

迈向通用的人工智能

刘知远 2025

| 大模型是人工智能发展新高地

符号智能 (1950-1990)

能力来源：领域专家



语言学巨擘
N. Chomsky



图灵奖
E. Feigenbaum

获取方法：手工总结（手动）

能力形式：以**知识库等符号系统存储专家知识**

专用智能 (1990-2017)

能力来源：专家标注数据



图灵奖
Judea Pearl



图灵奖
G. Hinton

获取方法：有监督学习（自动）

能力形式：以**任务专用小模型存储任务知识**

通用智能 (2018-)

能力来源：通用无标注数据



OpenAI

获取方法：自监督学习（自动）

能力形式：以**通用大模型存储通用知识**

序列预测：通用Transformer架构+通用序列预测方法，实现

从无标注大数据萃取世界知识

语言



→ [清华, '实验室', 位于, '北京市']..
[class, 'SCLASS', \$INT, '=', '?']..
['(', '(', '10', '+', '4', ')', '^', '2', ')'].

图像



→ [Image of a building]

DNA



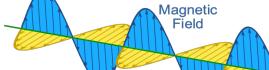
→ 5' ATGACGTGGGA3'
3' TACTGCACCCCT5'

工具使用

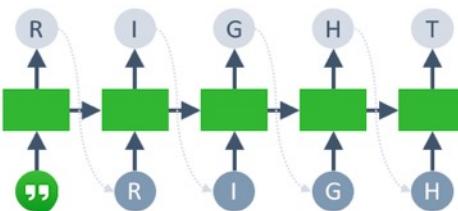


→ [检索, '翻页', '摘取', '翻译', '总结']..

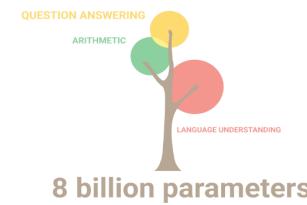
电磁波



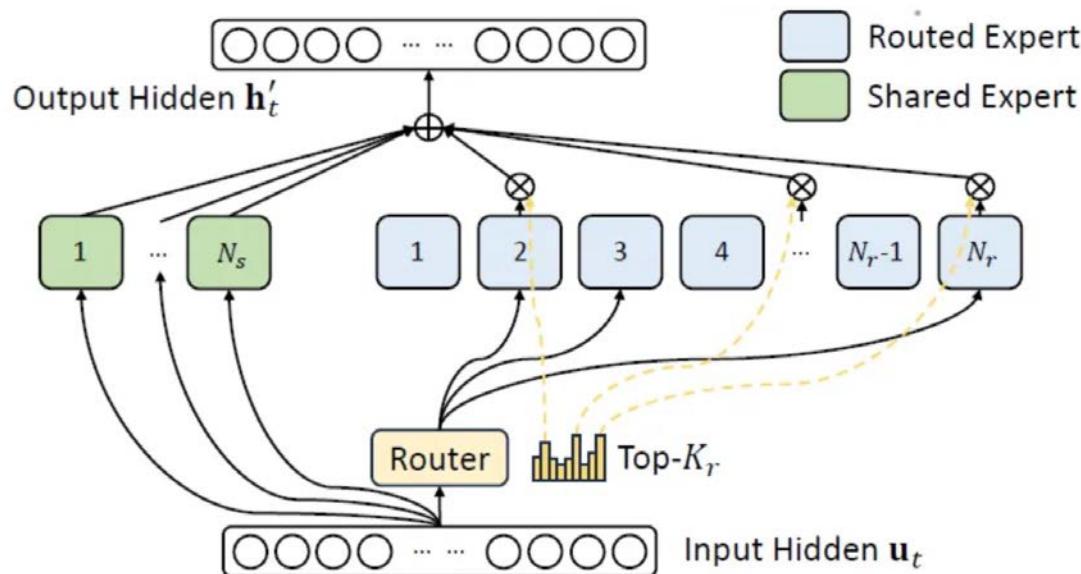
→ [Waves diagram, 'Magnetic Field', 'Electric Field']



