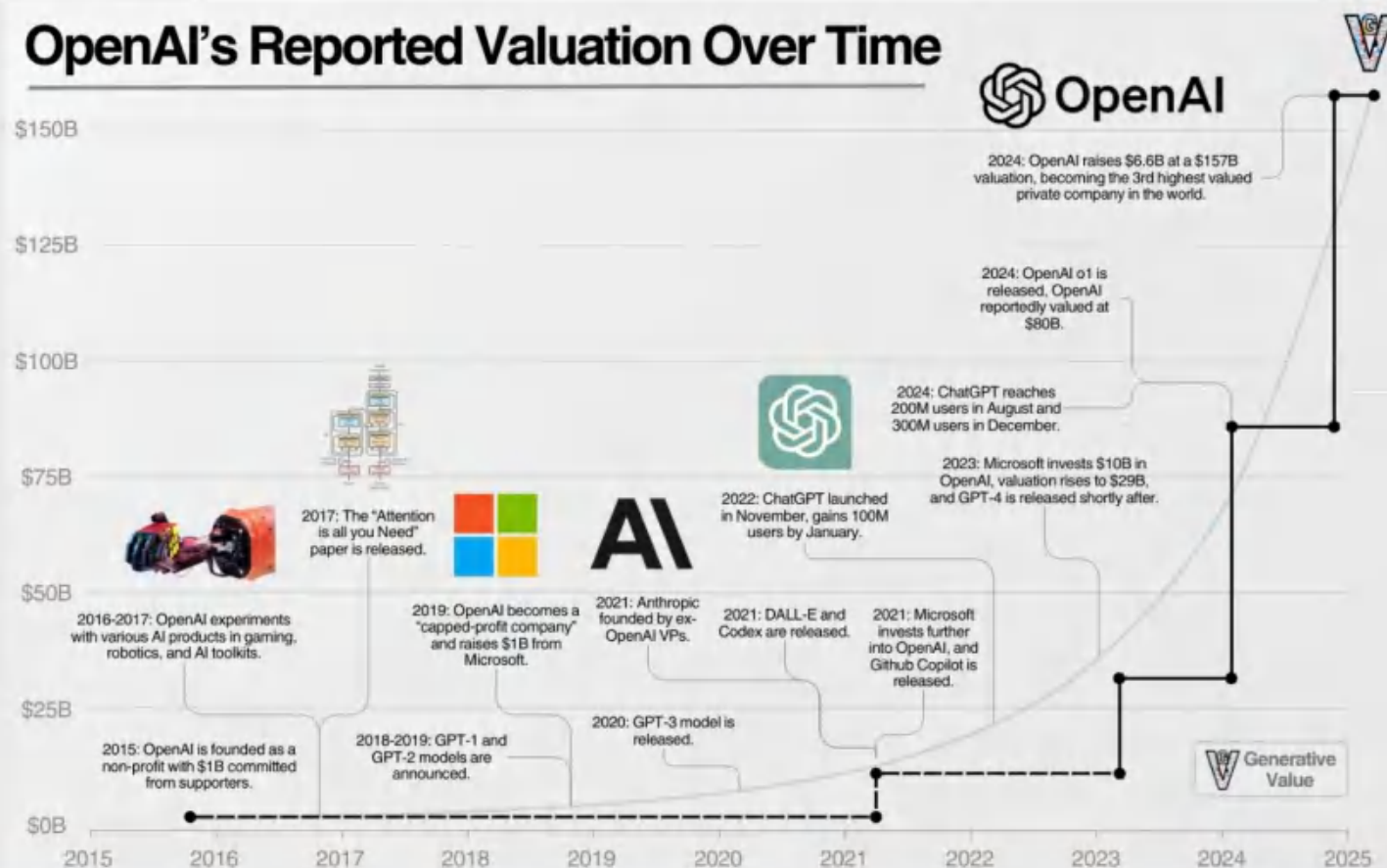
最宽泛MIT协议似乎就代表了态度
欢迎所有人但不做倾向？
API为什么充值都关了？
之后的定价和API有多重要？
或者完全不重要？

要对外融资么？

财务投资看起来应该是不用了
那战略投资呢？
如果站队似乎都是坏处……
问题再延展，如果是国家队呢？
接还是不接，是个问题……



OpenAI's Reported Valuation Over Time



过去2年行业的主叙事



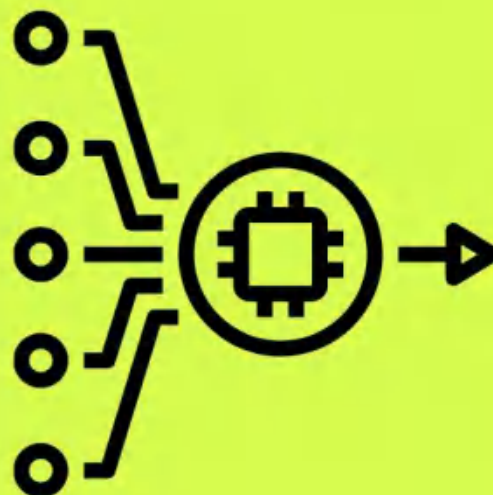
More Money



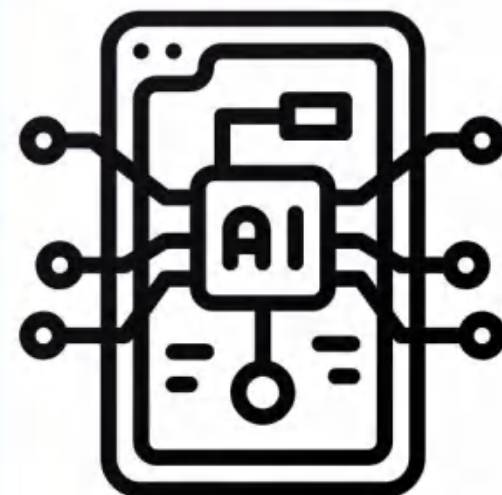
Better Infra



Better traning



Better AI LLM



简单粗暴的递进关系



THE MAGNIFICENT SEVEN

The Venture Capital frontier & the new AI Wild West

May 2024



SUNDAR
PICHAI

TIM
COOK

ELON
MUSK

MARK
ZUCKERBERG

SATYA
NADELLA

ANDY
JASSY

JENSEN
HUANG



七巨头看作一个整体

\$17.6 Trillion

MAGNIFICENT SEVEN MARKET CAP SOARS TO \$17.6 TRILLION IN 2024

Total Market Cap
Returns 2024
46.5%

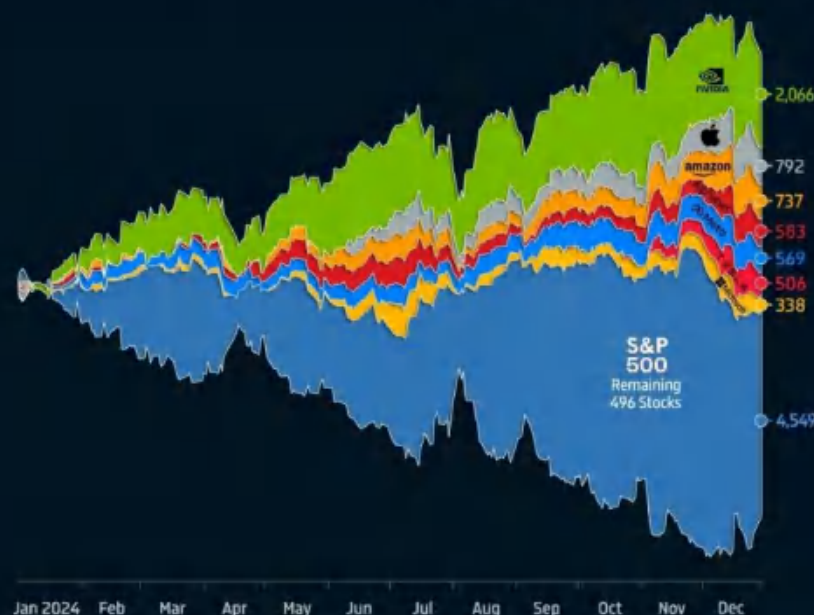


Sources: Nasdaq, Yahoo Finance

www.econovis.net @econovisuals BV

THE MAGNIFICENT SEVEN DRIVE 55% OF S&P 500 2024 GAINS WITH \$5.6 TRILLION GROWTH

Market Capitalization Addition (2024, Billion US\$)



Sources: S&P Global, Nasdaq, Yahoo Finance

www.econovis.net @econovisuals BV

万亿美金的公司

拉长20年时间看



MARKET CAP GROWTH TRENDS OF THE MAGNIFICENT SEVEN IN 2024

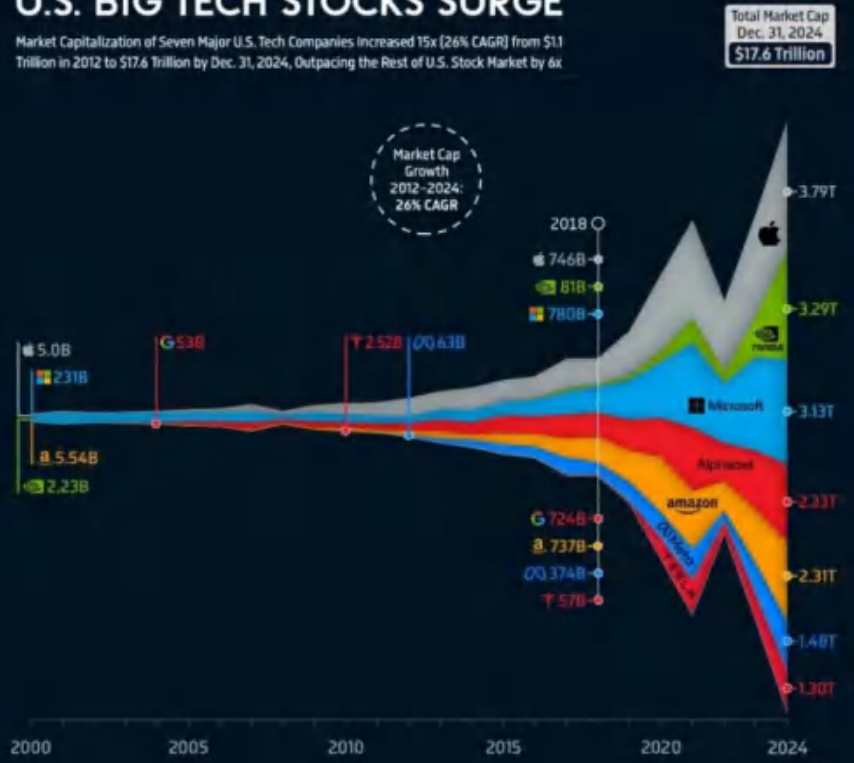


Sources: Nasdaq, Yahoo Finance

www.econovis.net @econovisuals EV

U.S. BIG TECH STOCKS SURGE

Market Capitalization of Seven Major U.S. Tech Companies Increased 15x (26% CAGR) from \$1.1 Trillion in 2012 to \$17.6 Trillion by Dec. 31, 2024, Outpacing the Rest of U.S. Stock Market by 6x



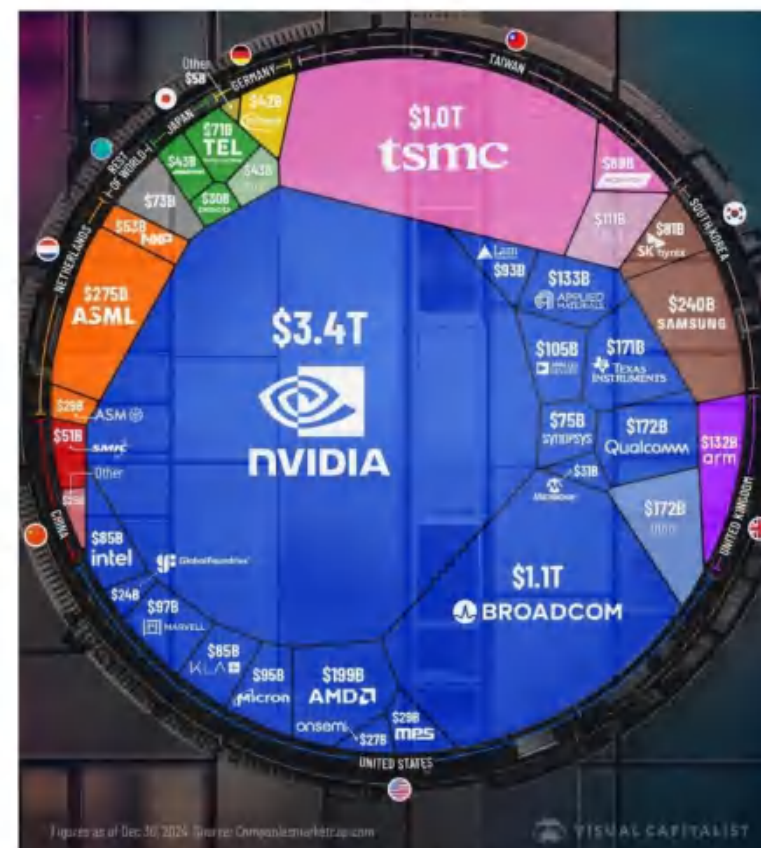
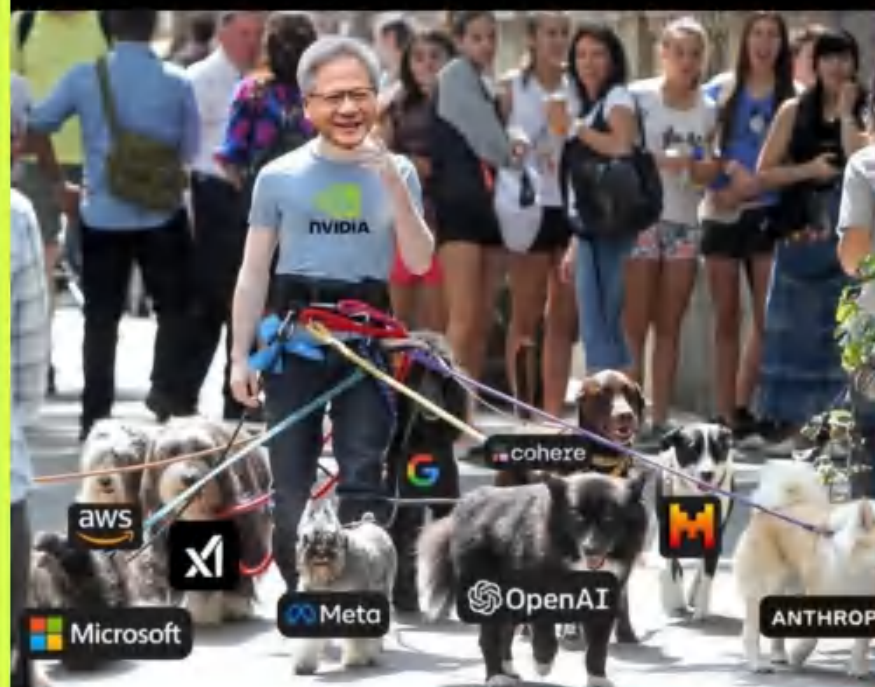
Sources: Nasdaq, S&P Global, Yahoo Finance

Research and visualization: Ehsan Soltani www.econovis.net @econovisuals EV

第一个共识:



The world runs on Nvidia.



