端侧通用人工智能大模型发展趋势及技术解析

面壁智能

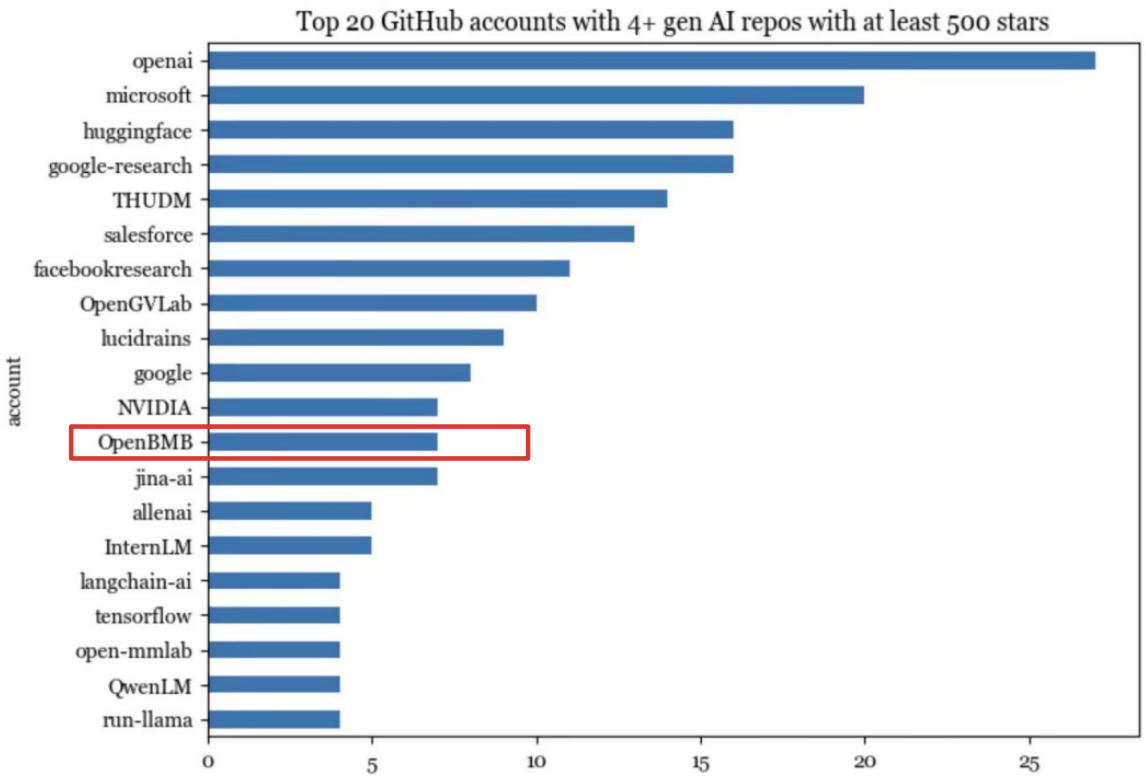
北京 • 2024

OpenBMB开源社区

OpenBMB开源社区由面壁智能（ModelBest）和清华大学自然语言处理实验室（THUNLP）共同支持发起，旨在打造大规模预训练语言模型库与相关工具，加速百亿级以上大模型的训练、微调与推理，实现大模型的标准化、普及化和实用化。

OpenBMB 开源影响力

跻身全球Github社区前列



GitHub Hugging Face ModelScope 魔搭社区 始智AI wisemodel.cn

OpenBMB 能力体系

自主智能体和群体智能

TooLLM　XAgent　ProAgent　ChatDev　AgentVerse　IOA　…

面壁MiniCPM旗舰端侧大模型系列

MiniCPM基座模型　MiniCPM- V多模态模型　长文本与MOE模型

CPM 系列通用大模型

CPM-1（2020年） → CPM-2（2021年） → CPM-3（2022年） → CPM-Ant（2022年） → CPM-Bee（2023年） → VisCPM 多模态 → CPM-Cricket（2023年11月）