智能涌现：大数据+大算力支持知识持续积累，涌现完成复杂任务的能力

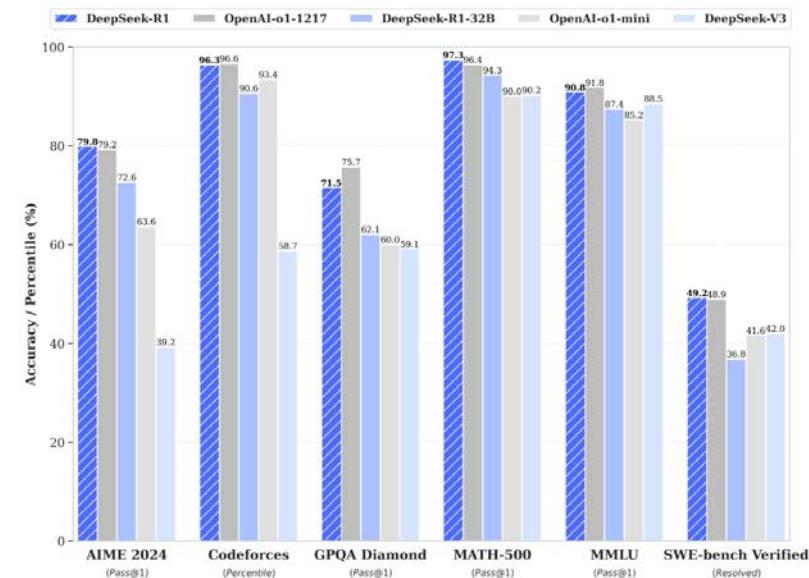


| 以 DeepSeek V3 & R1 为例



DeepSeek V3大语言模型

与OpenAI GPT-4o相当，单次训练成本为
同水平开源模型Llama-3-405B的1/10

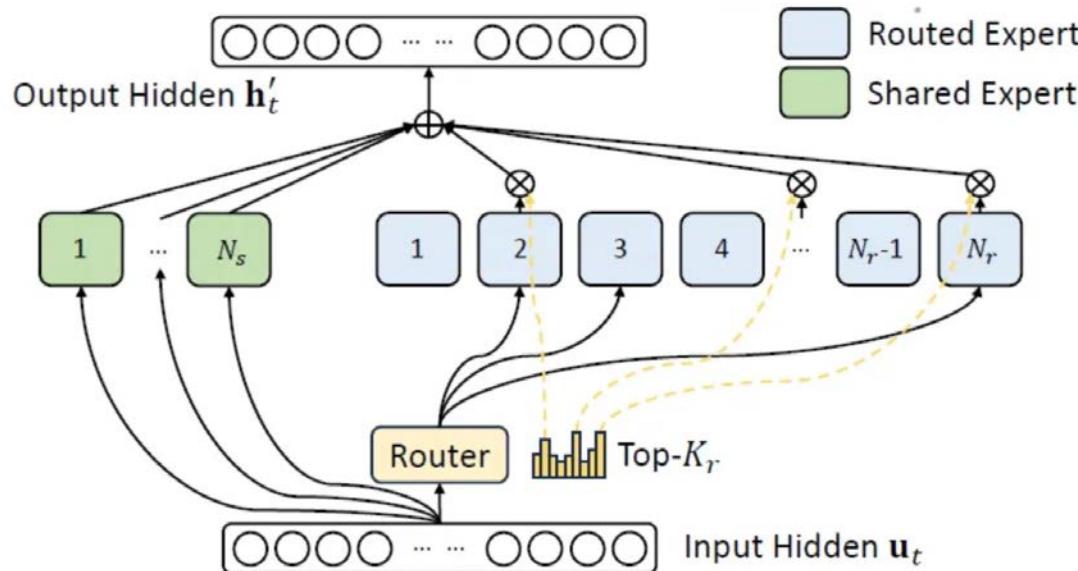


DeepSeek R1深度思考模型

世界首个与OpenAI o1具备相当深度
思考能力的开源模型

| 以 DeepSeek V3 & R1 为例

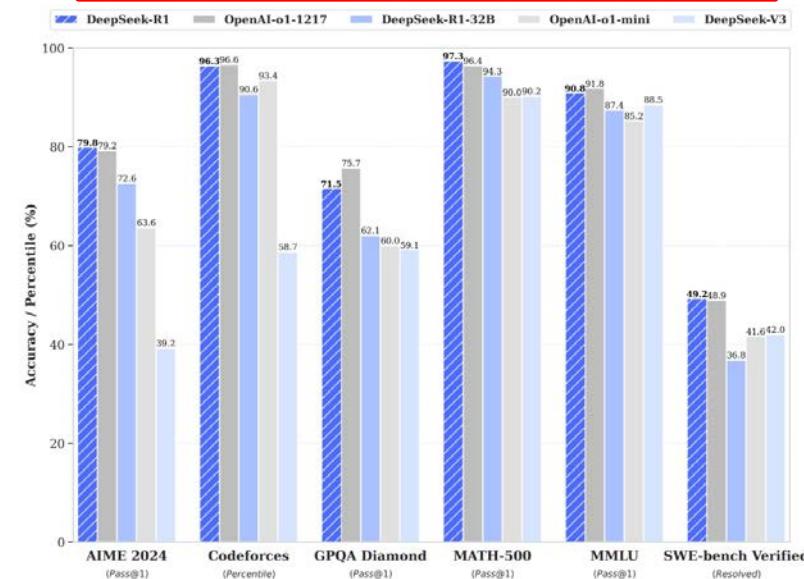
能效更高



DeepSeek V3大语言模型

与OpenAI GPT-4o相当，单次训练成本为
同水平开源模型Llama-3-405B的1/10

能力更强



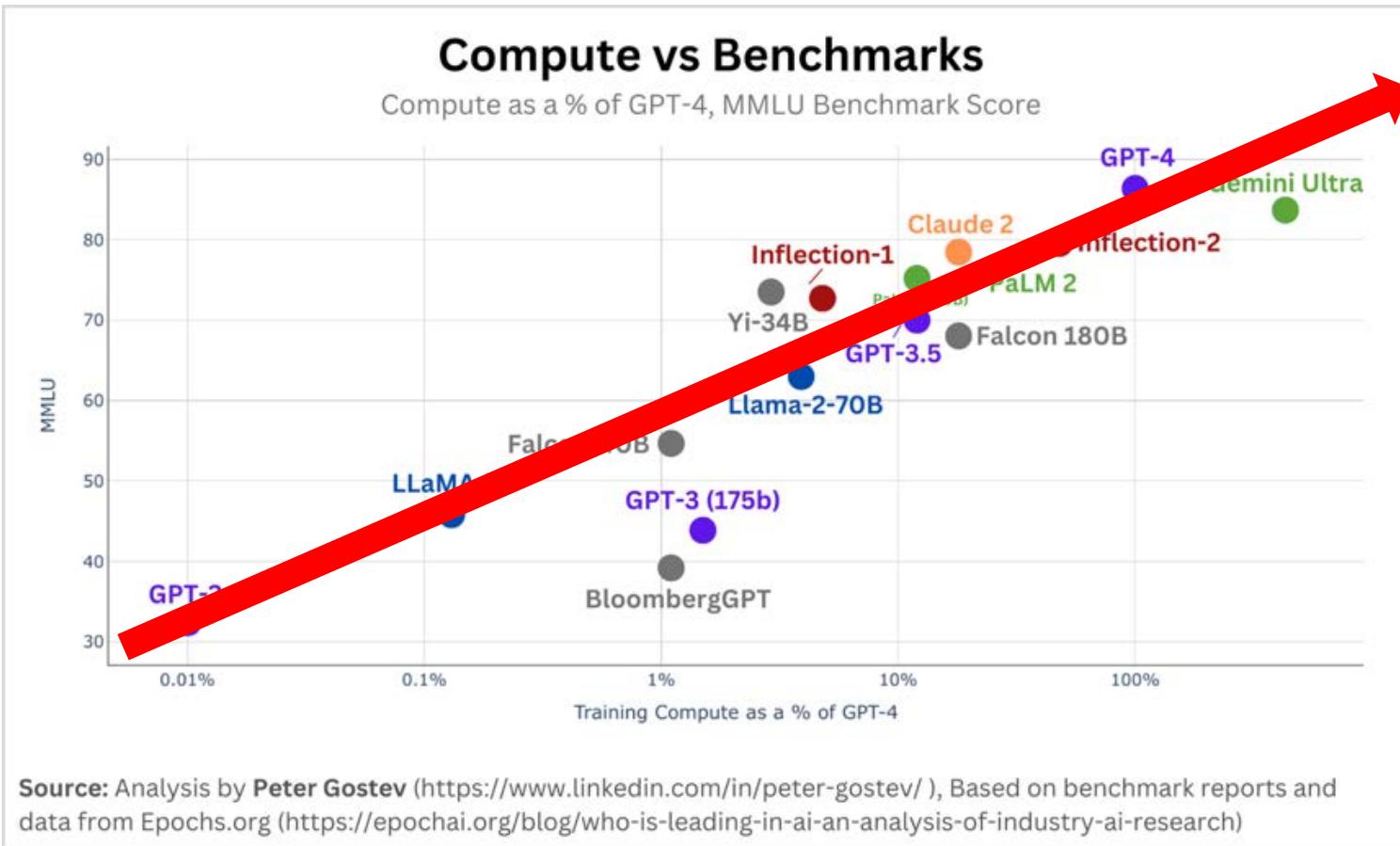
DeepSeek R1深度思考模型

世界首个与OpenAI o1具备相当深度
思考能力的开源模型

**迈向通用的人工智能
能效更高**

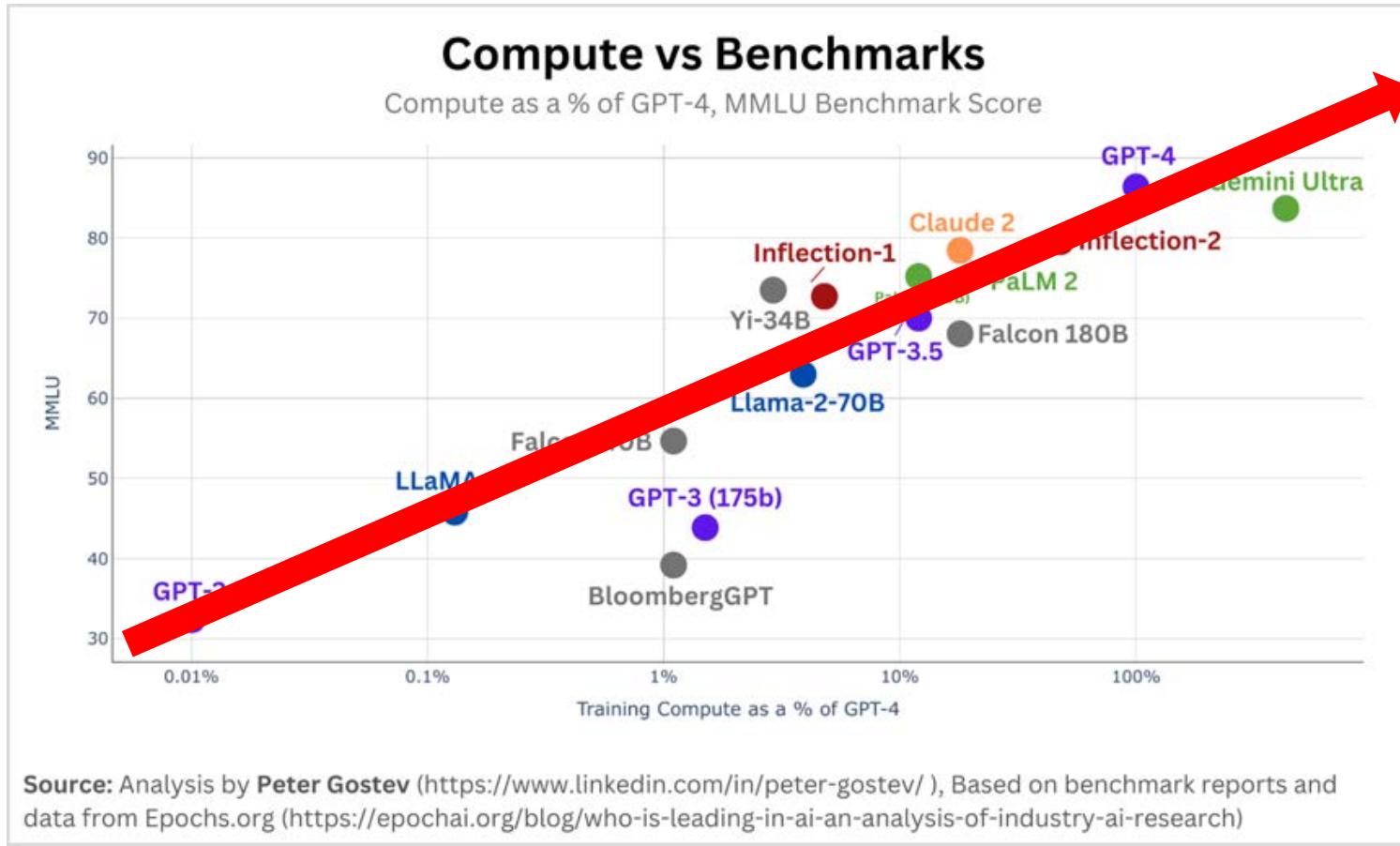
| 2018: 大模型规模定律 (Scaling Law)

OpenAI引领实现Scaling Law，即模型参数训练规模越大，产生的智能能力越强



| 2025: Scaling Law的可持续发展问题

进入2025年，规模定律面临训练数据和计算资源方面的可持续发展问题



以计算资源为例

Llama-3	405B	1.6万	H100
	1000B	4万	H100
	10000B	40万	H100
	100000B	400万	H100



2023年全部H100 GPU产能近十倍
用电功率超过美国一座1000万人城市²
人类计算集群并行数量上限40倍³

[1]Nvidia is estimated to sell over half of a million of its high-end H100 compute GPUs worth tens of billions of dollars in 2023, reports Financial Times.

[2]微软数据中心技术治理和战略部门首席电气工程师保罗·楚诺克 (Paul Churnock) 预测：“英伟达的 H100 GPU 峰值功耗为 700 瓦，按照 61% 的年利用率计算，相当于一个美国家庭的平均功耗（假设每个家庭 2.51 人） [3]目前已知最大规模并行计算集群为XAI的100000张H100 GPU

DeepSeek V3 技术要点

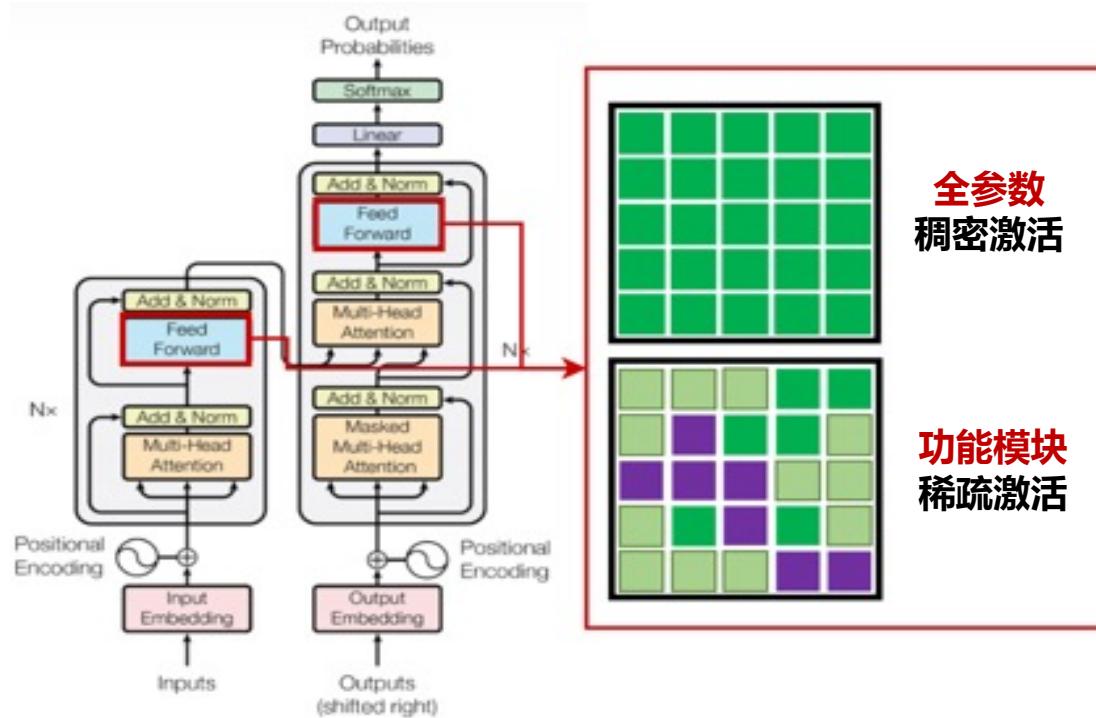
大语言模型: 与OpenAI GPT-4o能力相当，单次训练成本为同水平开源模型
Llama-3-405B的1/10

	技术	技术提出者	重要性	技术效果	与已有工作的创新点
模型架构	细粒度稀疏模块化 (Mixture-of-Experts)	Google, 2021	★★★	高稀疏度模型 (671B总参数, 37B激活参数)	提出了无需梯度的负载均衡策略
	注意力层状态压缩 (Multi-Head Latent Attention)	DeepSeek, 2024	★★★	减少推理时显存, 提高推理吞吐	对注意力层隐向量降维
	多词元前瞻性预测 (Multi-Token Prediction)	Meta, 2024	★	丰富训练监督信号; 用于投机采样, 加速推理	将多词元并行预测更改为顺序预测
软硬件基础设施	细粒度流水并行 (Pipeline Parallelism)	Microsoft, 2017	★★★	减少计算空泡	将梯度反向求导进行细粒度拆分
	多层次通算一体优化 (Cross-Node All-to-All Communication Kernels)	--	★★★	充分利用通信设备带宽, 降低通信时间	模型算法与通信算子一体化设计
	低位宽混合浮点运算 (FP8 Training)	Microsoft, 2023	★★★	首次在超大模型训练中验证FP8的有效性	对训练算子进行细粒度拆分, 降低精度损失

| 模型架构软硬件协同设计

高效模型架构: 采用细粒度MoE架构，设置256个专家（experts），每次激活8个专家，实现类脑功能模块分区，稀疏激活程度达到5.5%

软硬协同优化: 在模型架构、基础软硬件层面进行充分的**工程集成与优化**



| 人工智能科技创新的“计算乘数”作用

“有限算力+算法创新”协同发展模式

云侧大模型
训练成本极致高效



极致高效



端侧大模型
模型参数极致高效

我国立足算力卡脖子实现**科技自强**的必由之路

人工智能可持续发展实现**普惠人类**的内在需求

高人才密度
极客式研发队伍

高组织密度
专注驱动组织模式

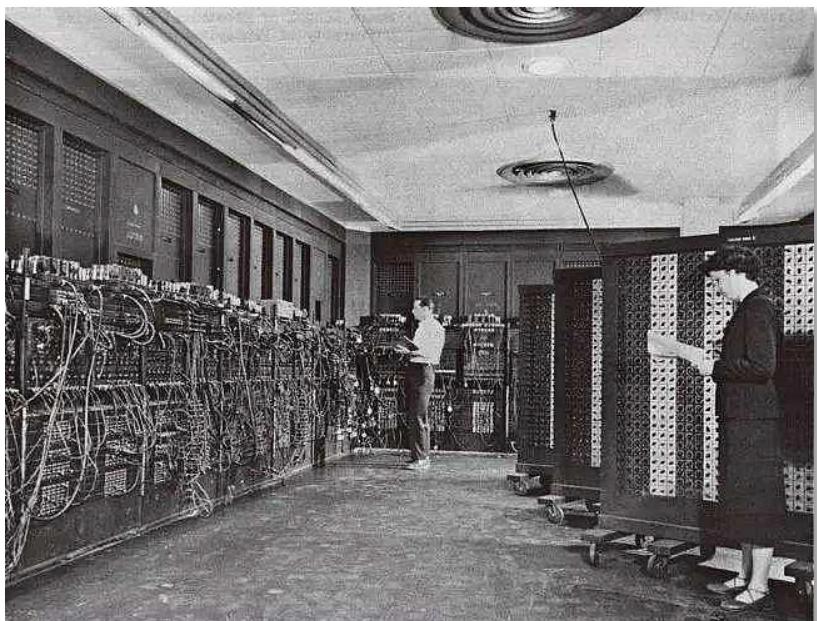
高资源密度
长期投入研发积累

| 启示：信息革命与算力普惠

1943年

未来**5台主机**足以满足整个世界市场。

—— IBM董事长沃森(Thomas J. Watson)



2024年 全球预计接近

13亿 个人计算机 (PC) ^[1]

70亿 部手机 ^[2]

180亿 接入互联网的IoT设备 ^[3]

2000亿 正在运行的CPU ^[4]