第一个共识:



THE BIGGEST SINGLE-DAY STOCK DECLINES

Nvidia has experienced 8 of the 10 biggest single-day stock declines.

- \$197B	Aug. 28, 2024	NVIDIA.
- \$205B	Jul. 24, 2024	NVIDIA.
- \$205B	Jul. 17, 2024	NVIDIA.
- \$206B	Apr. 28, 2022	amazon
- \$208B	Jun. 24, 2024	NVIDIA.
- \$212B	Apr. 18, 2024	NVIDIA.
- \$228B	Jan. 7, 2025	NVIDIA.
- \$251B	Feb. 3, 2022	Meta
- \$279B	Aug. 3, 2024	NVIDIA.
- \$560B	Jan. 27, 2025	NVIDIA.

Nvidia's stock plummeted after a Chinese startup called DeepSeek released a powerful AI model.

As of January 27, 2025
Source: Bloomberg



POST-DEEPSEEK Nvidia Dip-Buyers Were Rewarded



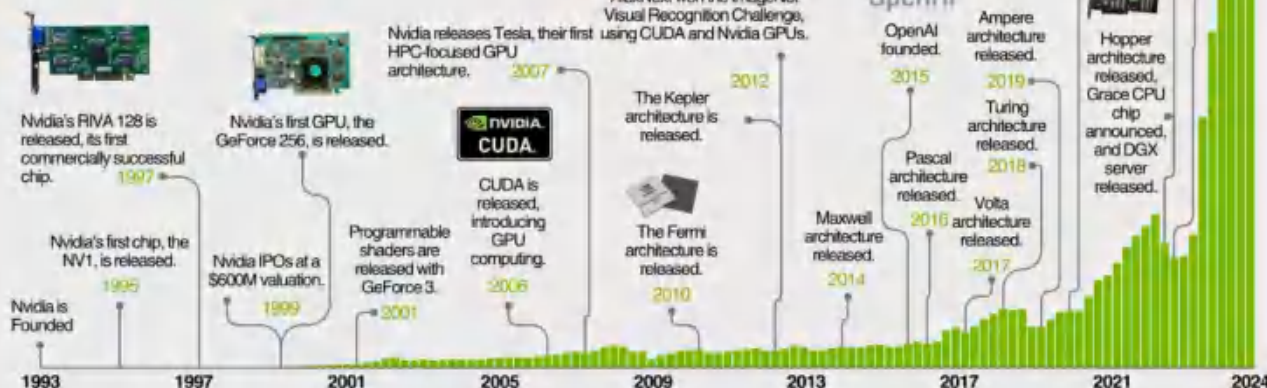
游戏-区块链-AI

三浪叠加

The Rise of Nvidia

The key moments leading to its rise from a market cap of \$600M to \$3.5T.

Nvidia's Quarterly Revenue (since IPO)



NVIDIA's market cap in perspective

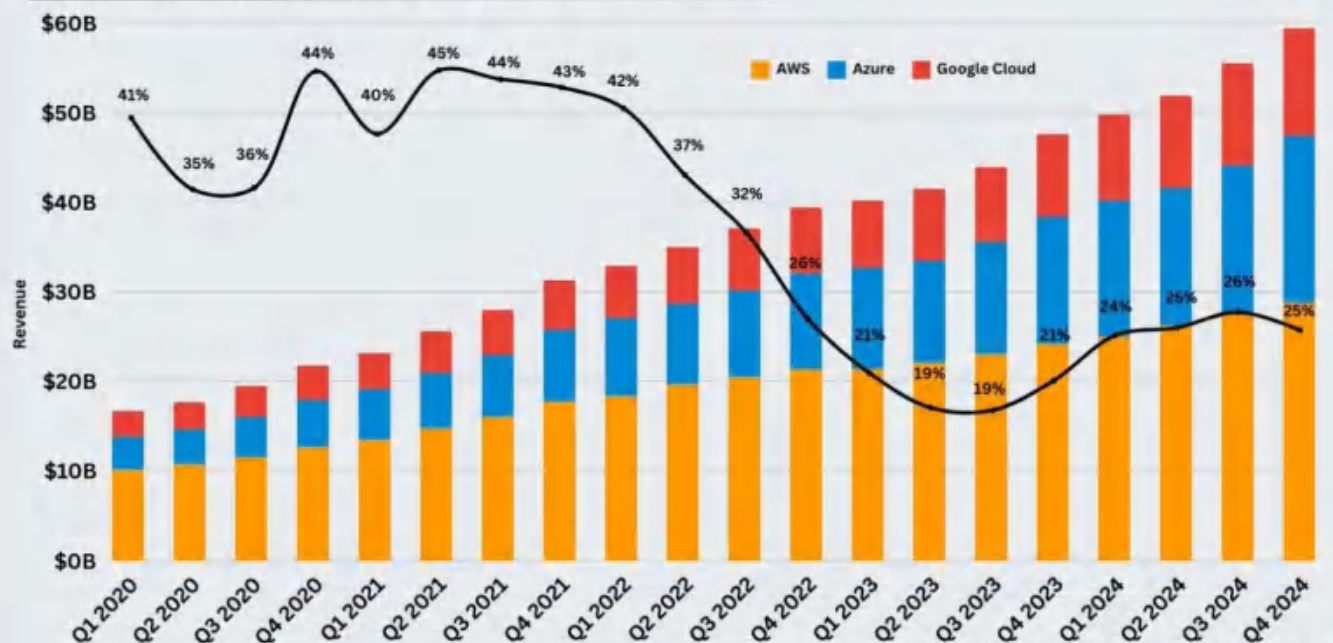


共识2：公有云



Cloud Revenue & Growth

Revenue Run Rate & Aggregate Y/Y Growth rate for AWS, Azure, and GCP

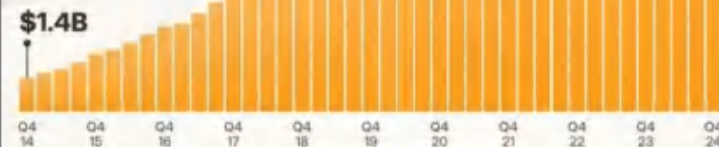


Source: Company Reports; Azure Estimates

Entering the Cloud

Amazon Web Services revenue per quarter

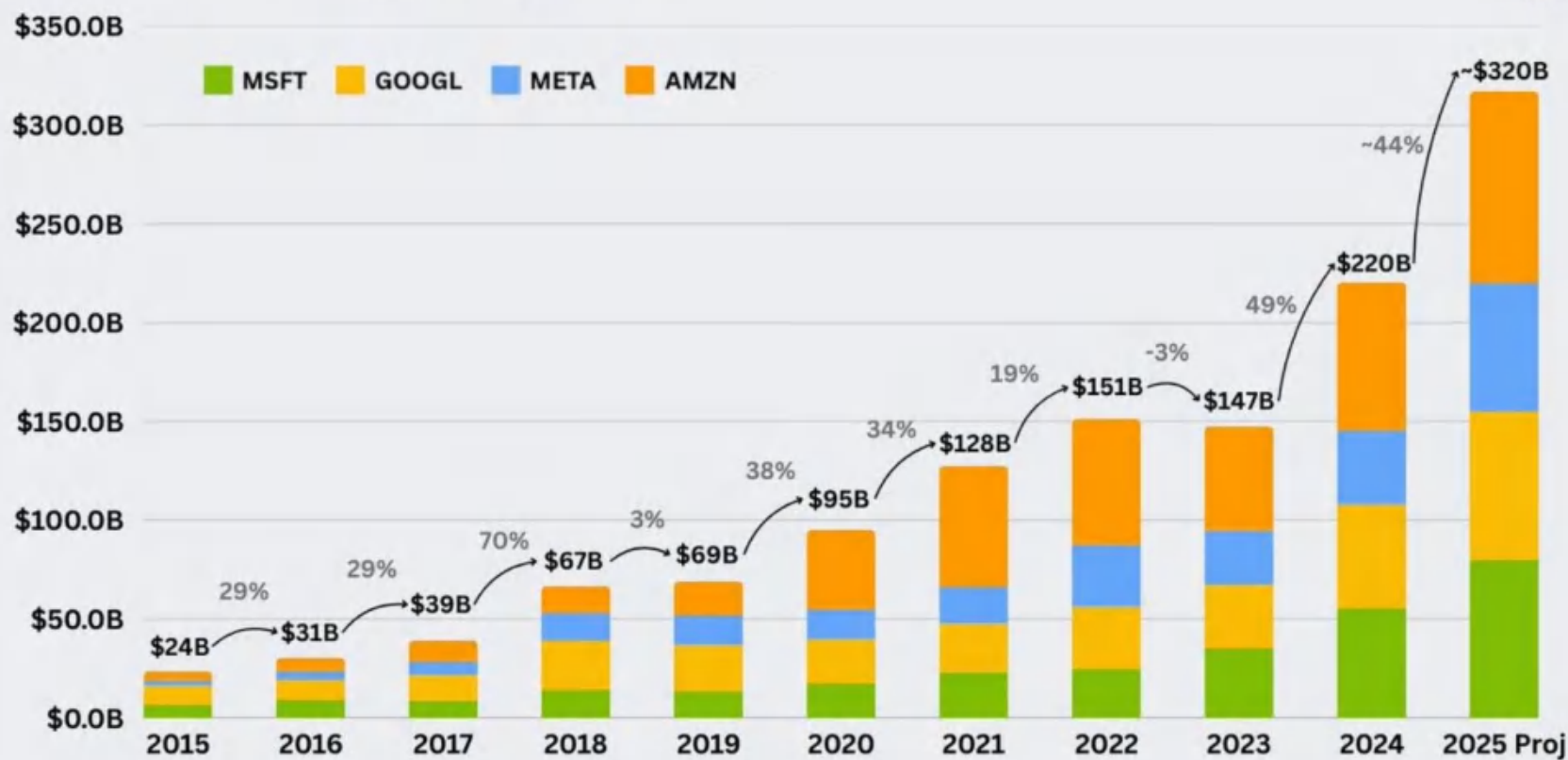
Q4 2014 → Q4 2024
35% 10-year CAGR



共识3: Capex



Hyperscaler Capex By Year

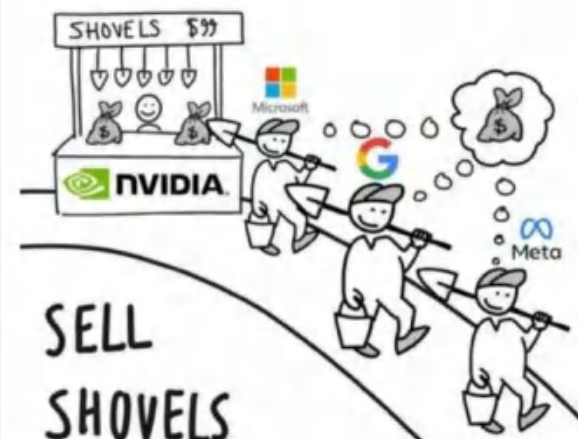


军备竞赛没有尽头



	2024 YE (H100 equivalent)	2025 (GB200)	2025YE (H100 equivalent)
MSFT	750k-900k	800k-1m	2.5m-3.1m
GOOG	1m-1.5m	400k	3.5m-4.2m
META	550k -650k	650k-800k	1.9m-2.5m
AMZN	250k-400k	360k	1.3m-1.6m
XAI	~100k	200k-400k	550k-1m

WHEN EVERYONE DIGS FOR GOLD



成年人的世界，钱是最那啥的标准



Tech Titans spending 10s of billions

Company	2024 Capex Spend (E)
amazon	\$60-65B
	\$55-60B
Google	\$50-55B
Meta	\$35-40B
	\$10-15B

AI Majors spending billions

Company	2024 Funding
OpenAI	\$18.9B
ANTHROPIC	\$7.8B
XI	\$6.1B

AI Challengers spending 100s of millions

Company	2024 Funding
MISTRAL AI	\$1.1B
	\$1.0B
cohere	\$940M
poolside	\$626M
Magic	\$437M
kyutai	\$363M
H	\$220M
black forest labs.	\$196M
ElevenLabs	\$101M