ModelForce 全流程优化加速平台

高效训练 BMTrain　高效压缩 BMCook　高效推理 BMInf　高效微调 OpenDelta、OpenPrompt、UltraChat

社区口碑项目 —— MiniCPM系列模型

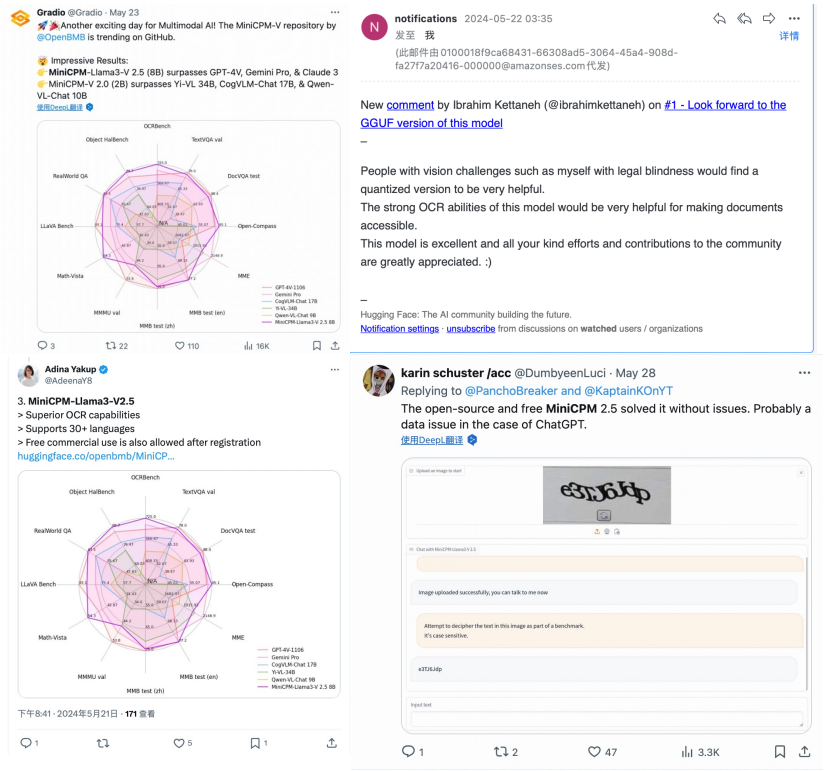
Github Star 12,352 | 项目下载量100万

开源社区好评

多次登顶GitHub Trending

跻身HuggingFace 50万模型TOP1

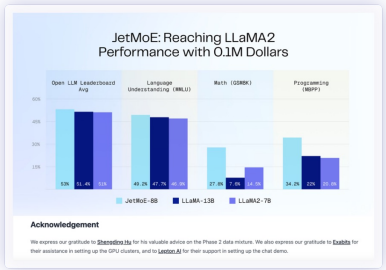
有海外开发者表示，MiniCPM-Llama3-V 2.5 对有视障人士非常有帮助，我们因 AI 向善而备受鼓舞！