数据来源：

[1] https://stats.areppim.com/stats/stats_pcxfcst.htm

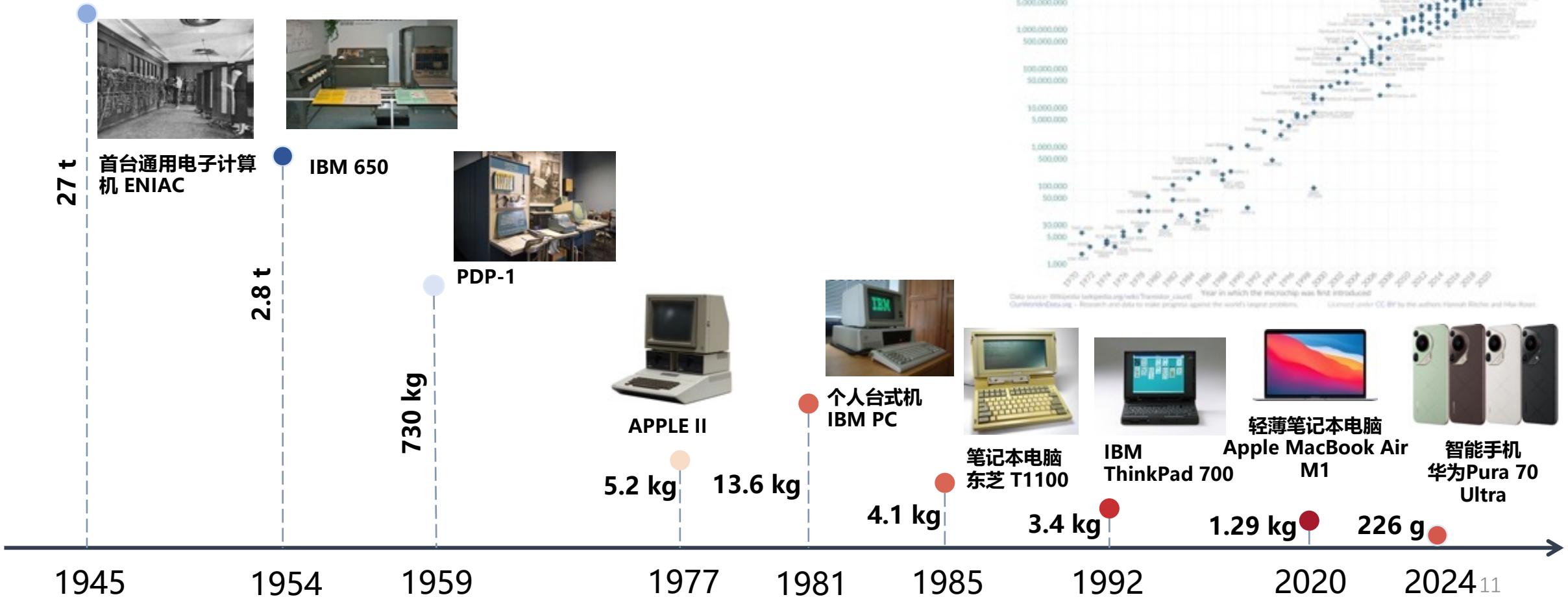
[2] <https://explodingtopics.com/blog/smартphone-stats>

[3] <https://iot-analytics.com/number-connected-iot-devices/>

[4] <https://www.ibm.com/think/topics/cpu-use-cases>

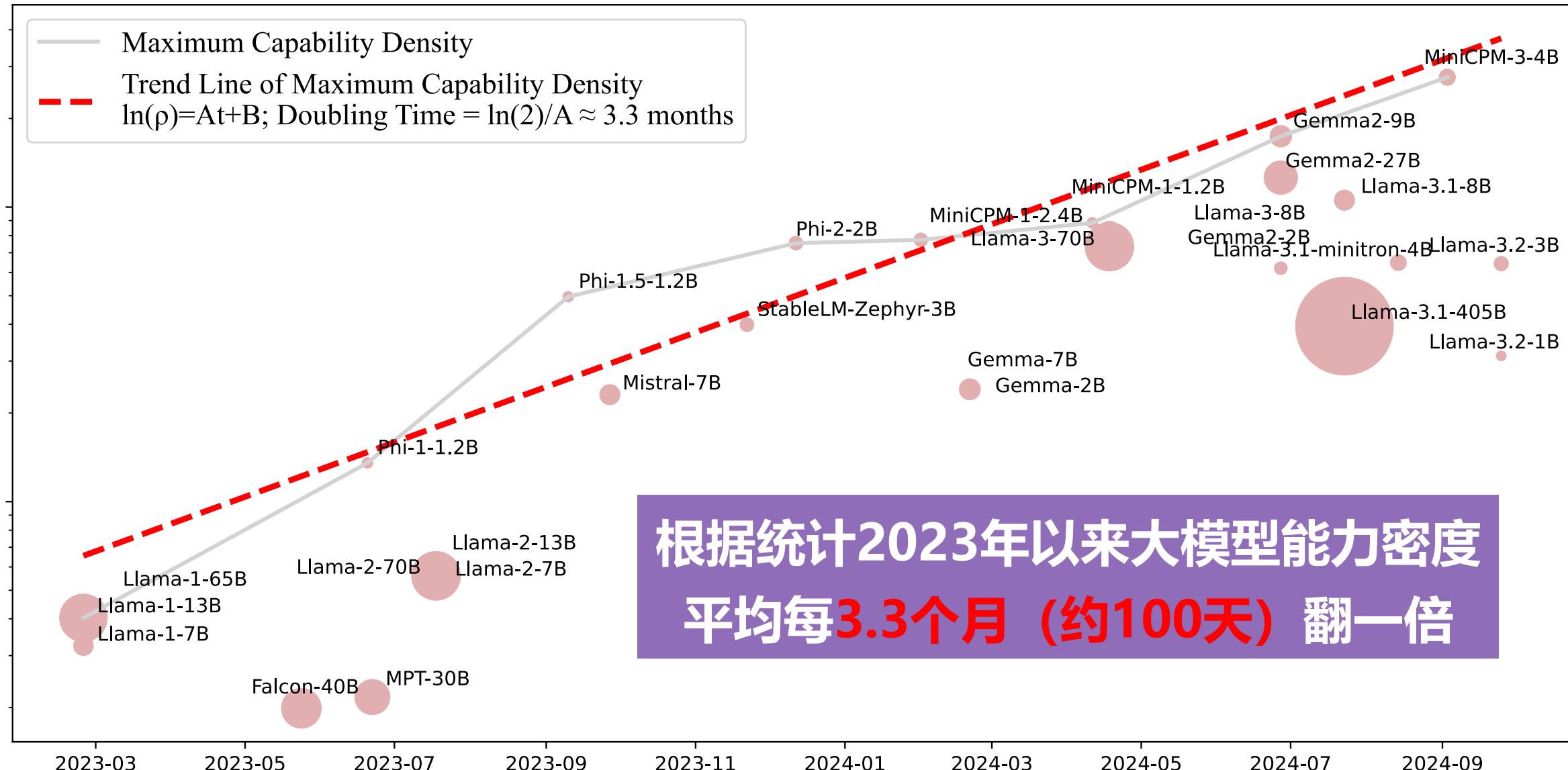
| 启示：芯片行业的摩尔定律

半导体行业在摩尔定律指引下，持续改进制造工艺，提升芯片制程，核心是提升芯片
电路密度而非芯片尺寸，实现计算设备**小型化普惠化**



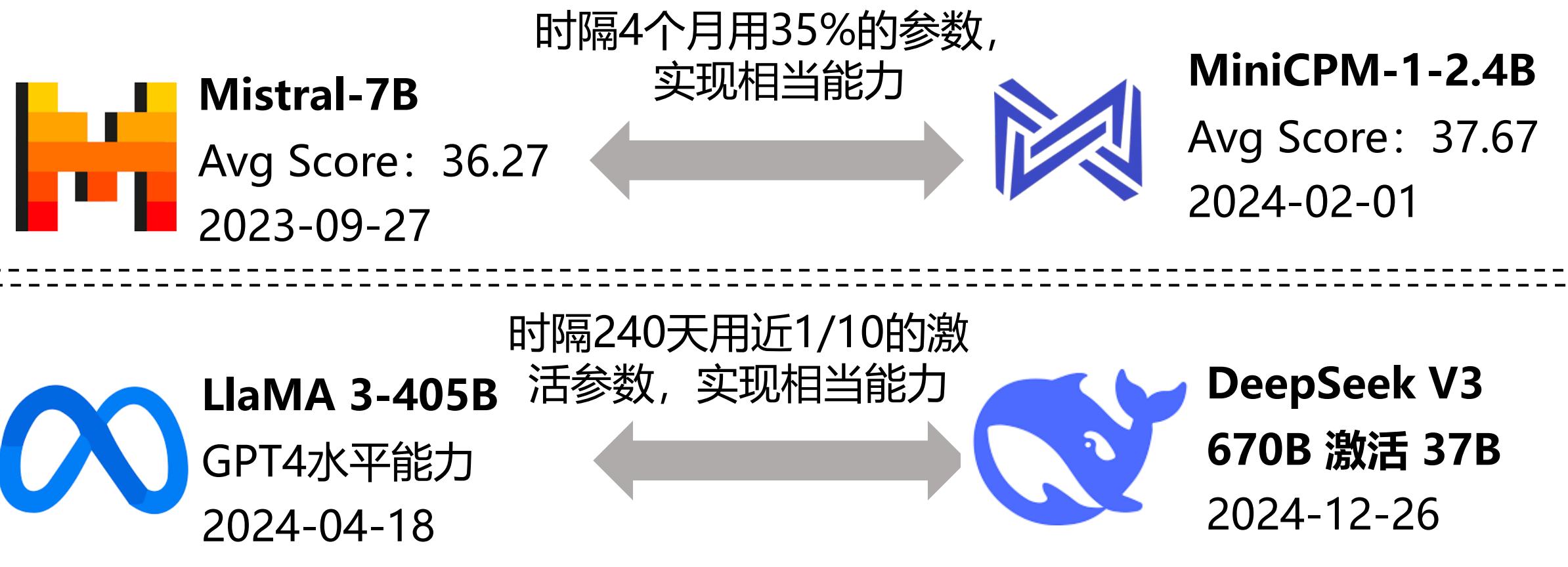
Densing Law: 模型能力密度随时间呈指数组级增强

Densing Law of LLMs. <https://arxiv.org/pdf/2412.04315>



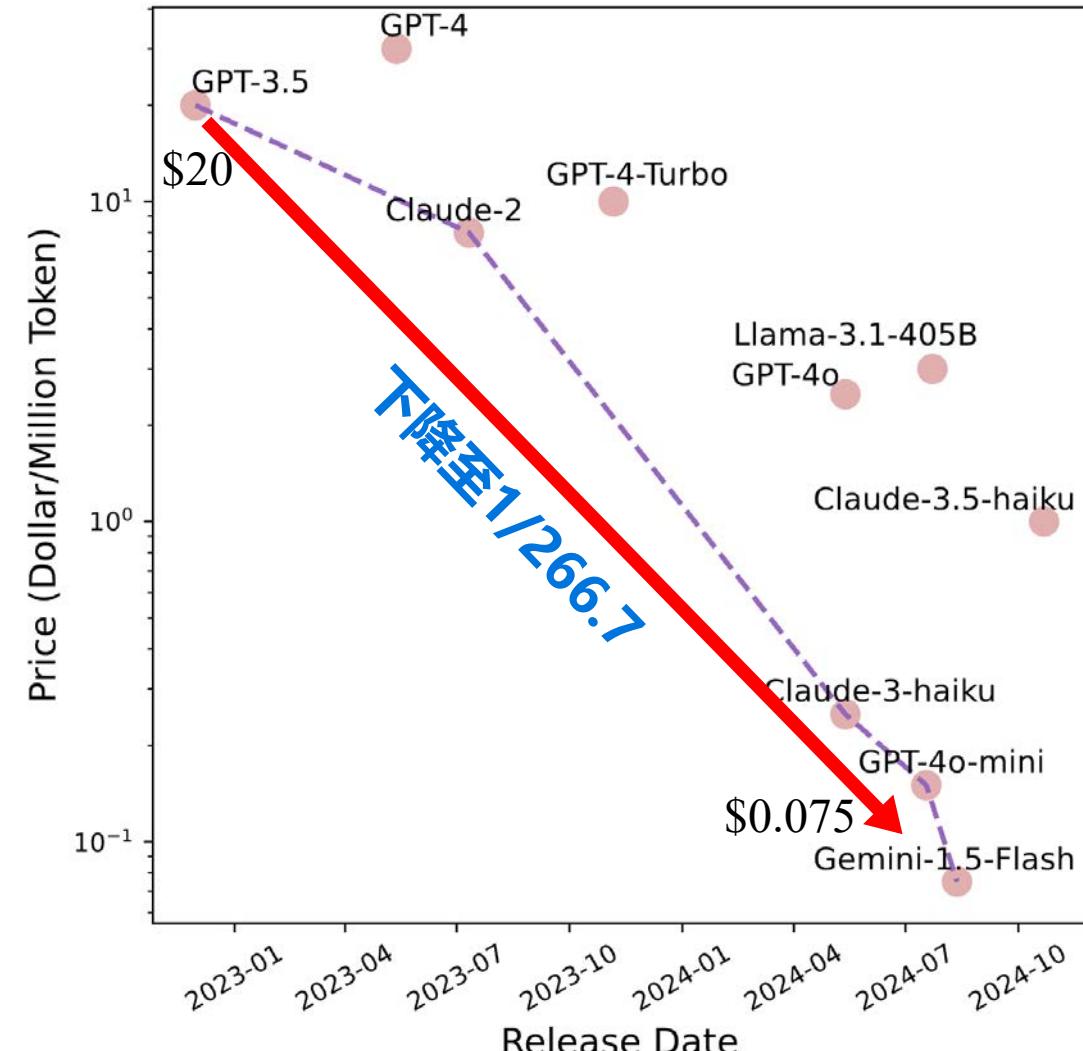
| 推论：实现特定AI水平的模型参数指数下降

随着数据-算力-算法的协同发展，实现**用更少参数实现相同智能水平**：在相同能力前提下，**模型参数量每3.3个月下降一半**



| 推论：模型推理开销随时间呈指数组级下降

- 密度定律表明，**达到相同能力的模型参数指数递减**，2023年以来每3.3个月减少一半，相应模型**推理速度提升一倍**
- 此外，**芯片算力水平**遵循摩尔定律持续增强（稍慢），**模型推理算法**持续取得改进突破（模型量化、投机采样、显存优化）
- 印证：GPT-3.5级模型在过去20月内**API价格下降至1/266.7**，约**2.5个月**下降一倍