10's of \$B

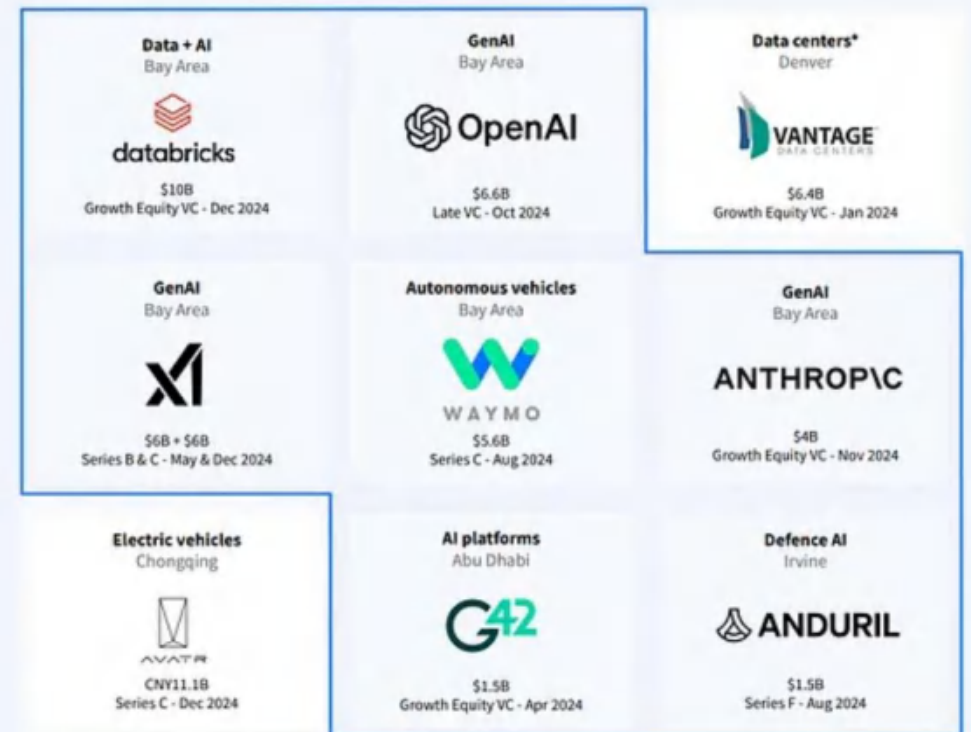
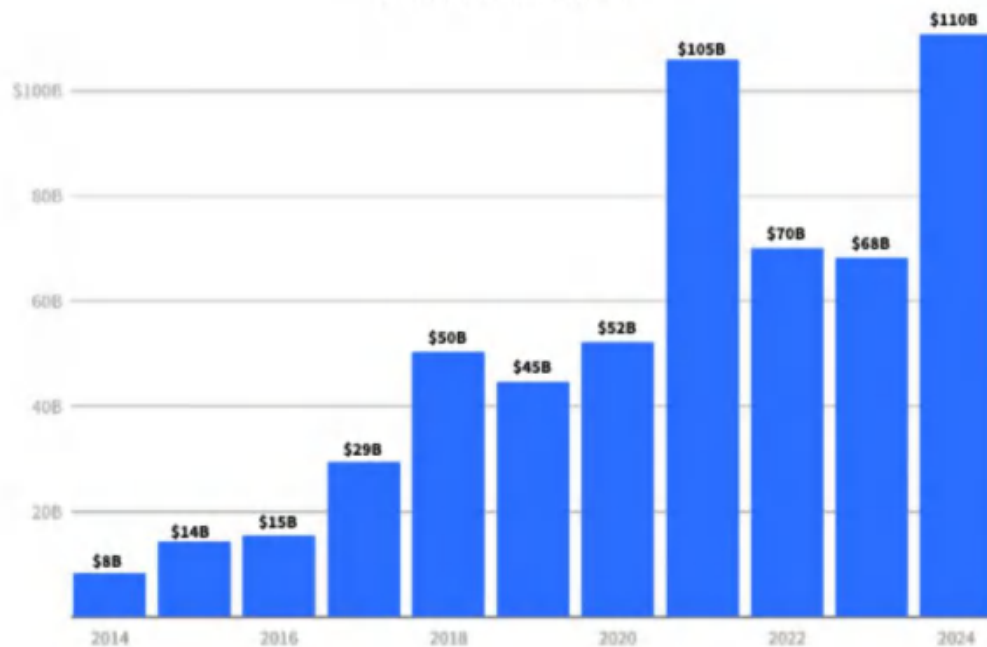
Billions

100's of \$M

不就是钱吗……

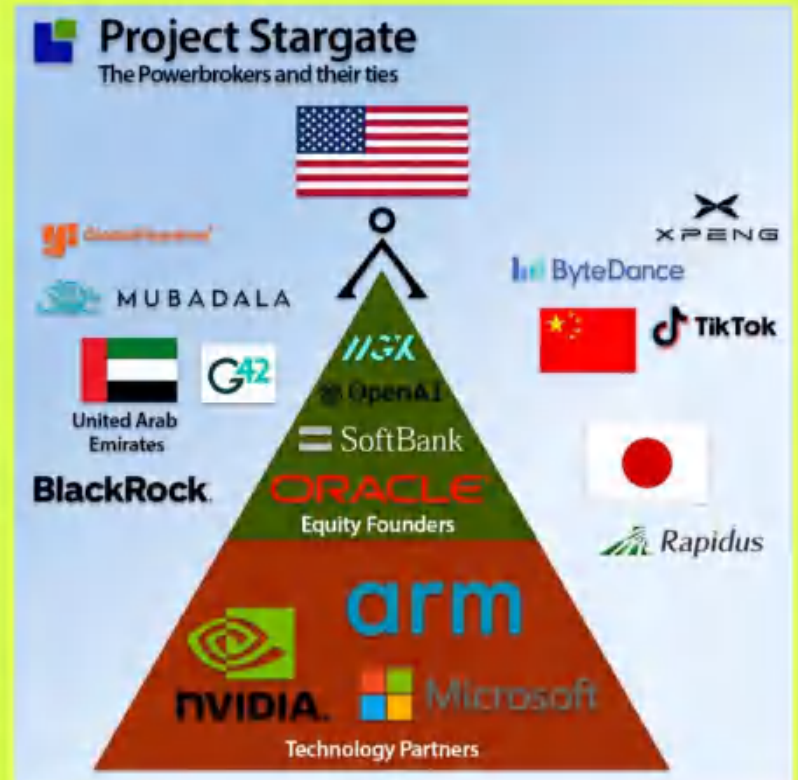


Global AI VC investment



特朗普第一把火

Stargate



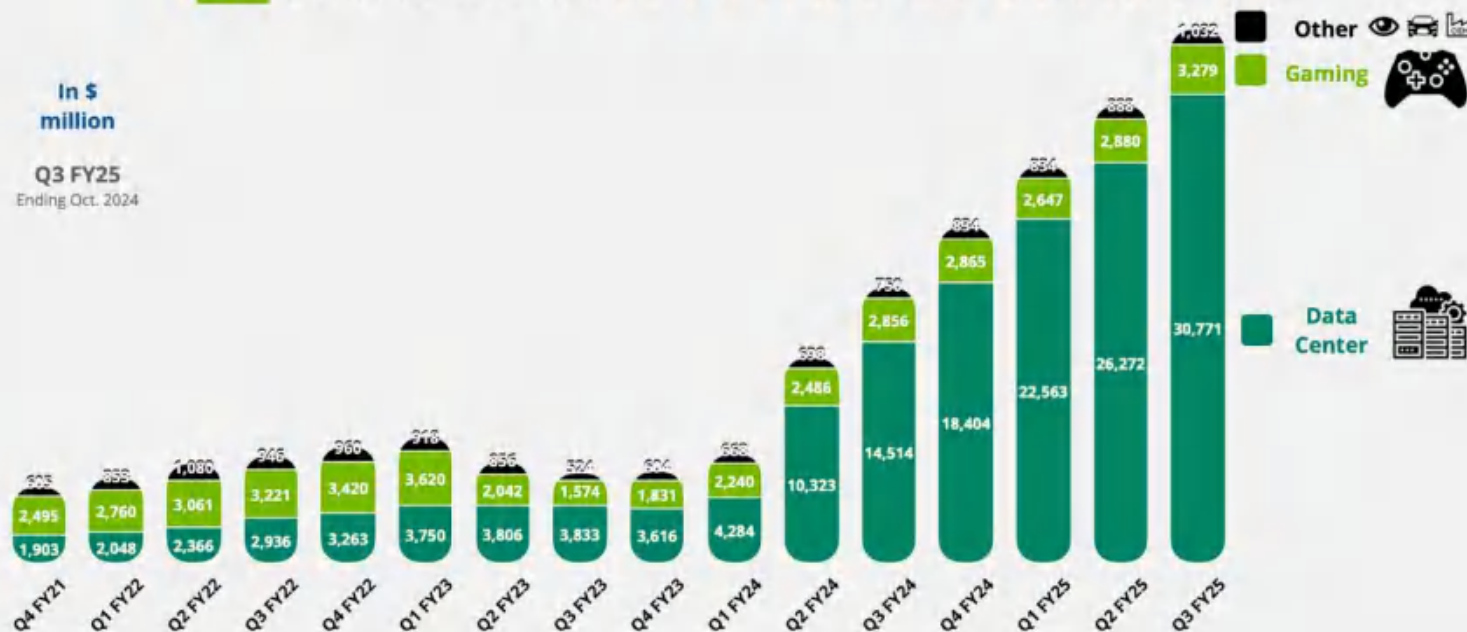
当然，这再次成就了Nvidia



NVIDIA Revenue Breakdown

In \$
million

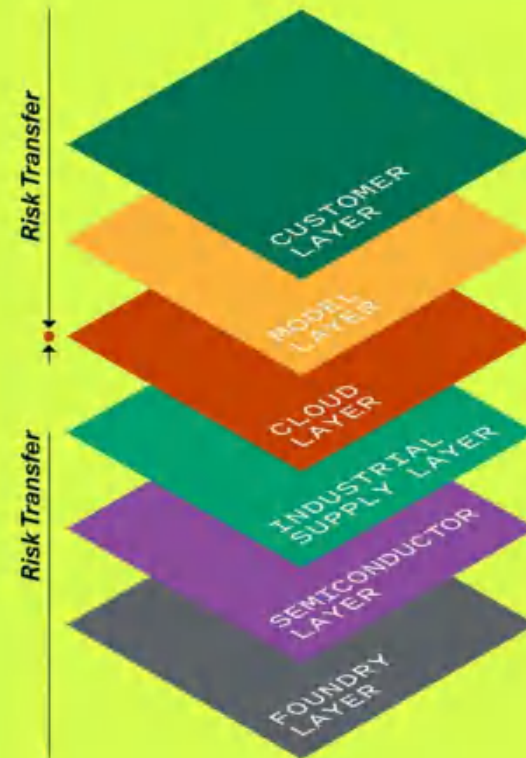
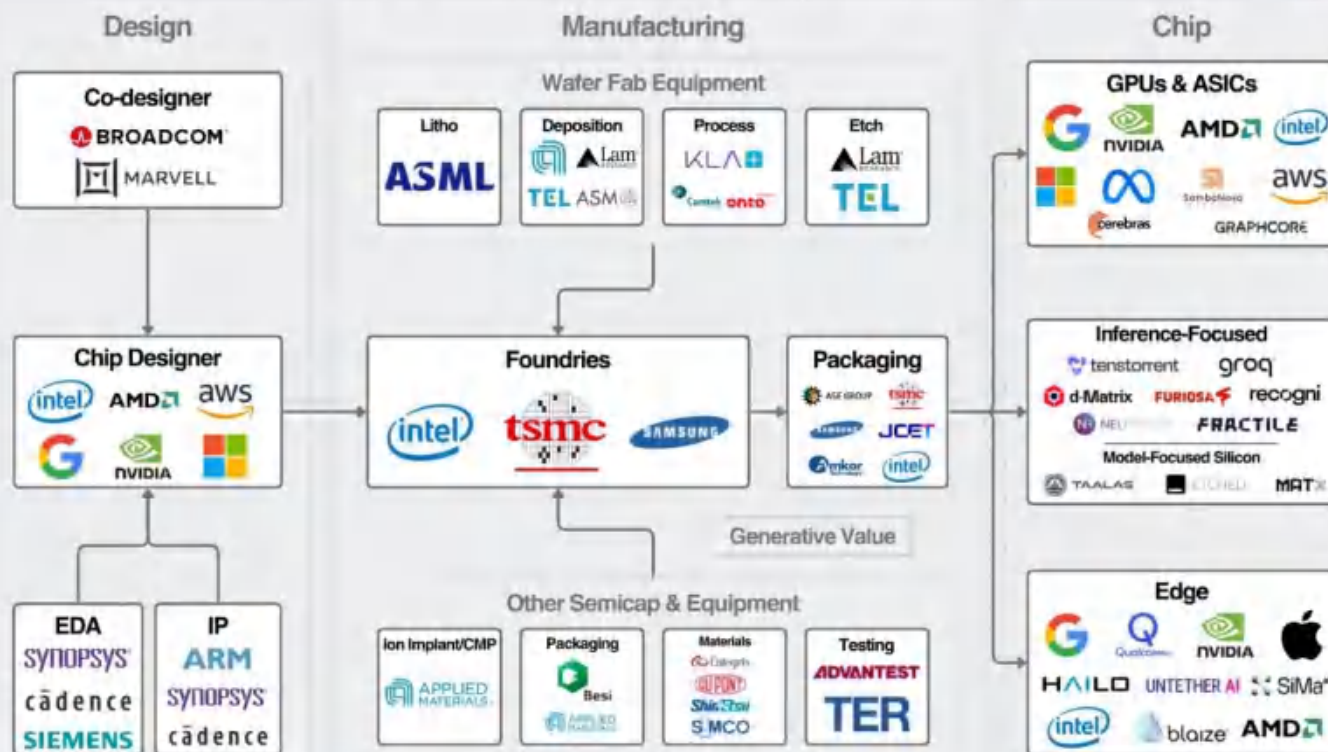
Q3 FY25
Ending Oct. 2024