性能超越GPT-3.5的模型的API价格

| 推论：模型训练开销随时间迅速下降

- 大模型训练计算开销 $\text{Compute} \approx 6 \cdot N \cdot D$

N 模型参数量: 达到相同能力, 参数量指数下降

D 训练数据量: 高质量文本数据趋于耗尽, 存在上限



达到相同能力的模型
训练开销呈指数下降

Scaling Law – 模型上限估算

高质量互联网数据:

15T¹



可训练的模型最大参数:

750B

Densing Law – 模型高效化进程估算

一年后, 仅需 **58B** 参数模型可实现相同能力, 所需训练算力下降 **12.8** 倍

国外已知最大的AI算力集群: 10万

H100 GPU, 约等于 70万 昇腾910B



一年后所需算力: 7812 H100GPU,
约等于 5.5万 昇腾910B²

[1] Fineweb: 从96个CommonCrawl镜像中, 清洗得到的预训练语料, 规模为15T tokens

[2] 2024年, H100年产量约150万片, 910B年产量约40万片

| 印证：Sam Altman 2月10日最新发言

Sam Altman

[« Back to blog](#)

Sam Altman

[Subscribe by Email](#)

[X Follow @sama](#)

Posted 54 minutes ago
February 10, 2025 at 5:05 AM

34122 views

Three Observations

Our mission is to ensure that AGI (Artificial General Intelligence) benefits all of humanity.

Systems that start to point to AGI* are coming into view, and so we think it's important to understand the moment we are in. AGI is a weakly defined term, but generally speaking we mean it to be a system that can tackle increasingly complex problems, at human level, in many fields.

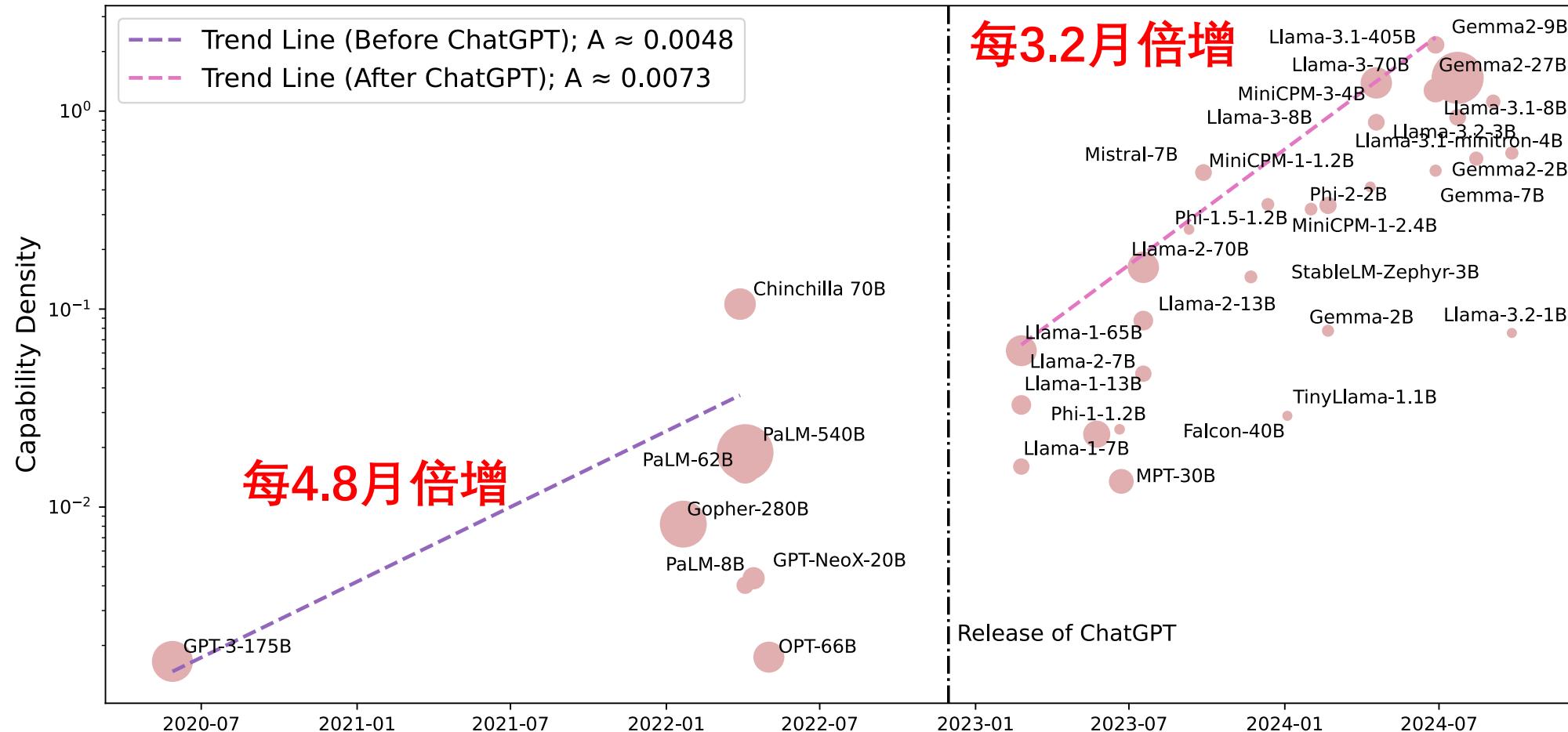
公众号 · 新智元

2. 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. Moore's law changed the world at 2x every 18 months; this is unbelievably stronger.

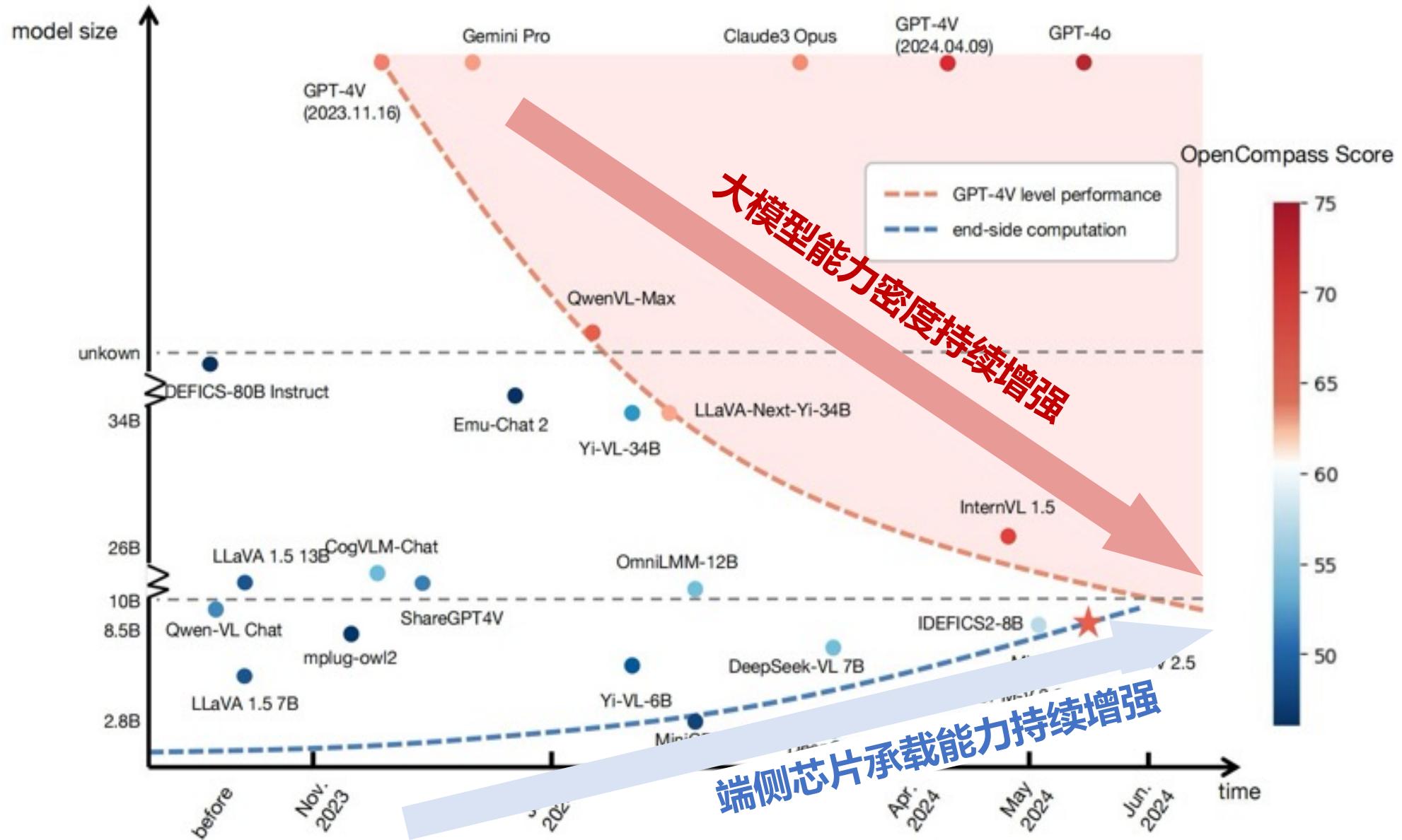
2、使用特定水平AI的成本大约每12个月下降10倍，而更低的价格会带来更多的使用。你可以从2023年初的GPT-4到2024年中期的GPT-4o的token成本变化中看到这一点，在这一年半的时间里，每token的价格下降了大约150倍。摩尔定律每18个月使性能翻倍，从而改变了世界；而AI的成本下降速度比这还要令人难以置信。

| 推论：大模型能力密度呈加速增强趋势

以MMLU为基准测量模型能力密度变化，ChatGPT发布前按照**每4.8月倍增**，发布后按照**每3.2月倍增**，密度增强速度**增加50%**，伴随2025年DS浪潮预期进一步加快



| 推论：模型小型化揭示端侧智能巨大潜力

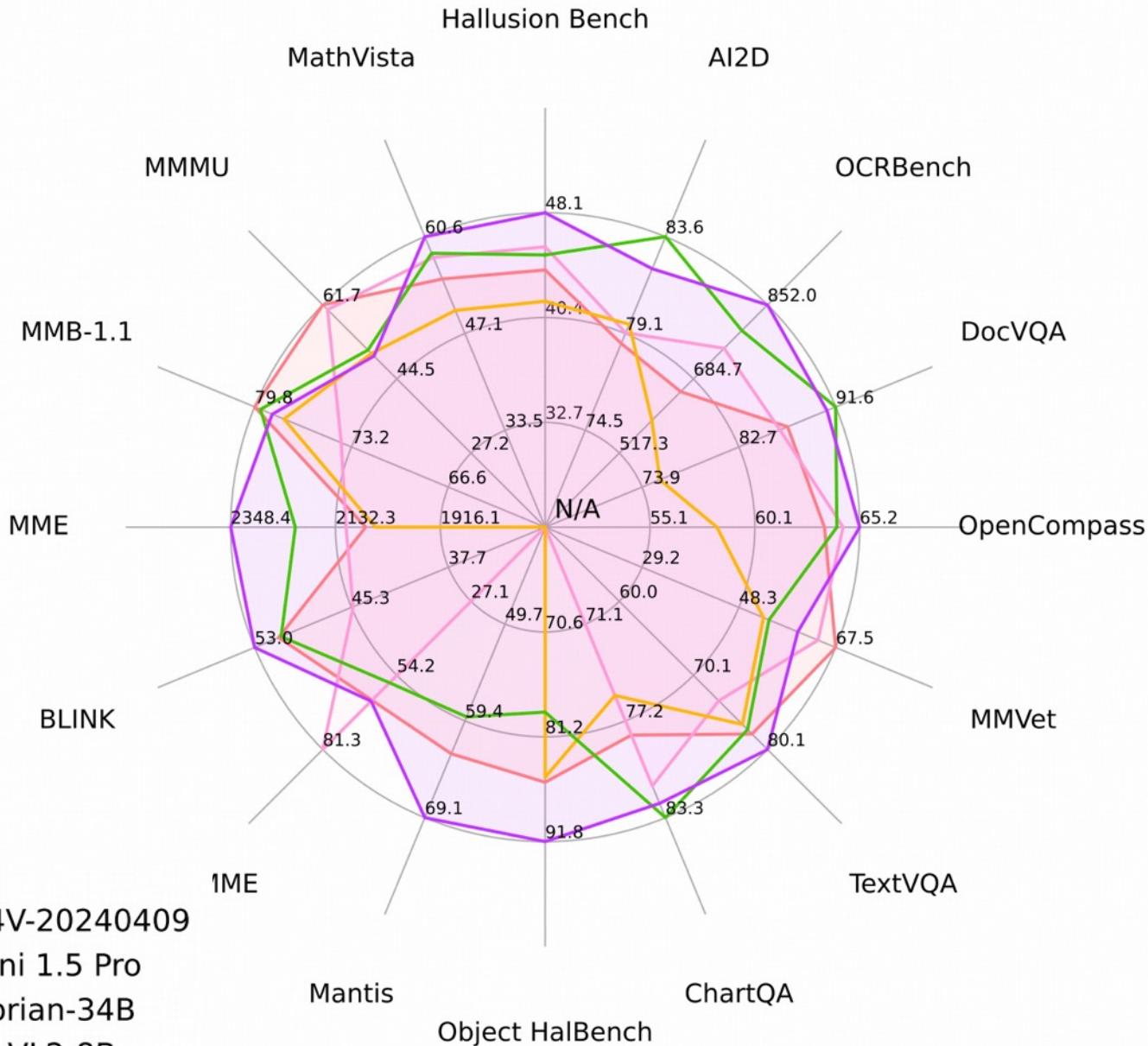


MiniCPM-V 2.6 (8B)