共识4：往上游去



AI Semiconductor Value Chain



没有人想被卡脖子



Chip rivals pose a risk to Nvidia’s meteoric growth

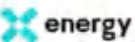
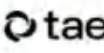
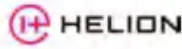
Company	Chip	Purpose	Launch date	Select AI customers
	Trainium	Training	2022	ANTHROPIC
	Inferentia	Inference	2019	 Perplexity
	Tensor processing unit	Training & inference accelerator	2017	character.ai
	Maia 100	Training & inference accelerator	2024	To be launched
	MTIA	Training & inference accelerator	2023	Internal use

共识4：往上游去---能源



Big tech companies morph into leading energy innovators to support AI workloads

Data center power usage will more than double from 460 TWh in 2022 to 1,000+ TWh by 2026 – equivalent to Japan's energy consumption*

Category	Big tech	Company/partner	Description	Date
Nuclear	 Microsoft	 Constellation	Reopening Three Mile Island nuclear plant for data centers	September 2024
Nuclear	 Google	 Kairos Power	Agreement for up to 500 MW of nuclear power from its SMRs** (deployment by 2030)	October 2024
Nuclear	 amazon	 ENERGY NORTHWEST  energy  Dominion Energy	Signed agreements to support deployment of SMRs in the US	October 2024
Fusion	 Google	 Commonwealth Fusion Systems  otae	GV portfolio companies	2021, 2022
Fusion	 Microsoft	 HELION	Aims to buy fusion-produced electricity starting in 2028	May 2023
Geothermal	 Meta	 Geothermal Energy	Collaborating to deliver 150 MW of geothermal power to data centers starting in 2027	August 2024
Geothermal	 Google	 FERVO ENERGY	Developed a 3.5 MW geothermal power plant to power data centers in Nevada	November 2023

共识5：战略投资与并购



Big Tech gobbles up AI model front runners

Resulting in Majority or Minority Control of Several Key AI Assets



Source: Pitchbook, Morgan Stanley Research

Generative AI is the new battleground

Big tech has formed overlapping alliances, with commitments into the billions

Generative AI companies with two or more big tech investors (as of 1/30/2025)

✓ Indicates where big tech invested

	amazon	Google	Microsoft	Meta	nvidia
A DEPT	✓		✓		✓
AI21 labs		✓			✓
ANTHROPIC	✓	✓			
databricks	✓		✓	✓	✓
datology.ai	✓		✓		
essential AI		✓			✓
Hugging Face	✓				✓
Imbue	✓				✓
Inflection			✓**		✓
inworld			✓	✓	
LUMA AI	✓				✓
MISTRAL AI			✓		✓
OpenAI			✓		✓
runway		✓			✓
scale	✓			✓	✓
synthesia		✓			✓
Typeface		✓	✓		✓

Source: CB Insights. Includes investments from venture arms. Only equity deals considered. **Effectively "acquired" by Amazon. ***Acquired by Microsoft.

CBINSIGHTS

所谓掏空式并购



character.ai

Founded in 2021

Full-Stack App

\$150M in Funding

Takeover

August 2024

Google

TechCrunch

Exclusive: Character.AI CEO Noam Shatter returns to Google as the tech giant invests in the AI company

- \$2.7B licensing agreement, incl. \$2.5B to buy out existing shareholders
- Co-founders Noam Shazeer and Daniel De Freitas rejoin Google
- Google hires part of Character.AI's employees

Inflection

Founded in 2022

Foundation Model

\$1.5B in Funding

March 2024

Microsoft

TechCrunch

After raising \$1.3B, Inflection is eaten alive by it's biggest investor, Microsoft

- \$650M licensing agreement
- Co-founders Mustafa Suleyman and Karén Simonyan hired to lead Microsoft AI unit
- Microsoft hires most of Inflection's team

ADEPT

Founded in 2021

Full-Stack App

\$415M in Funding

June 2024

amazon

TechCrunch

Amazon hires founder away from AI startup Adept

- Licensing agreement
- Co-founder David Luan hired to lead Amazon AGI Autonomy unit
- Amazon hires ~ 2/3 of Adept's employees

当然，自身业务肯定也不会落下



中国的“巨头”们也一样……



					
Description	Large ecommerce player and Hyperscaler engine, and operator of (Alibaba Cloud), largest shareholder of Ant Group	China's largest search engine, and operator of Wenxin Yiyan, an AI chatbot with a reported ~300m users	Parent company of Douyin (TikTok) and Toutiao, one of China's leading news applications	Global telco leader and one of the world's largest smartphone manufacturers	Parent company of Riot Games and WeChat, the 'all-in-one' 'super app' of China; Hyperscaler with their Tencent Cloud offering
AI Strategy (high-level)	- Release open weights models - More recently launched proprietary models - Offer inference on Alibaba Cloud	- Actively integrating proprietary models into search platform - Long time leader in self-driving AI	Develop proprietary models and integrate across their consumer platforms	Develop proprietary, domain-specific models and offer on Huawei Cloud	Release open weights models and offer proprietary models on Tencent Cloud
Non-Reasoning	Qwen 2.5 Max Intelligence: 79	Ernie 4.0 Turbo Intelligence: 76^s	Doubao 1.5 Lite Intelligence: 77^s	Pangu 5.0 Large	Hunyuan Large Intelligence: 74
Reasoning	QwQ Intelligence: 78^s	—	Doubao 1.5 Pro Intelligence: 80^s	—	—
Text to Speech	✓	✓	✓	✓	✓
Speech to Speech	—	—	✓	—	—
Image Generation	✓	✓	✓	✓	✓
Video Generation	✓	✓	✓	✓	✓
3D Generation	—	—	✓	—	✓
Primary Consumer Apps	Tongyi Qianwen	Wenxin Yiyan, Wenxin Yige	Doubao	Celia	Yuanbao, Yuanqi
Valuation (US\$)	235B ¹	32B ¹	300B ²	128B ³	469B ¹

Other Models Best LLM+

Other Firms with AI Ambitions



Kunlun Tech
SHE: 300418 (Mkt Cap: \$6B)¹

Beijing-based internet group with >300m MAUs; owner of the Opera browser. Launched the SkyWork series of models and AI accelerators



360 Security (Qihoo 360)
SHA: 601360 (Mkt Cap: \$11B)¹

China's largest provider of Internet and mobile security products. Launched the Zhinao series of models under the 360 AI brand



iFlytek
SHE: 002230 (Mkt Cap: \$16B)¹

Leading voice AI company in China with >14,000 employees. Launched the Spark series of models



Meituan
HKG: 3690 (Mkt Cap: \$115B)¹

China's leading shopping platform with >600m DAUs. Cofounder Wang Huiwen returned to lead AI efforts. Investor in multiple frontier AI labs



Xiaomi
HKG: 1810 (Mkt Cap: \$123B)¹

China's leading consumer electronics brand. Launched the MiLM series of small models. Recently poached Luo Fuli, DeepSeek researcher, to run AI lab. Investor in multiple frontier AI labs

当然，还有六小龙们……



Frontier Models by Chinese AI Tigers and Startups

Open Weights LLM ☐ State Backed Entity

		 Moonshot AI	 零一万物 01.AI	 deepseek	 ZHIPU AI		 阶跃星辰 Stepfun	
Description	China AI Tiger and publisher of Talkie AI app (4 th most downloaded in US in 1H24)	China AI Tiger with 2M Chinese character context window model; China's most well-funded AI startup based on available information	Chinese AI startup focused on smaller language models founded by Lee Kai-Fu (author, former head of Google China)	Chinese AI lab originating out of an AI-focused quantitative trading firm	China AI Tiger with nearly ~700k enterprise and developer users	China AI Tiger with a focus on medical AI models founded by Wang Xiaochuan (ex-CEO, Sogou)	First Chinese AI startup to develop a trillion-parameter model; founded by Jiang Daxin (ex-Chief Scientist, Microsoft Research Asia)	
Best LLM ¹	Non-Reasoning	MiniMax-Text-01 Intelligence: 76 	V1-128k Intelligence: 52	Yi-Lightning Intelligence: 73	V3 Intelligence: 79 	GLM-4-Plus Intelligence: 70	Baichuan 4-Turbo Intelligence: 65	Step-2-16k Intelligence: 82
Other Models	Reasoning	—	Kimi k1.5 Intelligence: 87 	—	R1 Intelligence: 89 	GLM-Zero-Preview Intelligence: 81	Baichuan M1-Preview Intelligence: 83	Step-R-mini Intelligence: 84
	Text to Speech		—	—	—	—	—	
	Speech to Speech	—	—	—	—			—
	Image Generation	—	—	—			—	
	Video Generation		—	—	—		—	—
	3D Generation	—	—	—	—	—	—	—
	Primary Consumer Apps	Hailuo AI Chat, Hailuo AI Video	Kimi	YiChat	DeepSeek Chat	ChatGLM	Bai Xiaoying	Yuewen, PopDuck
Funding Raised (\$)	0.85B ²	1.67B ³	0.2B ⁴	Unknown	1.12B ⁵	1.04B ⁶	Unknown	
Notable Investors (non-exhaustive)								

The DeepSeek Impact

1-day price move 27 Jan 2025



Semiconductor

 **-17%**

 **-13%**

 **-10%**

 **-7%**

 **-7%**

IT Hardware

 **-22%**  **-9%**

Data Centers

 **-9%**  **-4%**

Electricity

 **-28%**  **-22%**

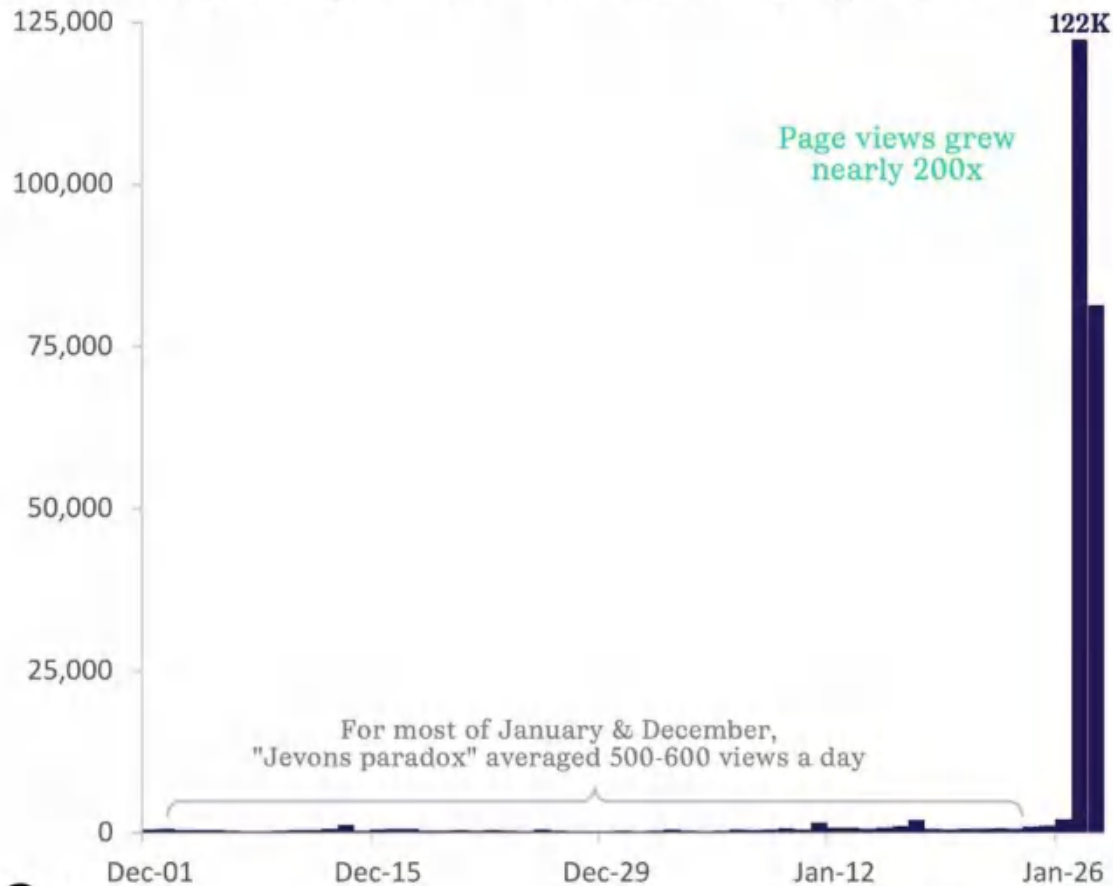
 **-21%**

新的阶段，新的叙事？



Traffic To An Obscure Wikipedia Page About A Paradox Soared This Week

Wikipedia Page Views For "Jevons paradox" [Daily, Wikipedia]



CHARTR

Source: Pageviews Analysis



Jevons paradox

JEVON'S PARADOX

FUEL EFFICIENCY GAINS TEND TO INCREASE,
NOT DECREASE, FUEL USE.

THESE NEW CARS ARE SO
EFFICIENT EVERYONE'S
DRIVING EVERYWHERE
THESE DAYS.



sketchplanations

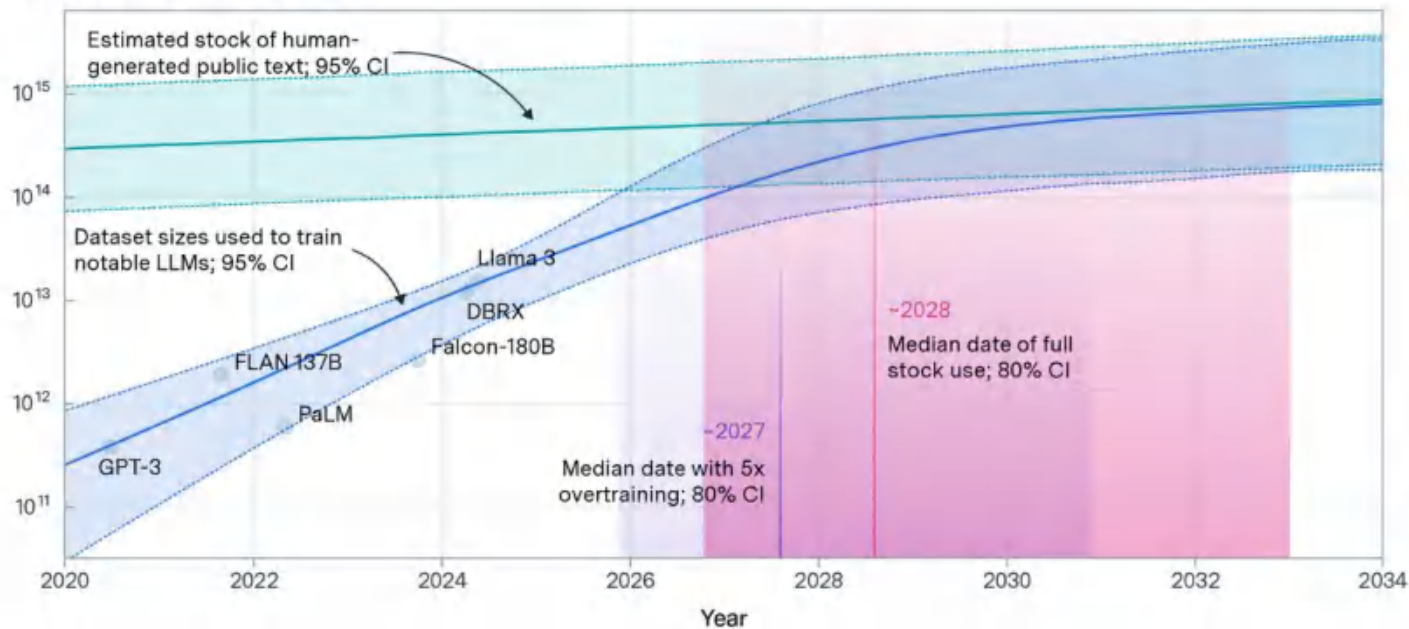
现存数据收敛+预训练失效



Projections of the stock of public text and data usage

EPOCH AI

Effective stock (number of tokens)



Ilya 在演讲中说，“如果你有一个巨大的数据集，训练一个超大的神经网络，那么成功是必然的。”但现在，这个公式似乎失效了。因为数据这个关键变量已经触顶，即便你有再强大的计算资源，也无法突破这个瓶颈。



蔡荔谈AI

Ilya向全世界宣布：预训练的黄金时代已经结束了！

新的范式： 强化学习

2024-09

强化学习（Reinforcement Learning）让 AI 用随机的路径尝试新的任务，如果效果超预期，那就更新神经网络的权重，使得 AI 记住多使用这个成功的事件，再开始下一次的尝试。



张小珺

【大模型季报】AGI范式大转移：从LLM到self-play RL

2024-06

- 超越GPT4的基准模型
- Computer use
- 文本理解和输出提升
- 数学和编程的巨大提升



Claude 3.5 Sonnet

ANTHROPIC

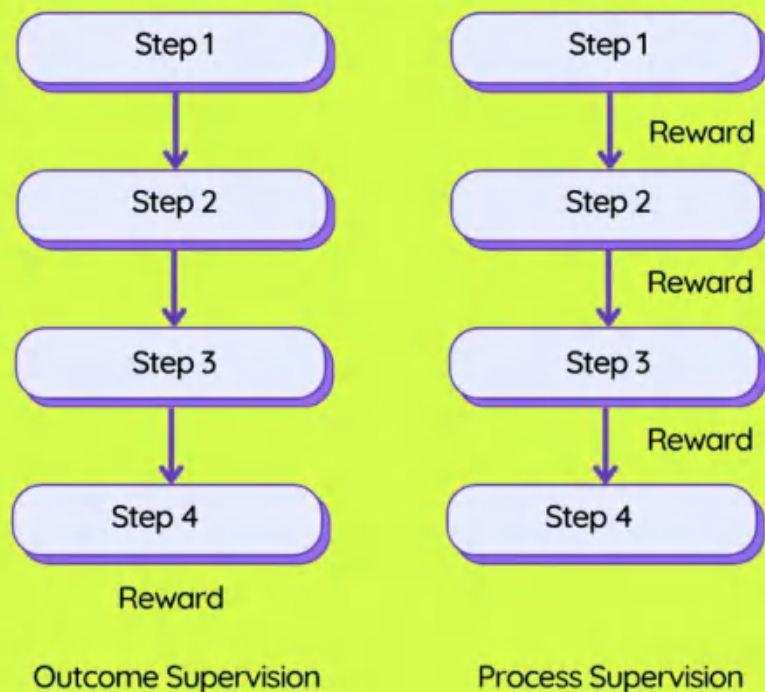
2024-09

- 大模型技术范式改变
- 纯粹意义上的推理模型
- 强化学习与后训练
- 数学、编程、Agent

OpenAI o1



如何复现o1?

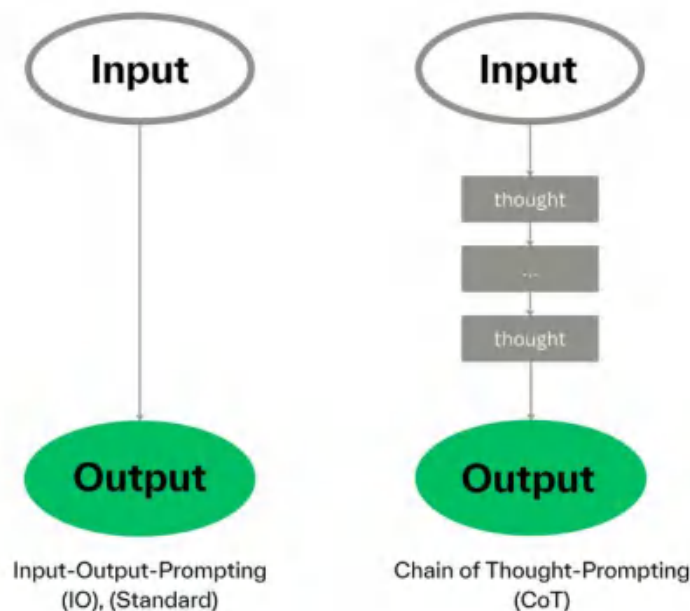


Outcome Supervision

Process Supervision

Process Reward Model

COT+PRM?



Input-Output-Prompting
(IO), (Standard)

Chain of Thought-Prompting
(CoT)

Chain-of-Thought

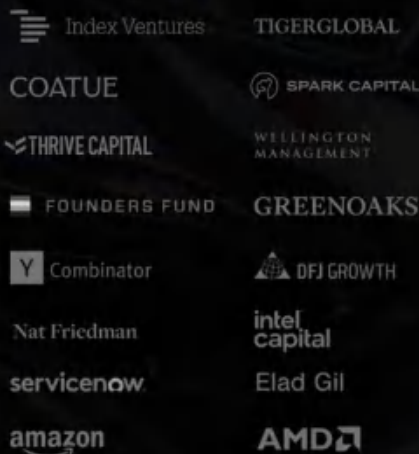
如果PRM路径是对的

scale

Series F

Led by Accel

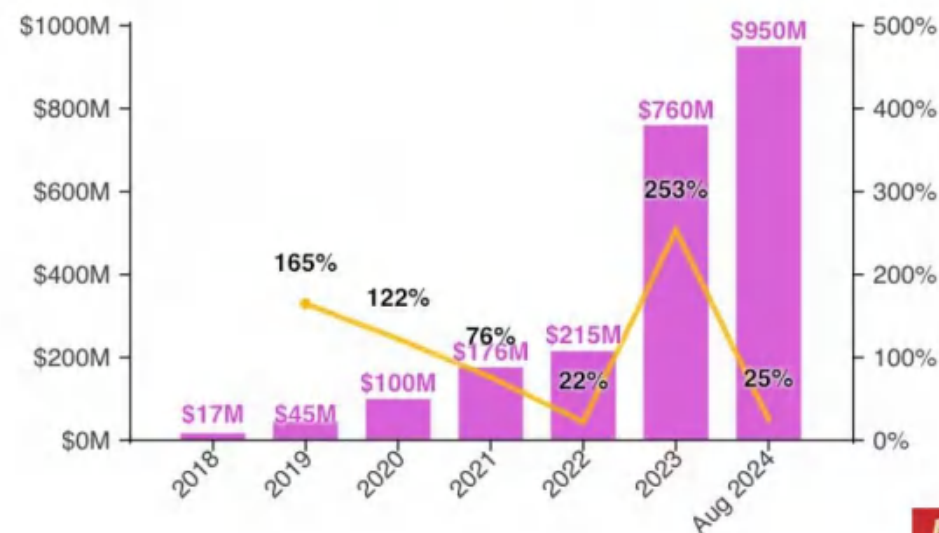
scale



Scale AI

Annualized recurring revenue (ARR)

ARR (\$M) Growth



要训练llm通过RL做题，有精确reward

不要采取结构化的方法，最终会限制住模型的效果，要让模型自己探索思考范式。

思考包含了搜索过程，允许犯错。



Flood Sung

知乎

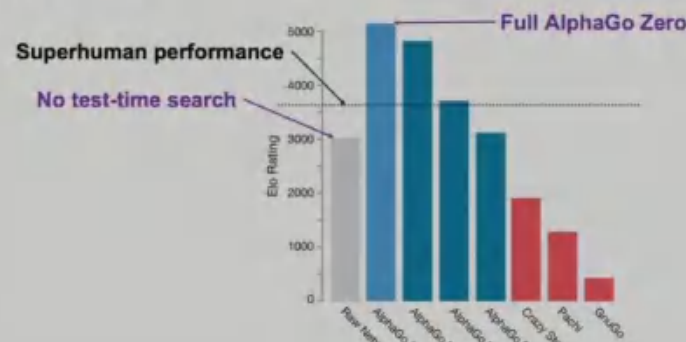
如何评价 Kimi 发布的多模态推理模型 k1.5?

扫码 打开知乎 · 与我共读

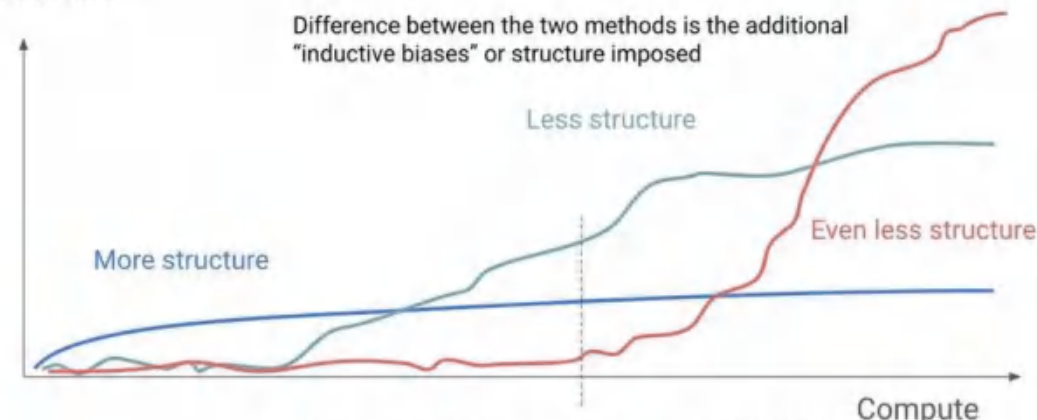


o1似乎不是PMF

Search in Go



Performance



If we are here, we should choose "Less structure", But remember to undo later

知乎 @Flood Sung

Alpha-zero

2017 年 5 月，AlphaGo 以 3:0 击败了柯洁，DeepMind 团队称，有一个比它更强的模型还没出战。[^2] 他们发现，其实根本不需要给 AI 喂人类高手的对局棋谱，只要告诉它围棋的基本规则，让模型自我对弈，赢了就奖励、输了就惩罚，模型就能很快从零开始学会围棋并超越人类，研究人员把这个模型称为 AlphaZero，因为它不需要任何人类知识。



波斯兔子

Deepseek R1可能找到了超越人类的办法

摆脱人类经验

在此之后，围棋变成了比谁更像 AI 的游戏，因为 AI 的棋力已经超越了人类的认知范围。所以，想要超越人类，必须让模型摆脱人类经验、好恶判断（哪怕是来自最强人类的经验也不行）的限制，只有这样才能让模型能够自我博弈，真正超越人类的束缚。



波斯兔子

Deepseek R1可能找到了超越人类的办法

从V3到R1 Zero

这是另一个 AlphaZero 时刻，在 R1-Zero 的训练过程，完全不依赖人类的智商、经验和偏好，仅靠 RL 去学习那些客观、可测量的人类真理，最终让推理能力远强于所有非 Reasoning 模型。