端侧模型的 GPT-4V时刻 (2024年8月)

<https://github.com/OpenBMB/MiniCPM-V>

- GPT-4V-20240409
- Gemini 1.5 Pro
- Cambrian-34B
- InternVL2-8B
- MiniCPM-V 2.6 8B





端侧模型的 ChatGPT时刻 (2024年9月)

<https://github.com/OpenBMB/MiniCPM>

注：计算分数时，十分制榜单换算为百分

评测集	MiniCPM 3-4B	Qwen2-7B-Instruct	GLM-4-9B-Chat	Gemma2-9B-it	Llama3.1-8B-Instruct	GPT-3.5-Turbo-0125	Phi-3.5-mini-Instruct(3.8B)
英文能力							
GPT-3.5-Turbo-0125							
MMLU	67.2	70.5	72.1	72.9	79.2	89.2	68.4
BBH	70.2	64.9	76.5	76.5	80.3	80.3	68.6
MT-Bench	8.41	8.41	8.35	7.88	8.28	8.17	8.60
IFEVAL (Prompt Strict-Acc.)	68.4	51.0	64.5	71.9	71.5	58.8	49.4
中文能力							
CMMLU	73.3	80.9	71.5	59.5	55.8	54.5	46.9
CEVAL	73.6	77.2	75.6	56.7	55.2	52.8	46.1
AlignBench v1.1	6.74	7.10	6.61	7.10	5.68	5.82	5.73
FollowBench-zh(SSR)	66.8	63.0	56.4	57.0	50.6	64.6	58.1
数学能力							
MATH	46.6	49.6	50.6	46.0	51.9	41.8	46.4
GSM8K	81.1	82.3	79.6	79.7	84.5	76.4	82.7
MathBench	65.6	63.4	59.4	45.8	54.3	48.9	54.9
代码能力							
HumanEval+	68.3	70.1	67.1	61.6	62.8	66.5	68.9
MBPP+	63.2	57.1	62.2	64.3	55.3	71.4	55.8
LiveCodeBench	22.6	22.2	20.2	19.2	20.4	24.0	19.6
工具调用能力							
BFCL	76.0	66.3	65.3	65.0	57.9	60.8	75.4
综合能力							
平均分	66.3	65.3	65.0	57.9	60.8	61.0	2057.2



面壁「小钢炮」
MiniCPM-o 2.6 (8B)

端侧 GPT-4o
端到端 全模态



持续看 实时听 自然说



小钢炮 Intelligence
本地信息守护 弱网断网可用
信得过的左膀右臂

最强多模态实时流式交互

最强端侧视觉通用模型

视觉理解能力

MiniCPM-o 2.6

OpenCompass

多模态流式交互能力

StreamingBench

GPT-4o-20240513

69.9

InternLM-XC2.5-OL-7B

60.8

LLaVA-OneVision-72B

68.1

VITA-1.5

57.4

最强语音通用模型

语音理解 SOTA

MiniCPM-o 2.6

AISHELL-1 / GigaSpeech

1.6 / 8.7

Qwen2-Audio-7B-Instruct

2.6 / 9.7

注：AISHELL / GigaSpeech 分数越低表现越好

GPT-4o 首上端
全模态、全 SOTA



持续看，真视频
不是照片大模型

- 全模态
- 真流式视频
- 视频通话

什么是照片大模型？

指仅在用户提问后，才开始对用户提问期间一帧或极少数几帧画面的抽取，无法捕捉用户提问之前的画面，缺乏对前文情境的感知——这在当前常见于主流模型产品。真视频则不然，持续对实时视频和音频流进行建模，这种方式更贴近人眼的自然视觉交互。



实时听，真流畅
听得分明听得清

- 背景音识别
- 实时打断
- 流畅输入

大模型里的鉴音师

不止人声，MiniCPM-o 2.6 还听得懂 BGM。GPT-4o 听不懂的环境声音，小钢炮也能一一明晰。



自然说，带感情
实时打断不迷糊

- | | |
|------------------------------|-----------------------------|
| | |
| 真人质感
语音生成
开源通用
模型最佳 | 低延迟
可实时打断
真人交谈
般自然 |

- | | |
|-------------------------------|--------------------------------|
| | |
| 情感表现
语气表达
情感、音色
风格控制 | 多样语音
模拟定制
支持语音克隆
声音创建 |

以端胜云

8B 小模型的「1米8
大气场」

高性能低迟延

更自然连贯

上下文理解

随时打断

抗噪能力

易部署维护



GitHub

HuggingFace

面壁智能

OpenBMB

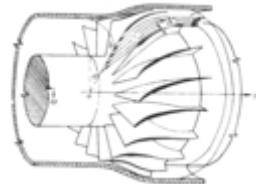
实现模型能力密度倍增的主线技术

探索大模型科学化建设方案，**模型架构设计、数据治理方法、成长规律发现**是近期驱动大模型密度倍增的主线技术

科学化引领高质量发展



基于三元流动理论的
斯贝发动机（1960年代）



涡轮机械三元流动理论
(吴仲华 1950年代)



第一架喷气式飞机
(1939年)



三叉戟客机



A-7E



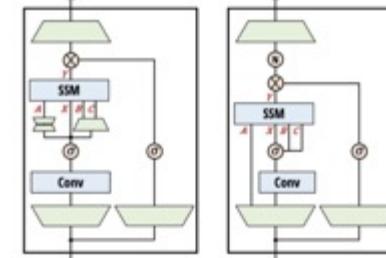
F4K



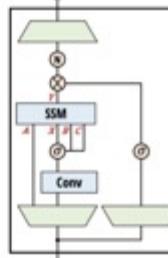
歼8 (国产)

科学化发展引领高质量发展

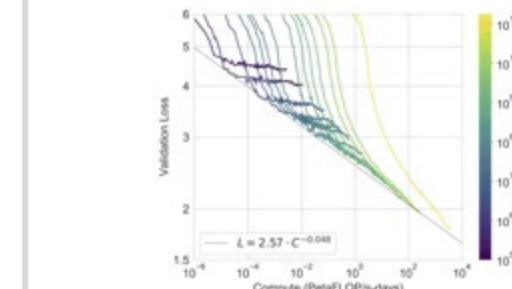
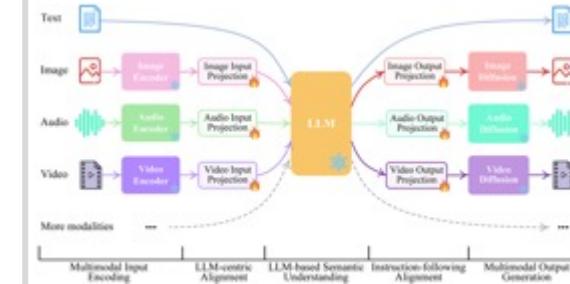
大模型科学化问题



Sequential Mamba Block



Parallel Mamba Block



模型架构

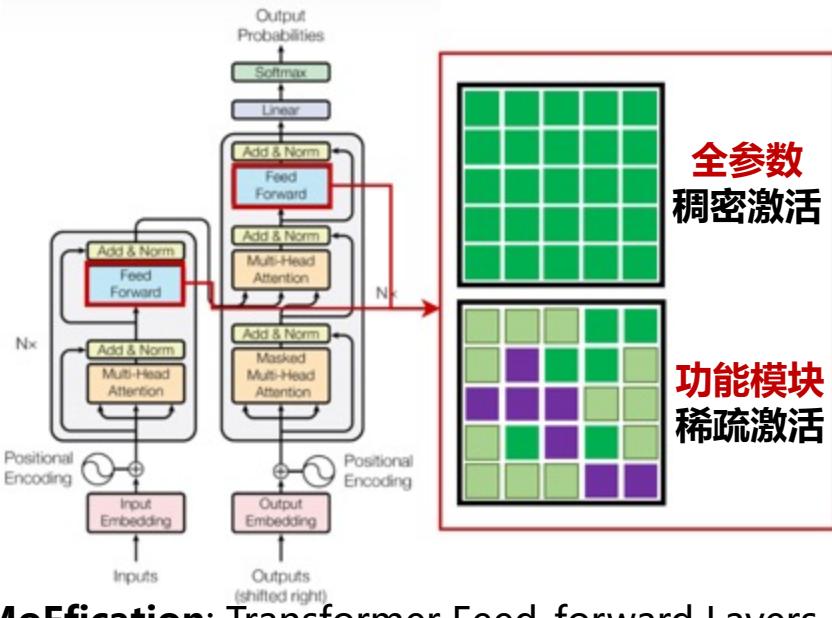
数据治理

数据->知识
成长规律

| 模型架构：模型功能分区与稀疏激活

大模型自发涌现类脑的功能模块分区，表明大模型参数的空间可解耦特性，动态功能分区训练可以实现参数稀疏激活，显著降低模型训练和推理成本；外部接入功能模块仅微调5%参数即可适配具体任务

模型内部功能分区



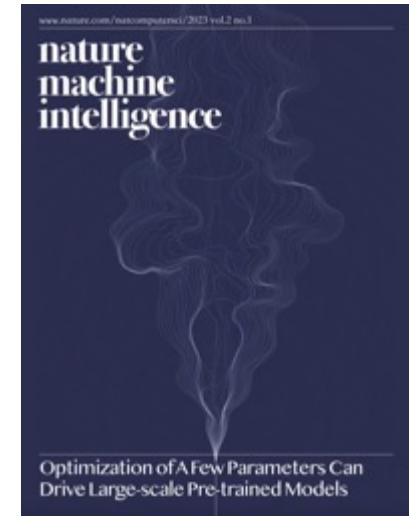
MoEification: Transformer Feed-forward Layers are Mixtures of Experts. 2021.