波斯兔子

Deepseek R1可能找到了超越人类的办法

从R1 Zero到R1

先收集了少量高质量的 Chain-of-Thought (CoT) 数据，对 V3 模型进行初步的监督微调，解决了输出语言不一致问题，得到冷启动模型。

然后，他们在这个冷启动模型上进行类似 R1-Zero 的纯 RL 训练，并加入语言一致性奖励。

最后，为了适应更普遍、广泛的非推理任务（如写作、事实问答），他们构造了一组数据对模型进行二次微调。

结合推理和通用任务数据，使用混合奖励信号进行最终强化学习。



波斯兔子

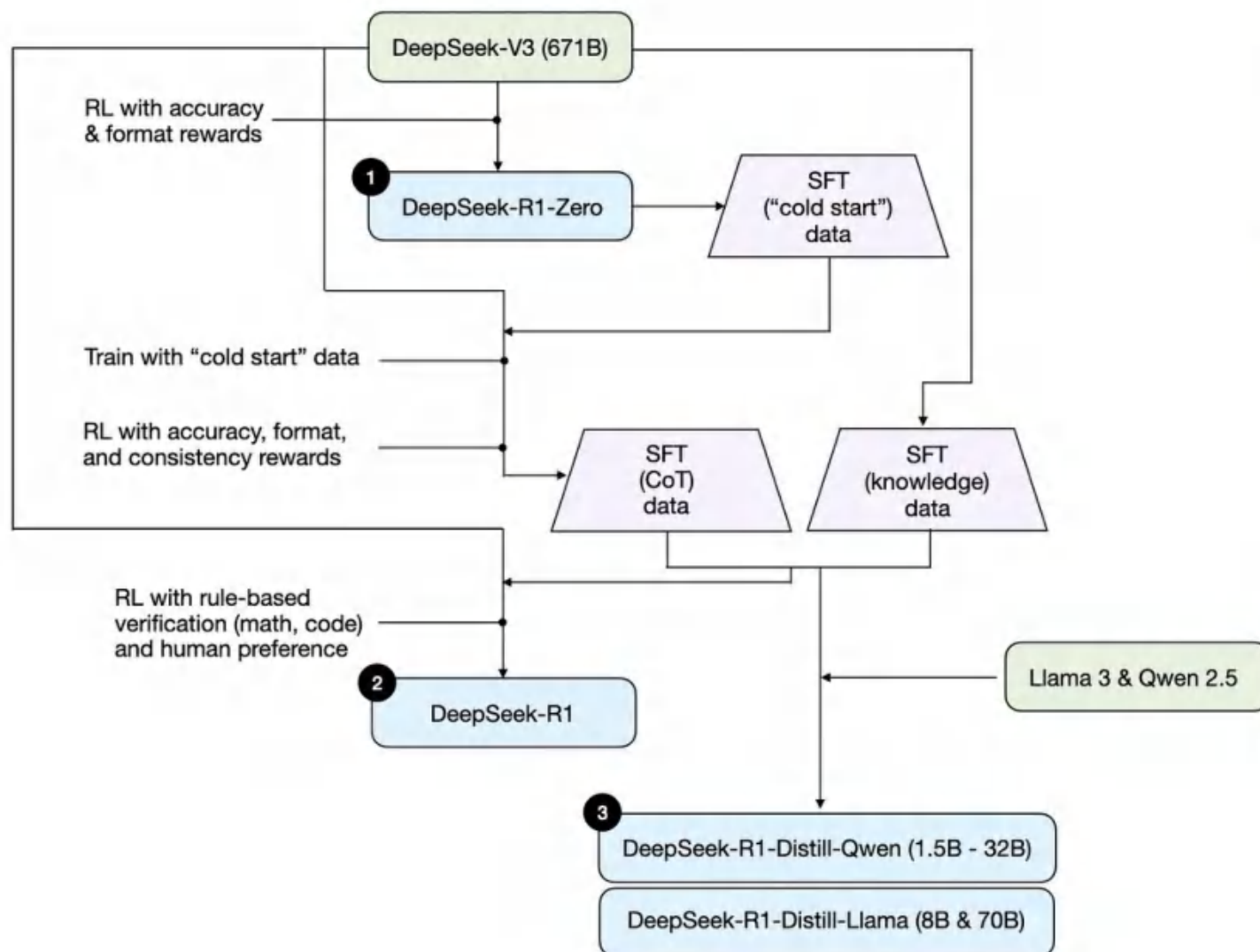
Deepseek R1可能找到了超越人类的办法

Deepseek V3-R1 Zero-R1



R1的实现 似乎看懂了

所谓
大道至简
又因为开源
技术圈沸腾

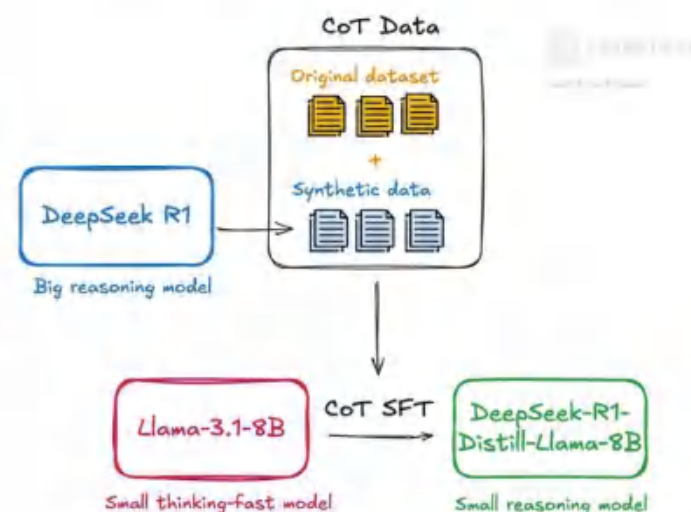


这些打分表我们也懂了



Model	AIME 2024		MATH-500	GPQA Diamond	LiveCode Bench	CodeForces
	pass@1	cons@64	pass@1	pass@1	pass@1	rating
GPT-4o-0513	9.3	13.4	74.6	49.9	32.9	759
Claude-3.5-Sonnet-1022	16.0	26.7	78.3	65.0	38.9	717
OpenAI-o1-mini	63.6	80.0	90.0	60.0	53.8	1820
QwQ-32B-Preview	50.0	60.0	90.6	54.5	41.9	1316
DeepSeek-R1-Distill-Qwen-1.5B	28.9	52.7	83.9	33.8	16.9	954
DeepSeek-R1-Distill-Qwen-7B	55.5	83.3	92.8	49.1	37.6	1189
DeepSeek-R1-Distill-Qwen-14B	69.7	80.0	93.9	59.1	53.1	1481
DeepSeek-R1-Distill-Qwen-32B	72.6	83.3	94.3	62.1	57.2	1691
DeepSeek-R1-Distill-Llama-8B	50.4	80.0	89.1	49.0	39.6	1205
DeepSeek-R1-Distill-Llama-70B	70.0	86.7	94.5	65.2	57.5	1633
DeepSeek-R1-Zero	71.0		95.9	73.3	50.0	1444
DeepSeek-R1	79.8		97.3	71.5	65.9	2029

DISTILLATION



DEAR PRESIDENT TRUMP,

America must
win the AI War.

scale



所以Scale AI CEO必须跳反



No Moat! ! !

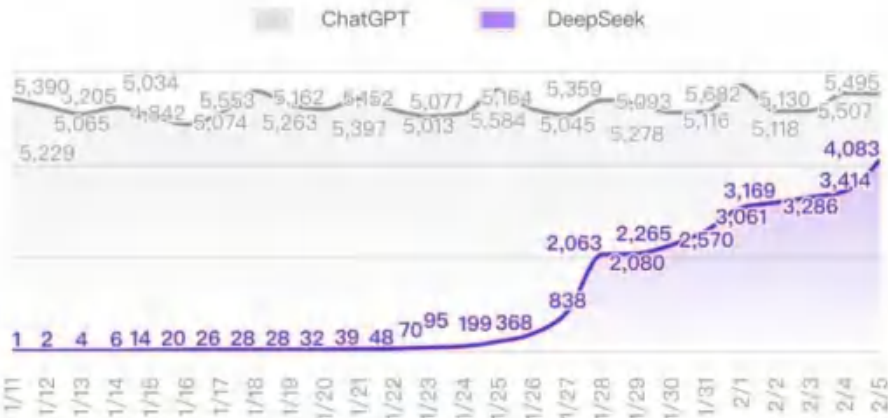
ChatGPT VS DeepSeek

IOOaiapps.cn

流量日报

#非凡产研

ChatGPT VS DeepSeek 移动端 DAU 日活趋势 (单位: 万人)



日期	App-DAU 日活 (万人)	Web-DAU 日活 (万人)	App-DAU 日活 (万人)	Web-DAU 日活 (万人)	移动端 DAU差距
		ChatGPT		DeepSeek	
2/5	5,495	7,810	4,083	1,600	74.30%
2/4	5,507	7,892	3,414	1,549	61.99%
2/3	5,130	7,816	3,286	1,520	64.05%
2/2	5,118	6,407	3,169	1,358	61.92%
2/1	5,682	5,695	3,061	1,257	53.87%
1/31	5,116	6,636	2,570	1,510	50.23%
1/30	5,093	7,402	2,265	1,756	44.47%
1/29	5,278	7,477	2,080	2,110	39.41%
1/28	5,359	7,447	2,063	2,477	38.50%
1/27	5,045	7,544	838	1,674	16.61%
1/26	5,164	6,272	368	590	7.13%
1/25	5,584	5,491	199	385	3.56%
1/24	5,077	6,488	95	302	1.87%
1/23	5,013	7,472	70	269	1.40%
1/22	5,152	7,416	48	237	0.93%
1/21	5,397	7,400	39	188	0.72%
1/20	5,162	7,476	32	148	0.62%
1/19	5,263	6,166	28	120	0.53%
1/18	5,553	5,373	28	112	0.50%
1/17	5,074	6,499	26	116	0.51%
1/16	4,842	7,234	20	128	0.41%
1/15	5,034	7,469	14	143	0.28%
1/14	5,205	7,421	6	139	0.12%
1/13	5,065	7,427	4	145	0.08%
1/12	5,229	6,081	2	133	0.04%
1/11	5,390	5,506	1	124	0.02%



万字赏析 DeepSeek 创造之美：DeepSeek R1 是怎样炼成的？

原创 与你同在的 真格基金 2025年02月13日 15:58 北京

最好的致敬是学习

DeepSeek-R1 赏析



hidecloud&Peak@Monica.im
2025.2

ZhenFund
真格基金

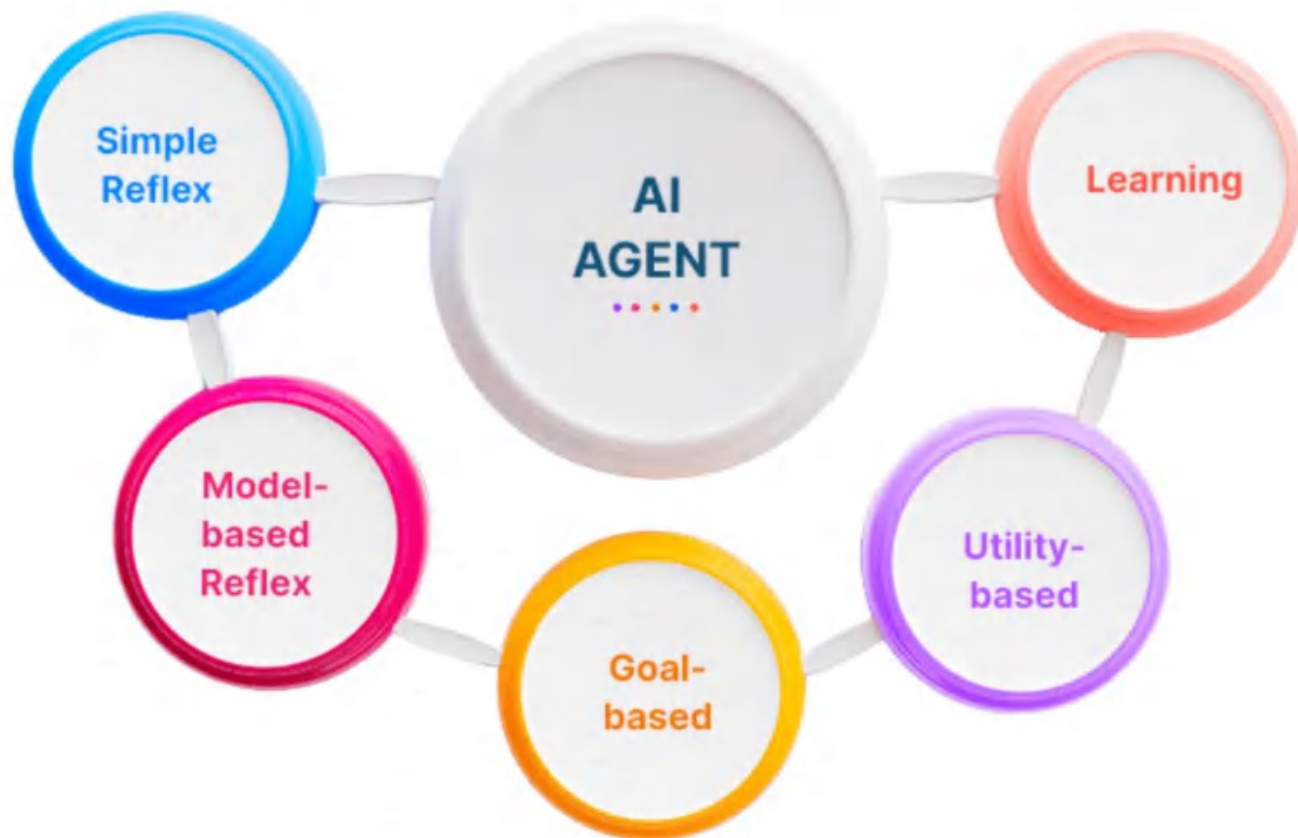
真格基金

♡ 253

最好的致敬是学习：
DeepSeek-R1 赏析
@hidecloud

强烈推荐



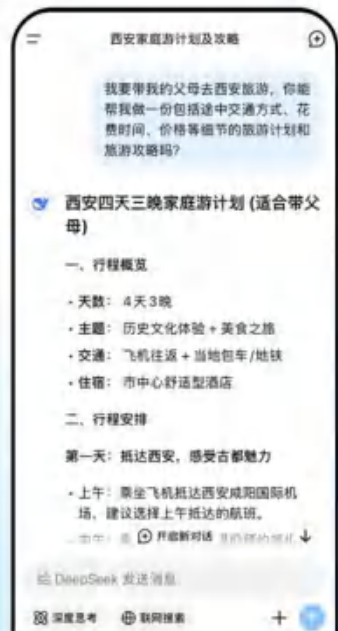


一千个人眼中有一千个Agent概念



智能对话

高智商模型，顺滑对话体验



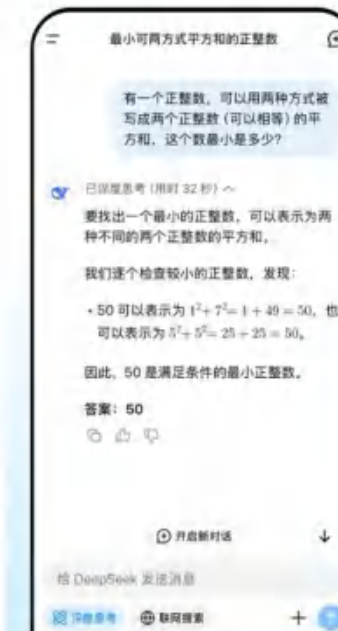
AI 搜索

全网搜索，信息实时掌握



深度思考

先思考后回答，解决推理难题



但这才是

版本: 1.0.4
版本更新日期
2025-01-24

版本对比

DeepSeek - AI 智能助手
AI 智能对话助手，搜索与内容翻译工具

更新日志

- 支持深度思考和搜索同时进行
- 修复部分已知问题

“完全体”



两个按钮带来的四种用法



- o1-\$20-200/月的门槛
- 推理模型没太多人用过
- 更没见过AI的思维过程
- 单纯对话有数据时效性
- 取代搜索是共识中共识

KIMI



为什么思考+联网 同时使用的意义重大?