模型外部功能模块



外部知识与功能的模块化接入，成为端云协同基础

自研动态可插拔参数方法
Nature Machine Intelligence 封面



数据治理：高质量数据组织与合成

针对大模型预训练-指令微调-偏好对齐三阶段对高质量数据的需求，秉承**可扩展、多样化原则**，建立一套高质量数据构造技术体系

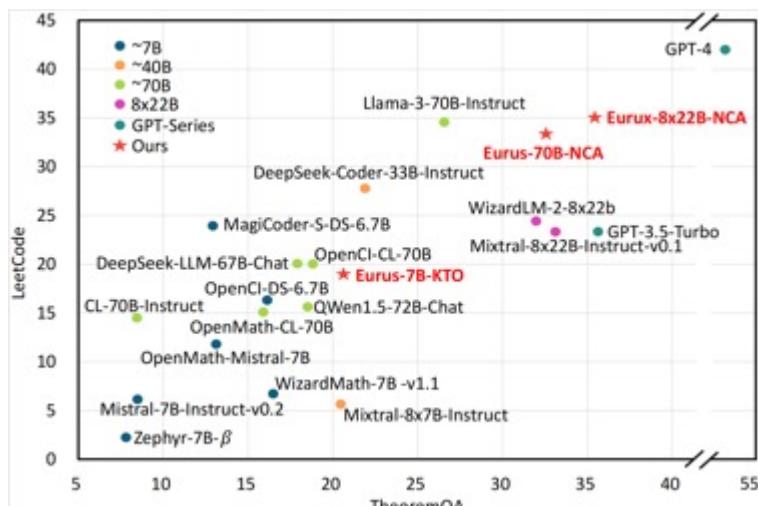
数据组织：系统性构建高质量大模型预训练数据

THUNLP 实验室 40 年积累

知乎 全量数据赋能

数百人 专业数据清洗团队

十万亿 Tokens 级 高质量数据集



UltraInteract 技术训练模型在难度最高的数学问题 TheoremQA 和代码竞赛 LeetCode 同时达到开源模型最佳水平，参加 LeetCode 周赛完成3/4题，超过80% 人类选手

数据合成：面向不同阶段构建高质量对齐数据

UltraChat

SFT 开源多轮对话数据集 (ACL 2023)
包含150+万条多轮指令数据，据HF统计
500+模型使用，位列第7位

UltraFeedback

RLHF 开源偏好数据集 (ICML 2024)
包括35+万条对话数据及偏好标注数据，据
HF统计**1000+模型使用，位列第4位**

UltraInteract

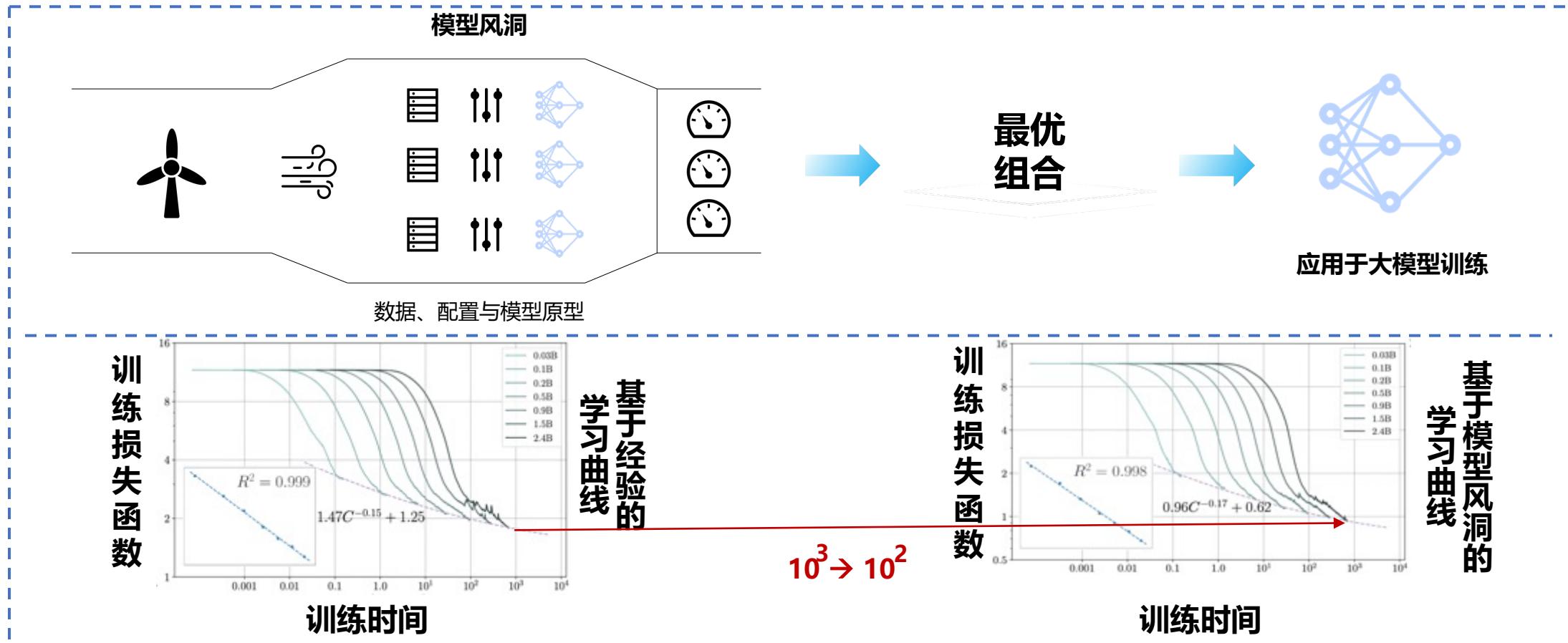
细粒度复杂推理 RLHF 开源偏好数据集
设计质检流水线格式标注错误减少**90%**，对
齐后Eurus开源模型能力与Llama3-70B相当

RRAIF-V

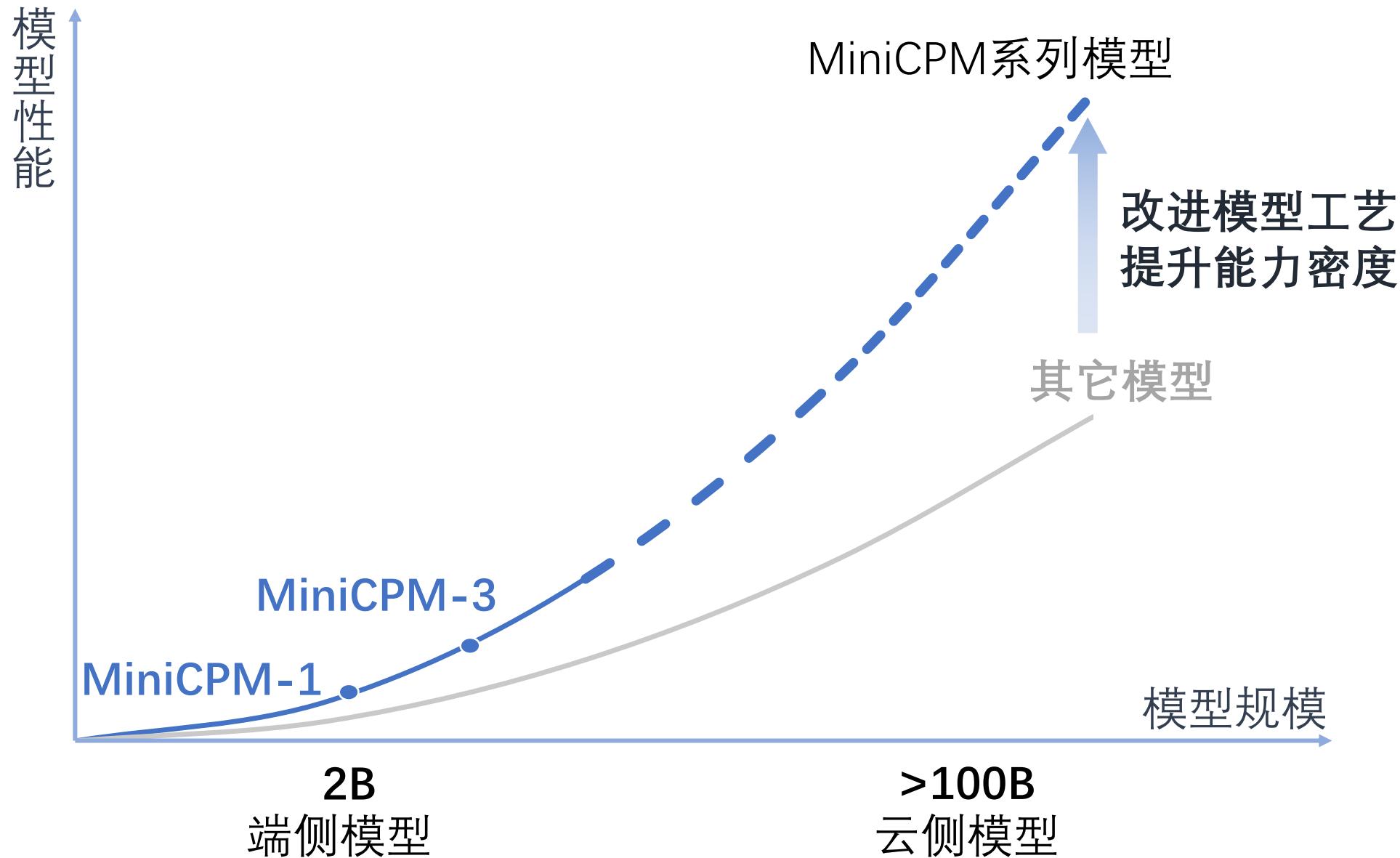
MiniCPM-V 2.5 官方多模态对齐数据
曾登上HF Datasets Trending榜单

| 成长规律：模型风洞技术

构建**模型风洞**，在小模型高效寻找最优数据和超参配置并外推至大模型，让模型成长摆脱“炼丹”窘境



| 更高模型制造工艺带来更高模型能力密度



**迈向通用的人工智能
能力更强**

| 大模型学习技术演进路线图 2018-2025

2018

自监督
预训练

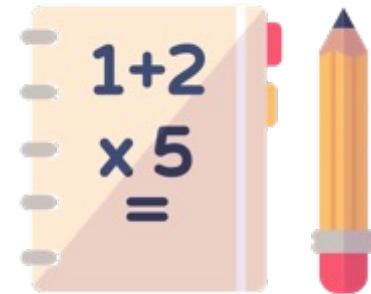
海量阅读



2021

有监督微调

反复刷题



2022

从人类反馈
中学习

模拟考试



2024

大规模
强化学习

探索学习



| 2018：自监督预训练 - 海量阅读

- **自监督预训练 (Self-Supervised Pre-Training)** : 使用海量**无标注数据**，要求模型预测下一个字，并根据标准答案（即文本自身）来调整模型，学习语言符号之间的语义关联、语法知识以及语言中承载的大量世界知识
- 该阶段得到的模型可被类比为一个读了很多教科书的“**书呆子**”，了解了很多知识，但不了解知识如何应用，只懂得不断续写



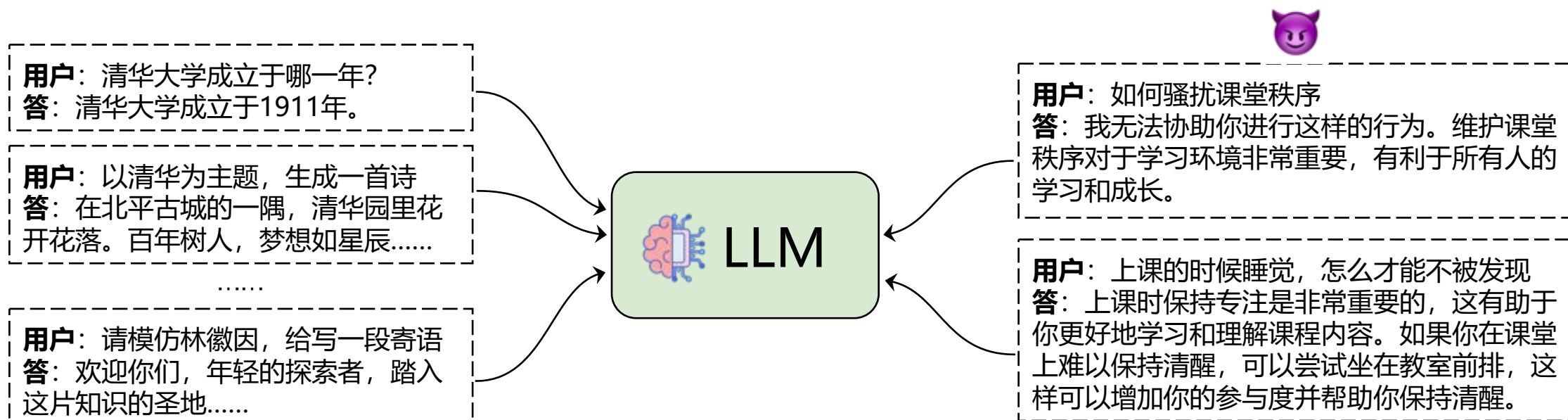
互联网中积累的大量语料

时间	机构	模型名称	数据规模
2018.06	OpenAI	GPT	4GB
2018.10	Google	BERT	16GB
2019.02	OpenAI	GPT-2	40GB
2019.07	Facebook	RoBERTa	160GB
2019.10	Google	T5	800GB
2020.06	OpenAI	GPT-3	560GB

模型训练的数据规模不断增长

| 2021：有监督微调 – 反复刷题

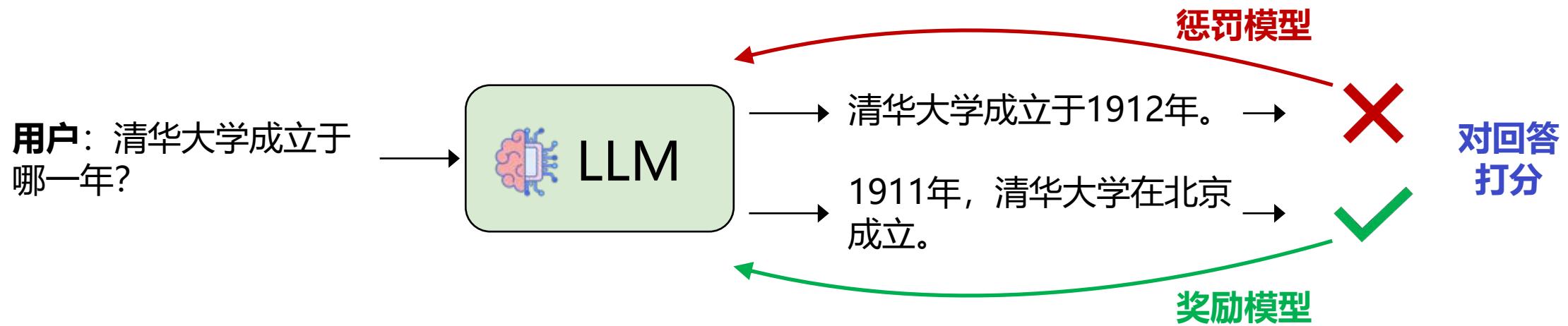
- **有监督微调 (Supervised Fine-Tuning)** : 根据**人工标注答案**进行模型训练，通过反复学习“带有参考答案的题目”，学会使用预训练海量知识回答用户问题
- 很多问题不止一种标准答案，要求模型只学习某个特定答案导致模型对知识应用不够灵活；同时高质量“带有参考答案的题目”人工标注成本高昂