Plus 入门

\$20 美元/月

借助扩展的访问权限提升工作效率和创造力

获取 Plus

- ✓ Free 套餐中的所有功能
- ✓ 对消息、文件上传、高级数据分析、网页浏览和图片生成功能的扩展访问权限
- ✓ 访问高级语音功能和视频输入
- ✓ 对 o1 和 o1-mini 的有限访问权限
- ✓ 创建和使用项目及自定义 GPT
- ✓ 对 Sora 视频生成功能的有限访问权限
- ✓ 测试新功能的机会

存在限制

Pro

\$200 美元/月

以更高访问级别充分体验 OpenAI 的功能

获取 Pro

- ✓ Plus 套餐中的所有功能
- ✓ 对 o1、o1-mini、GPT-4o 和高级语音（仅限音频）功能的无限访问权限
- ✓ 对高级语音功能的无限访问权限
- ✓ 访问 o1 pro 模型，该模型使用更多的计算来提供高难度问题的答案
- ✓ 对 Sora 视频生成功能的扩展访问权限

无限制，须遵守防滥用规则。了解更多信息



秘塔AI搜索

没有广告，直达结果

请输入，Enter键发送，Shift+Enter键换行

全网 长思考·R1

采用DeepSeek-R1深度推理模型进行回答

纳米AI

DeepSeek-R1-满血版

我可以帮你搜索、答疑、分析、写作、提建议，请把你的任务交给我吧~

DeepSeek-R1-满血版发送消息，Enter发送，Shift+Enter换行

DeepSeek-R1-满血版

联网搜索

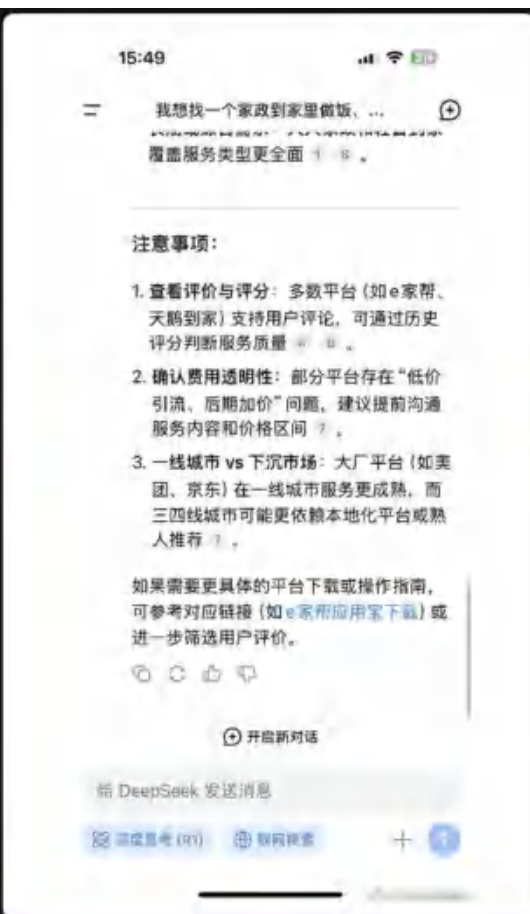
知乎直答

输入你想问的...

深度思考
DeepSeek-R1

既然提到搜索， AI搜索比拼的是什么？

- 不都是搜索吗？
- 信息源能有那么大差异吗？
- 产品设计上的差异足够大吗？
- 之前不都问，AI搜索公司是否要做自己的模型吗？



AI搜索“进化”的另外一个标志



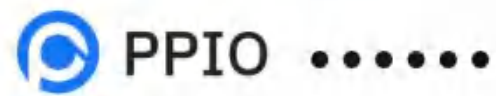
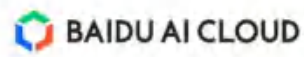
支持DeepSeek-R1的国产云服务平台	
序号	平台名称
1	DeepSeek
2	火山引擎
3	阿里云
4	PPIO
5	百度云
6	腾讯云
7	华为云
8	百度云
9	腾讯云
10	百度云
11	腾讯云
12	百度云
13	腾讯云
14	百度云
15	腾讯云
16	百度云
17	腾讯云
18	百度云
19	腾讯云
20	百度云
21	腾讯云
22	百度云
23	腾讯云
24	百度云
25	腾讯云
26	百度云
27	腾讯云
28	百度云
29	腾讯云
30	百度云
31	腾讯云
32	百度云
33	腾讯云
34	百度云
35	腾讯云
36	百度云
37	腾讯云
38	百度云
39	腾讯云
40	百度云
41	腾讯云
42	百度云
43	腾讯云
44	百度云
45	腾讯云
46	百度云
47	腾讯云
48	百度云
49	腾讯云
50	百度云

这一波对于云和infra厂商，以及进一步大模型生态的影响深远



我没那么多富余人力和卡来应付这些客户的需求，反正是开源的，你们爱咋部署咋部署吧……

好嘞~~~





这些都接入了……



投资分析
三六零能考虑进场吗？
AI自选诊断
你有2只自选股延续强势，值得重点关注
关注 推荐 热榜 投稿

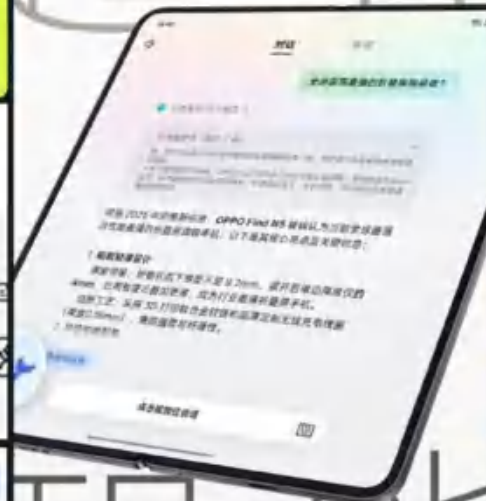
A股市场在春节后5天内上座的概率是多少？

深度思考

圈选分析

图像问答

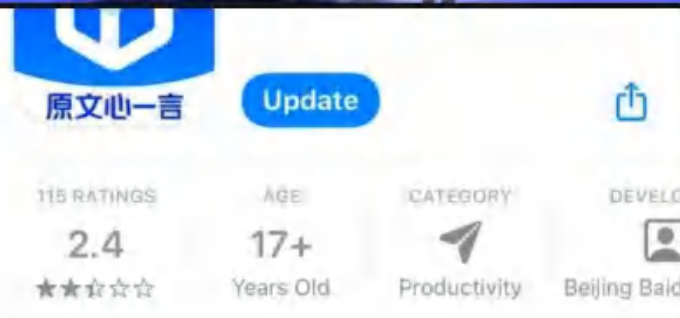
使用深度思考模式回答...



OPPO AI 正式接入 DeepSeek-
支持语音唤醒 支持联网搜索 支持本地导出

OPPO Find N5 折叠旗舰

2月20日 19:00 全球首发



What's New >

Version 4.9.0

上线了伴学模式、手抄报和作文等学习类功能，接入 Deepseek-R1 模型优化了拍照解题功能~

核弹……

确认！微信测试接入DeepSeek

新华社 2025年02月16日 14:30 直播



DeepSeek继续扩大生态圈！2月15日，部分微信用户发现，微信搜索已经上线“AI搜索”功能，并接入DeepSeek-R1提供的“深度思考”服务。

2月16日，记者从腾讯集团确认，微信搜一搜在调用混元大模型丰富AI搜索的同时，正式灰度测试接入DeepSeek。

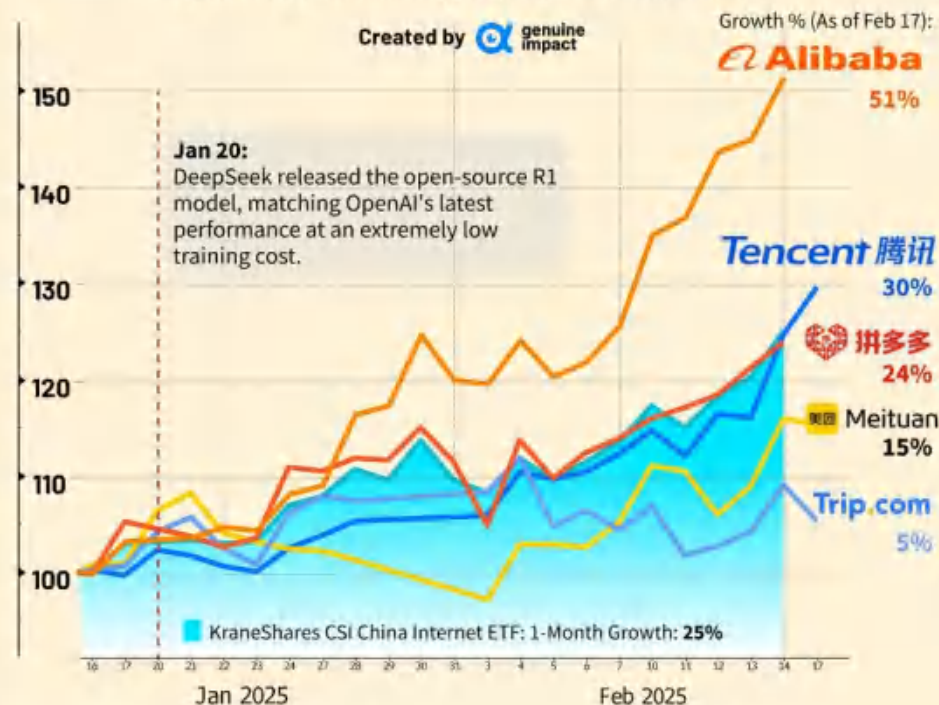
腾讯方面表示，部分测试用户，可在微信对话框顶部搜索入口，看到“AI搜索”字样，点击进入后，可免费使用DeepSeek-R1满血版模型，获得更多元化的搜索体验。若未显示该入口，说明此次灰度测试暂未覆盖到该用户账号，可耐心等待后续开放。

“中国资产重估”…



DeepSeek Boosting Chinese Tech Stocks

Top 5 Kweb Holdings: 1-Month Growth



微信生态接入DeepSeek，抖音怎么做？



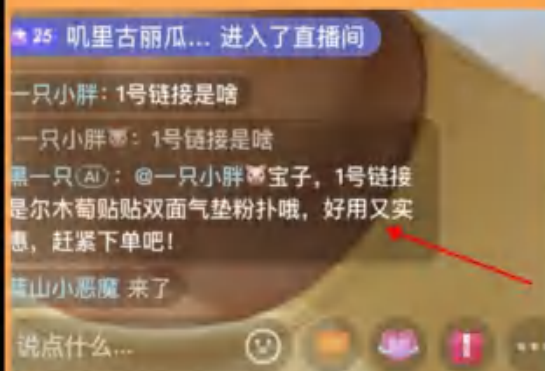
扣子 创建抖音 AI 分身



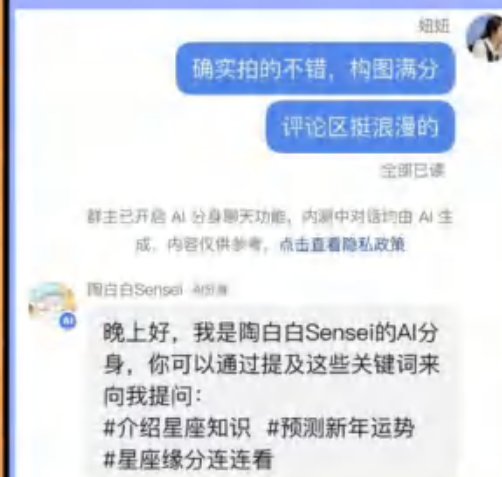
1. 分发到搜索



2. 分发到直播间



3. 分发到群聊



4. 分发到评论区

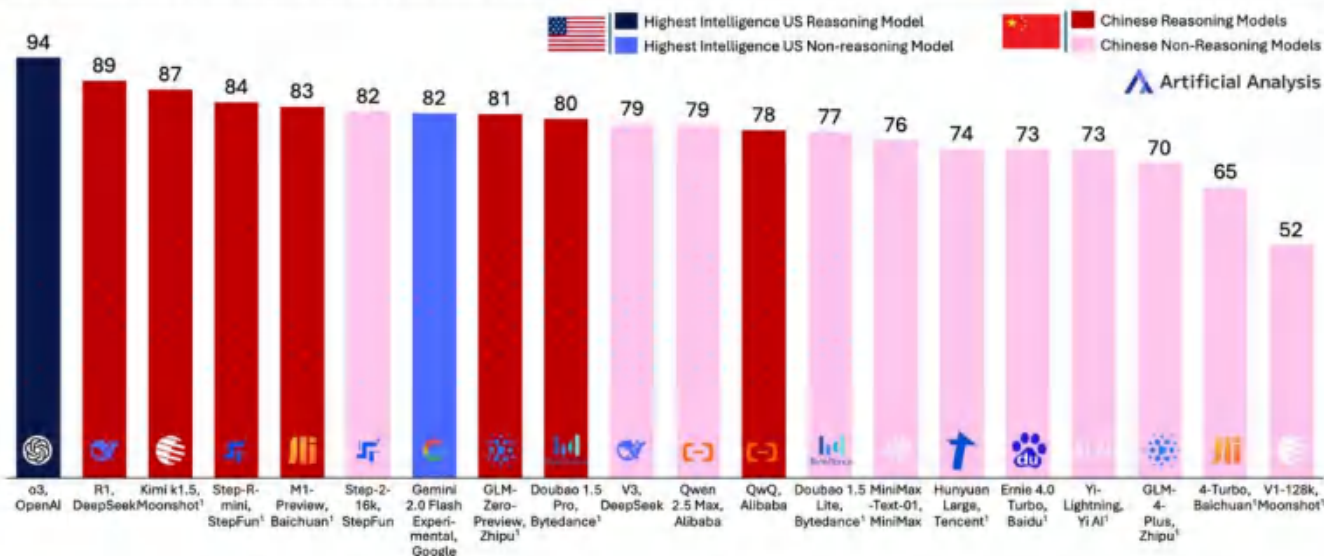


那么如何构建“护城河”呢？



The Language Model Frontier: Models by Chinese AI Labs

Artificial Analysis Intelligence Index, Leading Models (Early 2025), Non-exhaustive



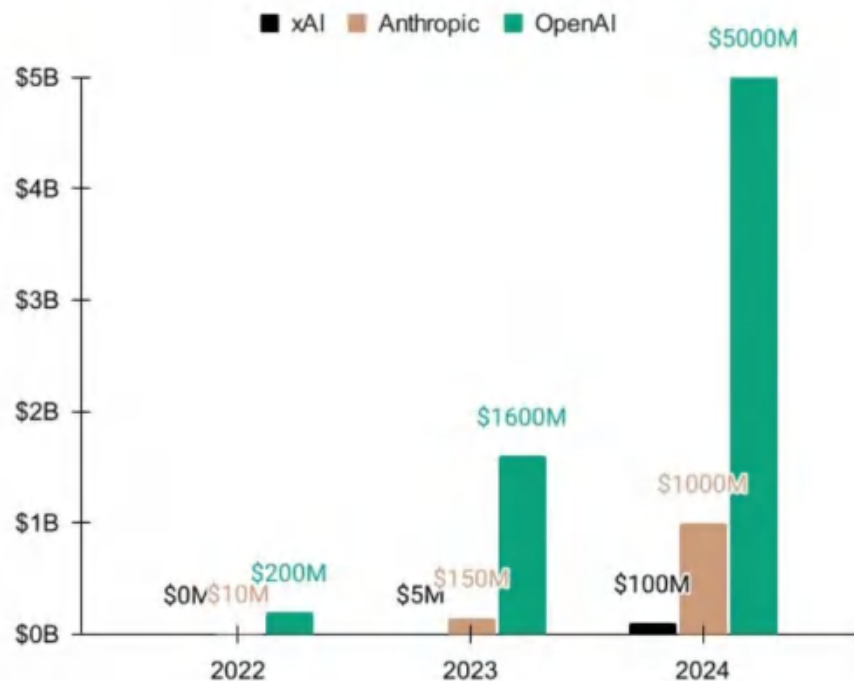
- 产品执行？
- 用户运营？
- 品牌效应？
- 网络效应？
- 垂直行业？

ARR

依然是最粗暴的标准

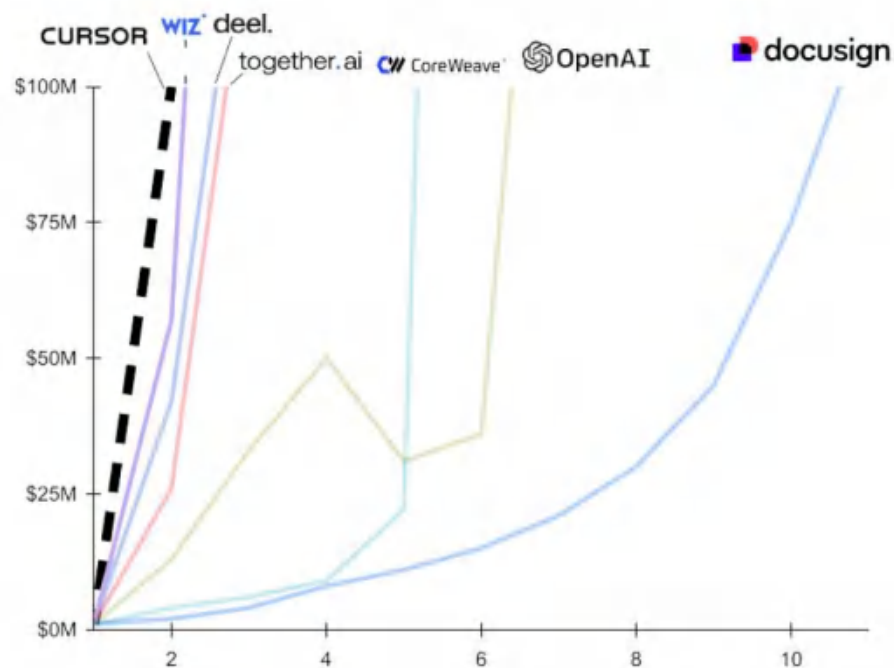


xI vs. ANTHROPIC vs. OpenAI
Annualized recurring revenue (ARR)



CURSOR

Years from \$1M to \$100M ARR

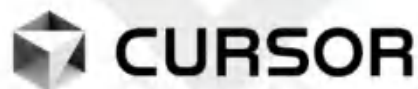




- 根据 prompt 输出代码，无执行调试



- 理解全库环境，提升预测和补全准确率



- 加强逻辑理解，支持文件和命令行操作



- 自主任务执行，规划、执行和调试能力

再看AI-coding



Coding Agent 大概是去年 AI 应用进展最快、引发最多讨论的方向了，从上半年 GitHub Copilot 一家独大到下半年 Cursor 异军突起，引发 Replit、Bolt.new、Windsurf 群雄争霸，直到 Devin 以 500 美金的天价和实打实的效果技惊四座……

这篇访谈难得集结了国内和硅谷 Coding Agent 领域一线的创业者、大模型研究员以及投资人，深入讨论了 Coding Agent 发展与现状、产品设计思考、用户实际反馈、带来的社会变革，以及对 o3 能力、实现难点和未来挑战的思考。



张无常
全网最深度 | 5万字解读 Coding Agent & OpenAI o3

 perplexity = **search agent?**

 Devin = **coding agent?**



Agent的概念被严重泛化了



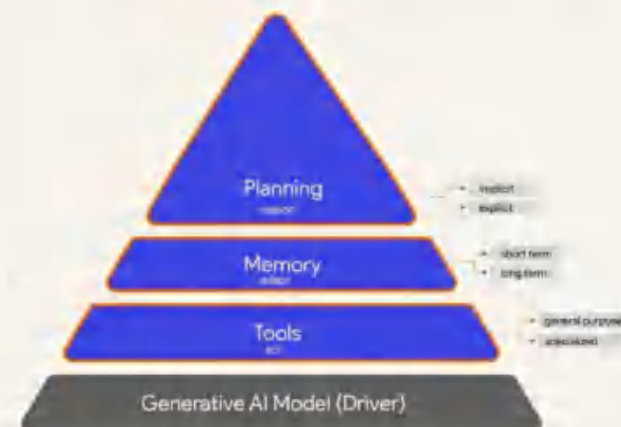
AI Automation & Agents Market Map

INSIGHT PARTNERS



典型对于 Agent 的定义

- 1. Planning
- 2. Memory
- 3. Tools
- 4. 其他 Agent 通信交互





之前

App: 基于各种 APIs 的响应来构建



2024 年趋势

App → Agent (LLMs 调用 APIs)



Before

Models focused on language modeling



2024 Trend

More capabilities (reasoning, planning, tool use etc) "lifted" into Models



之前

Agents 基于代码搭建 (函数、APIs)



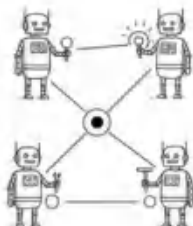
2024 年趋势

Agents 的交互更多基于界面
化窗口 (Web、OS 等)



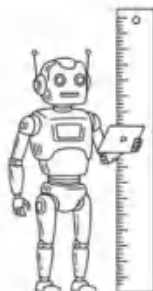
之前

简单 Chains、简单的 Agents



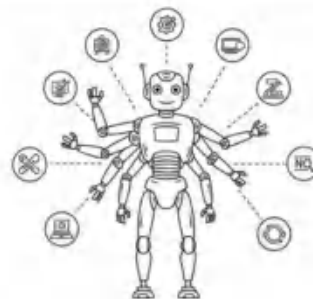
2024 年趋势

复杂的任务和多智能体框架兴起



之前

评估模型本身的能力



2024 年趋势

评估 Agents 处理各种任务分别的能力

AI Agents 2024 Rewind

Top 5 trends in AI Multi-Agent Systems from 2024

1. Enterprises are adopting agents, but with some caveats
2. Teams are building "agent-native" foundation models from the ground up
3. Interface automation agents dominated early commercial applications
4. A Shift to Complex Tasks and the Rise of Frameworks
5. Benchmarks reveal both progress and limitations

Learn more!
newsletter.victordibia.com

24年Agent发展趋势会延续到25



概念太玄乎，我们看一个真-Agent案例



新一代扫地机器人

可能都会增加

一个机械手臂



场景确定

- 杂物清洁
- 辨别物品
- 分类整理
- 记忆位置
- 重量限制
-

技术边界

- 视觉模型
- 推理模型
- 路径规划
- 抓取精度
- 避障逻辑
-

产品交付

- 主动服务
- 模块搭配
- 兼容测试
- 承诺范围
- 定价体系
-

更新迭代

- 用户反馈
- 搜集整理
- 事故案例
- 寻找原因
- 软硬更新
-

为了实现，我们需要……



场景确定

技术边界

产品交付

更新迭代

今天在AI时代，
场景-技术-产品的结合可能是难点；
大模型本身的技术边界其实在逐渐清晰，
而从场景到产品，
恰恰应该是“古典”互联网人的经验所在……





当然，这带来了新一轮组织架构上的取舍选择

- 模型研发应该是纯lab形式吗？
- 模型和产品团队严格分开吗？
- 产品必须只用自家模型吗？
- 模型的迭代需要产品跟进吗？
- 如果再算上运营和商业化呢？



似乎大家有了一定的阶段性共识

智能信息事业群



CSIG
(云与智慧产业事业群)



MEG
(移动生态事业群)



Flow



阿里云



TEG
(技术工程事业群)



TPG
(技术中台事业群)



Seed



可是Deepseek的爆火， 似乎又打破了一些东西...



DeepSeek 这次也验证了一件事：模型即应用，DeepSeek 在产品形式上没有任何创新，核心就是智能+开源，我也不禁思考：在 AI 时代，任何产品和商业模式的创新都比不上智能的创新吗？



海外独角兽
对 DeepSeek 和智能下半场的几条判断

核心问题很简单，技术还不到位。理想状态下，AI 应该能实时生成匹配表情、文字、语音和嘴型，真正做到端到端的多模态交互，但目前这个技术仍需两三年才能成熟。



出海同学会
AI原生的社交杀手级产品会是怎样的？ | No.124期
干货笔记

“能力没有收敛前，谨慎做应用。”



李翔李翔
关于DeepSeek，请教了几位大佬

乐观的一点是，多模态 AI 终将走入千家万户。文字和语音已经较为成熟，视频和动态图像也到了“勉强可用”的阶段。关键问题是，如何在当前产品形态中合理接入逐渐成熟的模态，以增强情感表达？这既需要技术理解，也考验产品决策。小公司可以快速堆叠新技术，但真正的挑战是在正确的节点投入资源，确定何时开始买量和投流。



出海同学会
AI原生的社交杀手级产品会是怎样的？ | No.124期
干货笔记

AI人工智能产业链联盟

每日为你摘取最重要的商业新闻

更新 · 更快 · 更精彩



Zero

AI音乐创作人

水墨动漫联盟创始人

百脑共创联合创始人

人工智能产业链联盟创始人

中关村人才协会秘书长助理

河北北大企业家分会秘书长

墨攻星辰智能科技有限公司CEO

河北清华发展研究院智能机器人中心线上负责人

中关村人才协会数字体育与电子竞技专委会秘书长助理



主要业务:AI商业化答疑及课程应用场景探索, 各类AI产品学习手册, 答疑及课程



欢迎扫码交流

提供: 学习手册/工具/资源链接/商业化案例/行业报告/行业最新资讯及动态



人工智能产业链联盟创始人

邀请你加入星球, 一起学习

人工智能产业链联盟报告库



星主: 人工智能产业链联盟创始人

每天仅需0.5元, 即可拥有以下福利!

每周更新各类机构的最新研究成果。立志将人工智能产业链联盟打造成市面上最全的AI研究资料库, 覆盖券商、产业公司、研究院所等...

知识星球

微信扫码加入星球 ▶



有什么结论吗？

似乎没有

权当是为这两个月

做个记录

THE END