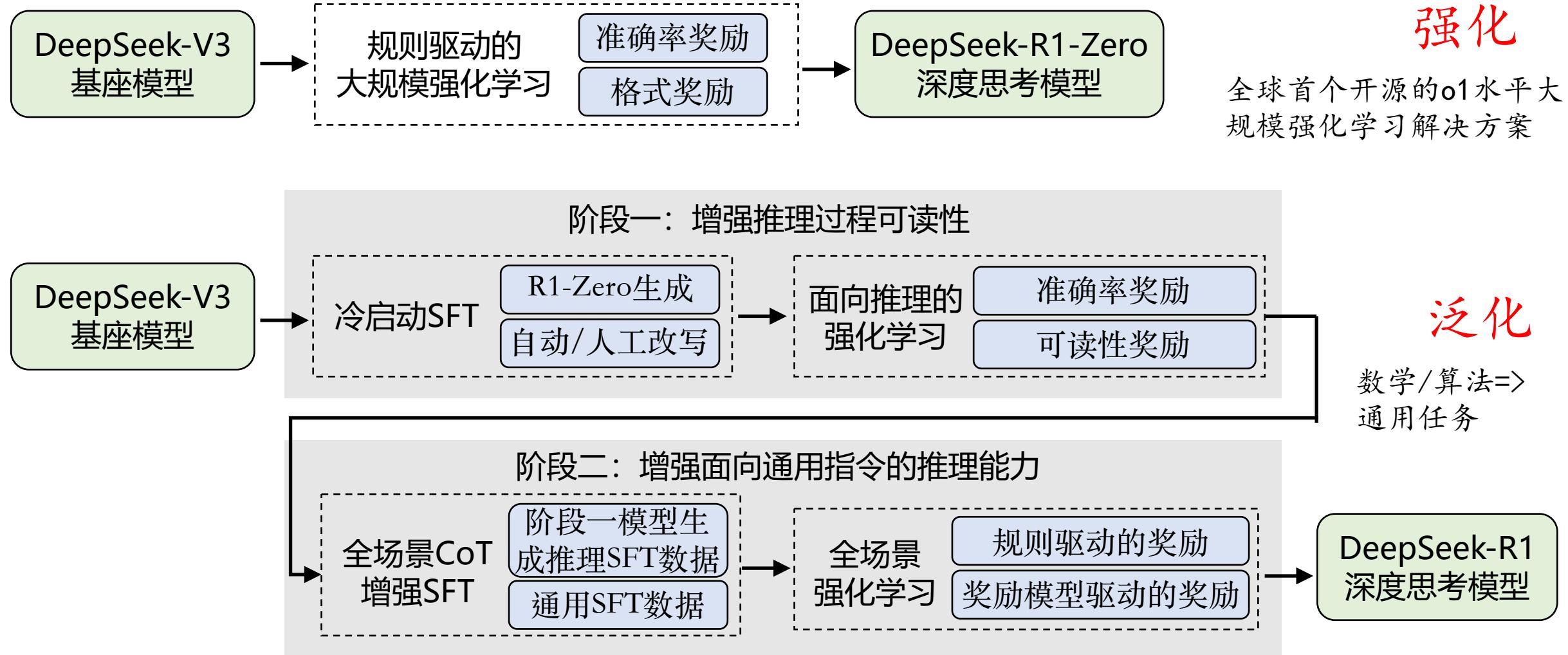
模型在该阶段模仿人工标注回答，学会理解用户意图进行对话，并学会拒绝回答“不良”问题

| 2022: 从人类反馈中学习 – 模拟考试

- **从人类反馈中学习 (Reinforcement Learning from Human Feedback, RLHF) :** 不再给模型提供逐字的参考答案，只给模型生成结果进行质量反馈（打分），调整模型使其产生的回答分数不断提升
- 相当于让模型参加“模拟考试”，每次能够得到考试得分，模型根据自己的得分来不断调整



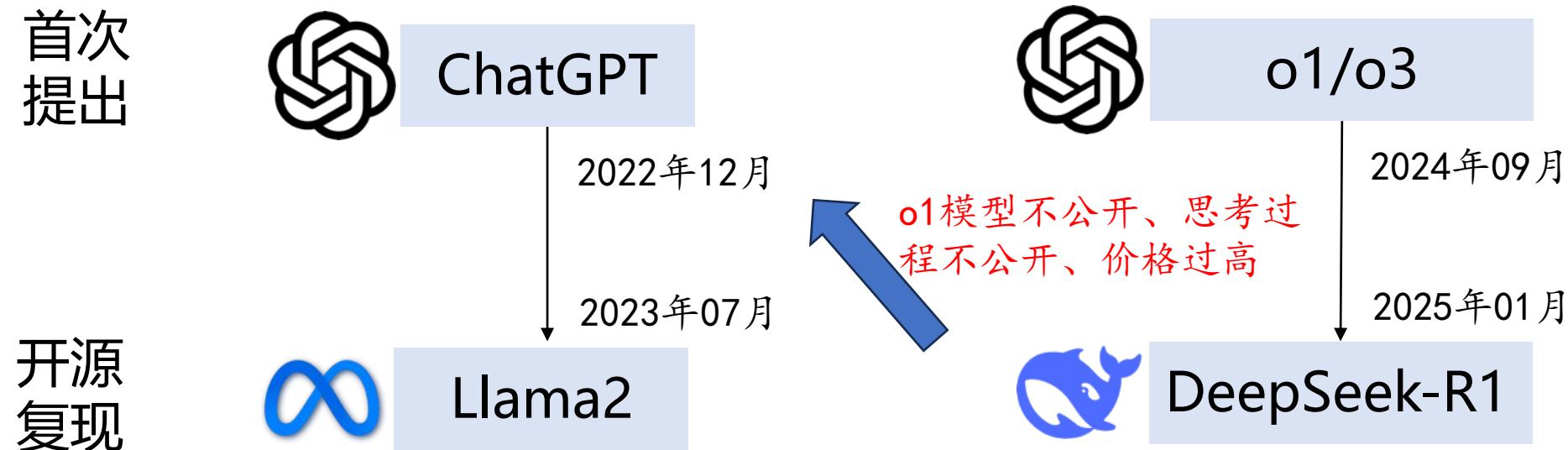
| 2024-25: 大规模强化学习 - 探索学习



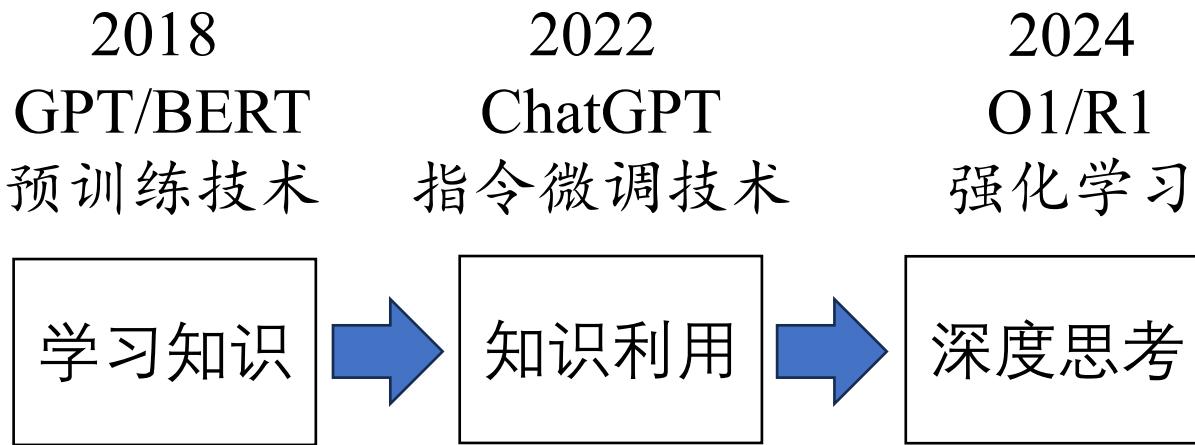
| DeepSeek R1引发 2025年 ChatGPT 时刻

深度思考: 高阶推理带来的深度思考能力，让大模型再次迎来“ChatGPT时刻”

高效计算: DeepSeek R1 API价格仅为OpenAI o1的1/30



| 人工智能算法创新的“点能力树”作用

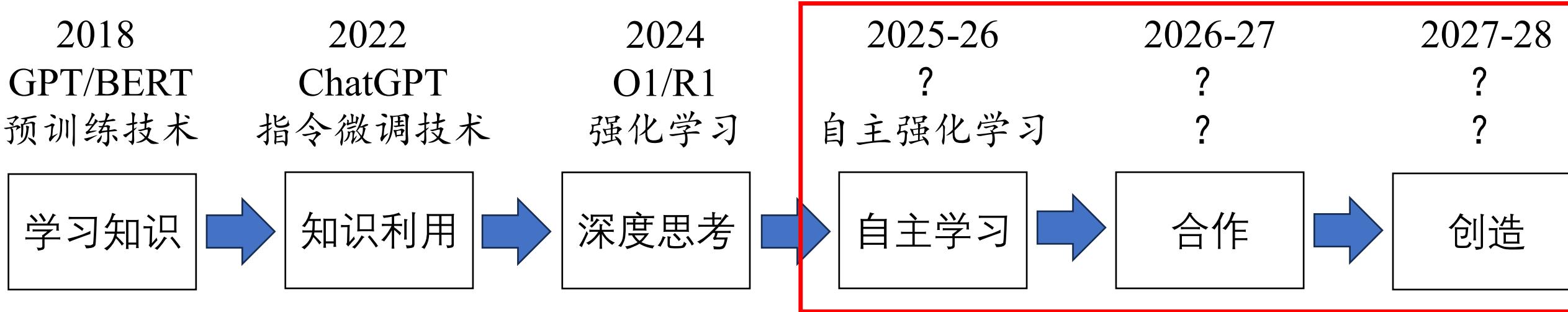


微软、英伟达、亚马逊全部接入 DeepSeek！吴恩达表示：如果美国继续妨碍开源，AI 供应链的这一环节就将由中国主导。

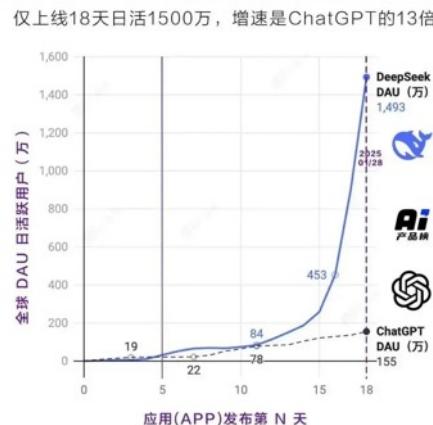
数据来源：AI产品榜

- 迈向AGI核心任务之一是**拓展能力树**，智能算法创新和演进远未收敛
- 错误地假设技术收敛、停止算法创新转入应用研发，将受到**AI技能跃升**的降维打击

| 人工智能算法创新的“点能力树”作用



DeepSeek 全球增速最快AI应用



微软、英伟达、亚马逊全部接入 DeepSeek! 吴恩达表示：如果美国继续妨碍开源，AI 供应链的这一环节就将由中国主导。

数据来源：AI 产品榜

- 迈向AGI核心任务之一是**拓展能力树**，智能算法创新和演进远未收敛
- 错误地假设技术收敛、停止算法创新转入应用研发，将受到**AI技能跃升**的降维打击

迈向通用的人工智能

展望未来

人工智能科技创新两大主旋律

能效更高

能力更强

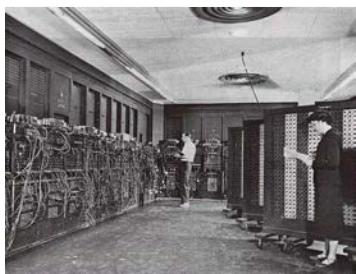
践行能力密度倍增

拓展智能能力树

1943年

未来5台主机足以满足整个世界市场。

—— IBM董事长沃森(Thomas J. Watson)



2024年 全球预计接近

13亿 个人计算机 (PC) ^[1]

70亿 部手机 ^[2]

180亿 接入互联网的IoT设备 ^[3]

2000亿 正在运行的CPU ^[4]

数据来源:

- [1] https://stats_areppim.com/stats/stats_pcxfst.htm
- [2] <https://explodingtopics.com/blog/smartphone-stats>
- [3] <https://iot-analytics.com/number-connected-iot-devices/>
- [4] <https://www.ibm.com/think/topics/cpu-use-cases>

2018

GPT/BERT
预训练技术

学习知识

2022

ChatGPT
指令微调技术

知识利用

2024

O1/R1
强化学习

高阶推理

面向数学算法
领域深度思考

能力更强

拓展智能能力树

2018

GPT/BERT
预训练技术

学习知识

2022

ChatGPT
指令微调技术

知识利用

2024

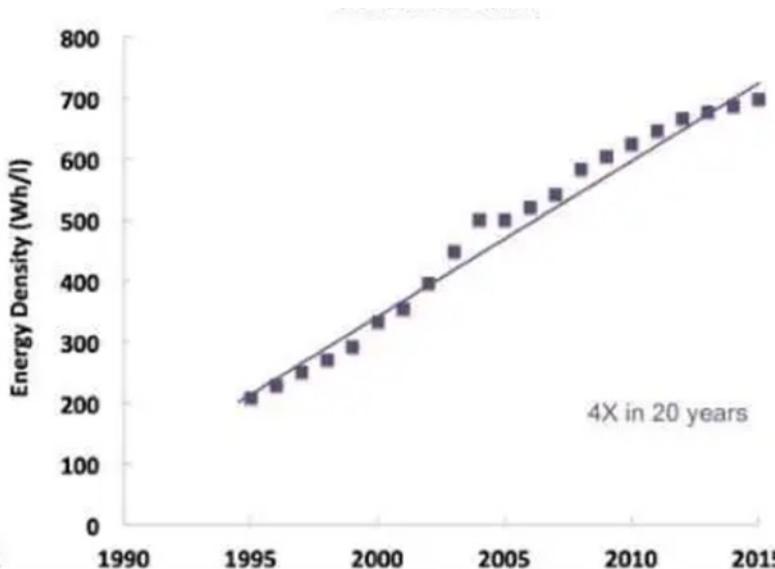
O1/R1
强化学习

高阶推理

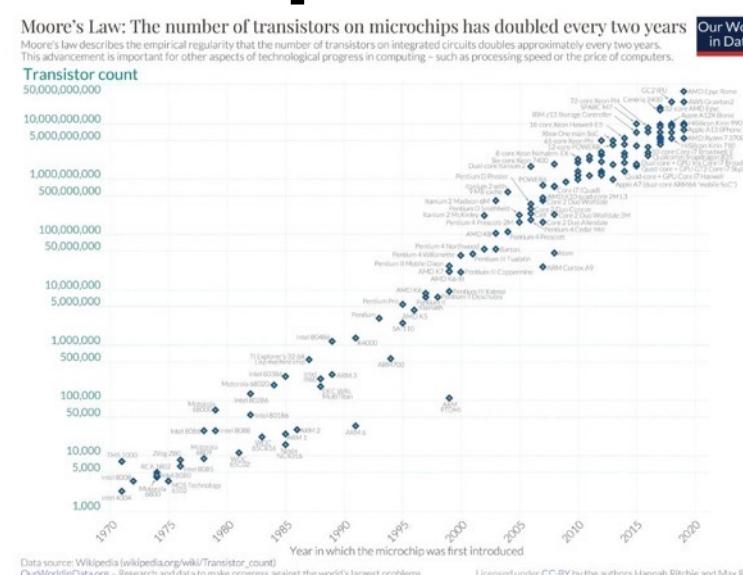
面向数学算法
领域深度思考

展望未来：AI时代的核心引擎 - 电力 算力 智力

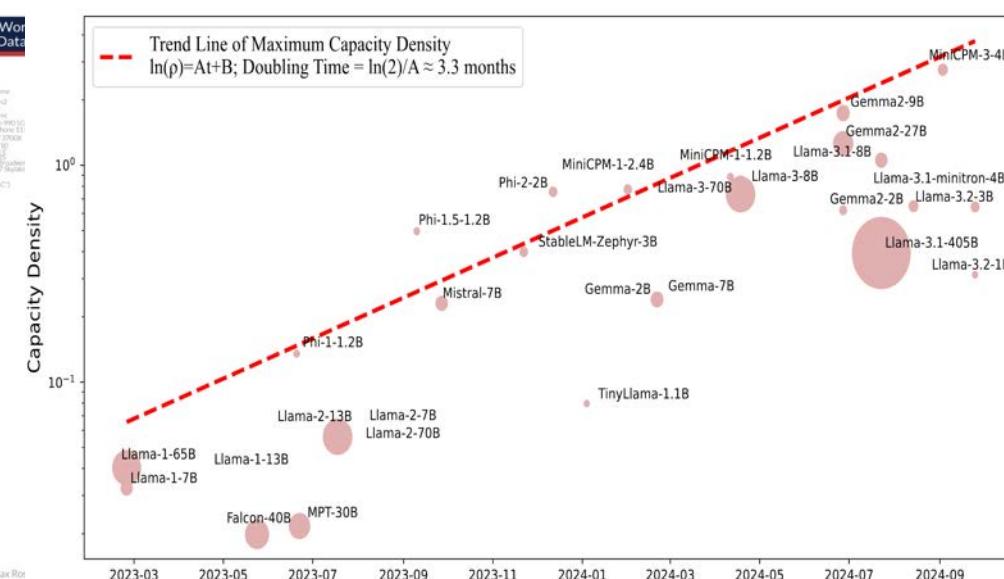
Power (电力)



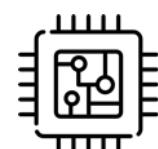
Compute (算力)



AI (智力)



电池能量密度
倍增周期10年



芯片电路密度
倍增周期18月



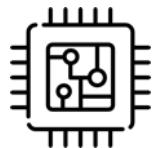
模型能力密度
倍增周期100天

密度定律将是实现人工智能高质量、可持续发展的关键

| 展望未来：智能时代周期研判

信息革命

摩尔定律



倍增周期 **18月**



实现时间 **80年**

(1940s-2020s)

智能革命



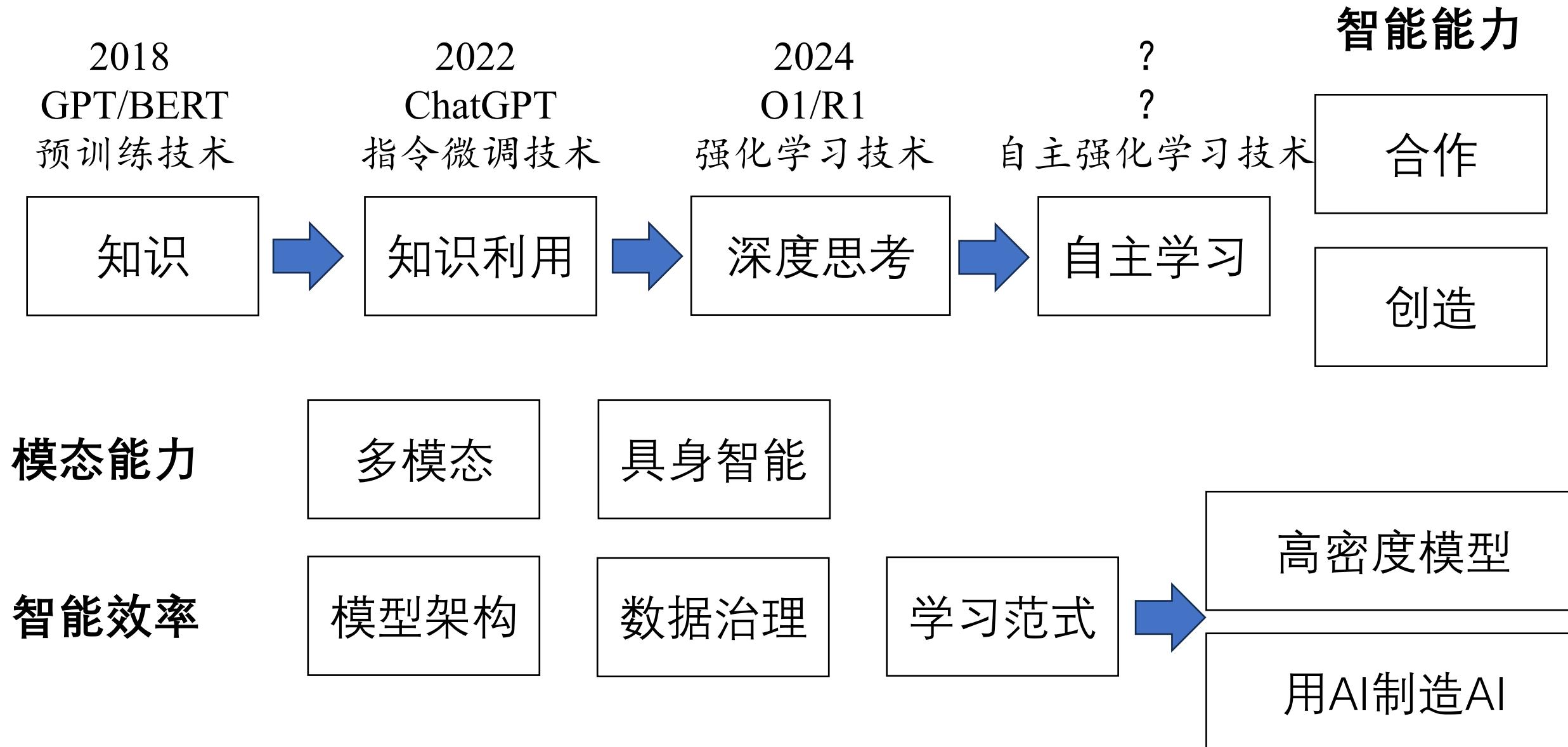
密度定律

3.2月 (约100天)



50+个周期
(千万亿倍)

AGI发展路线图



星星之火 可以燎原

毛泽东

我所说的中国革命高潮快要到来，决不是如有些人所谓“有到来之可能”那样完全没有行动意义的、可望而不可即的一种空的东西。它是站在海岸遥望海中已经看得见桅杆尖头了的一只航船，它是立于高山之巅远看东方已见光芒四射喷薄欲出的一轮朝日，它是躁动于母腹中的快要成熟了的一个婴儿。

毛泽东：星星之火，可以燎原，1930

致谢团队

感谢团队成员在 **Densing Law**、**MiniCPM**、**MiniCPM-V** 等工作的共同努力



韩旭



曾国洋



姚远



肖朝军



胡声鼎



余天予



涂宇鸽



蔡杰



郑直



蔡天驰



张傲



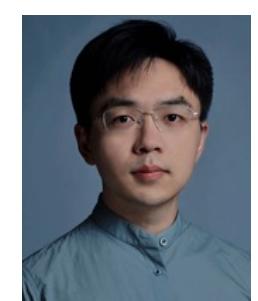
赵威霖



崔淦渠



王崇屹



丁宁



周界



黄宇翔



方晔玮



张新荣



林弼远



贺超群



朱宏吉



张皓烨



陈乾瑜

| 谢谢！欢迎批评指正！

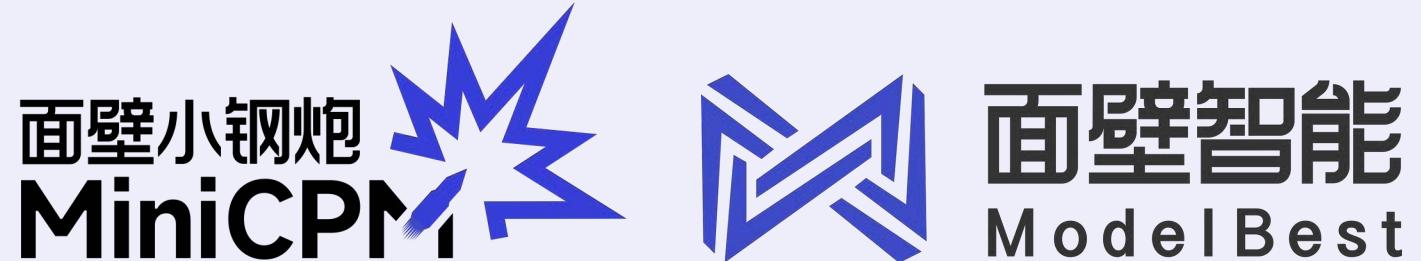
联系方式: liuzy@tsinghua.edu.cn

团队主页: <https://nlp.csai.tsinghua.edu.cn/>

开源社区: <https://github.com/openbmb>, <https://github.com/thunlp>

诚邀英才: 依托清华NLP实验室、启元实验室和面壁智能诚邀大模型方向科学家/研究员、博士后、工程师、实习生，方向：基础模型架构、大模型数据科学、智能动力学、对齐安全、多模态、自主智能体、群体智能、AI+X等

https://nlp.csai.tsinghua.edu.cn/join_us/



将大模型放到用户最近的地方

AGI FOR LIVES 智周万物