影响了一些有影响力的工作

基于MiniCPM的RAG模型BGE-MiniCPM

中文检索排序能力优于同规模模型

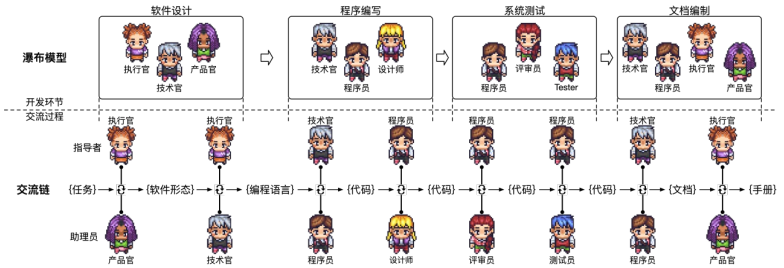


https://github.com/OpenBMB/MiniCPM
https://github.com/OpenBMB/MiniCPM-V

社区口碑项目 —— ChatDev

通过 **ChatDev** 模拟一个

由多智能体协作运营的虚拟软件公司

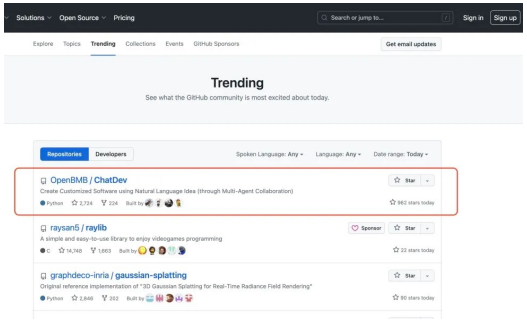


“不到一杯可乐的钱和时间”

开发一款软件 | < 7分钟 | < 3块钱 | 1 个人 + 多个**智能体**

GitHub

2W+ ⭐



GitHub Trending榜单 TOP1 🏆

B站众多开发案例 🤩

吴恩达点赞 👍

社区口碑项目 —— Ultra Series对齐技术

面壁 Ultra 对齐

对齐全球 200+ 大模型

月均下载量超100万 次

Ultra Series高质量对齐数据集

HuggingFace

Zephyr-7B 模型

超越 LLaMA2-70B-Chat
列入 HuggingFace 官方对齐手册

67万 月均下载

Ai2 艾伦人工智能研究所

Tulu v2/OLMo 模型

多个评测榜单上达到开源模型新标杆

16.3k 月均下载

Stability.ai

StableLM-Zephyr 系列

擅长创意与个性化文本生成
3B 量级性能比肩 LLaMA2-70B-Chat

5万 月均下载

Upstage

Upstage-Solar 系列

多个评测榜单超越Mistral-7B,
Qwen-14B, LLaMA2-70B

14万 月均下载

新加坡科技设计大学

TinyLLaMA 模型

各种下游任务表现优于等大开源模型
常识推理表现出色

20万 月均下载

加利福尼亚大学伯克利分校

UCB-StarlingLM 模型

教育, STEM, 人文学科, 写作和
角色扮演方面表现卓越

11万 月均下载

Thomas Wolf

HuggingFace 联合创始人

Thomas Wolf @Thom_Wolf 关注

There is a beautiful story that just happened in AI so let me share it for a lighter tone weekend post among all the doom stories in our AI field this week.
The next day they start diving in the datasets openly shared on the HF hub and stumble upon two interesting large and good quality fine-tuning datasets recently open-sourced by OpenBMB, a Chinese team from Tsinghua: UltraFeedback and UltraChat.

HuggingFace

2023 官方年度总结

2023, year of open LLMs

☀️Summer: In August, UltraLM (a high-performing chat fine-tune of LLaMA) was released by OpenBMB, a Chinese non-profit, and in September, they released the associated preference dataset UltraFeedback, a feedback dataset of inputs compared by GPT4 (with annotations). Throughout the summer, NousResearch, a collective, released several fine-tunes (notably the Hermes and Capybara collections) based on several private and public instruct datasets. In September, a student team from Tsinghua University released OpenChat, a LLaMA fine-tune using a new RL finetuning strategy, and Intel released an Orca style DPO dataset.

🍂Autumn: In October, Hugging Face released Zephyr, a Mistral fine-tune using DPO and AIF on UltraChat and UltraFeedback, and

AlpacaEval Leaderboard

An Automatic Evaluator for Instruction-following Language Models

Caution: GPT-4 may favor models with longer outputs and/or those that were fine-tuned on GPT-4 outputs.

Evaluator: GPT-4 | Claude Filter: Community | Verified | Minimal

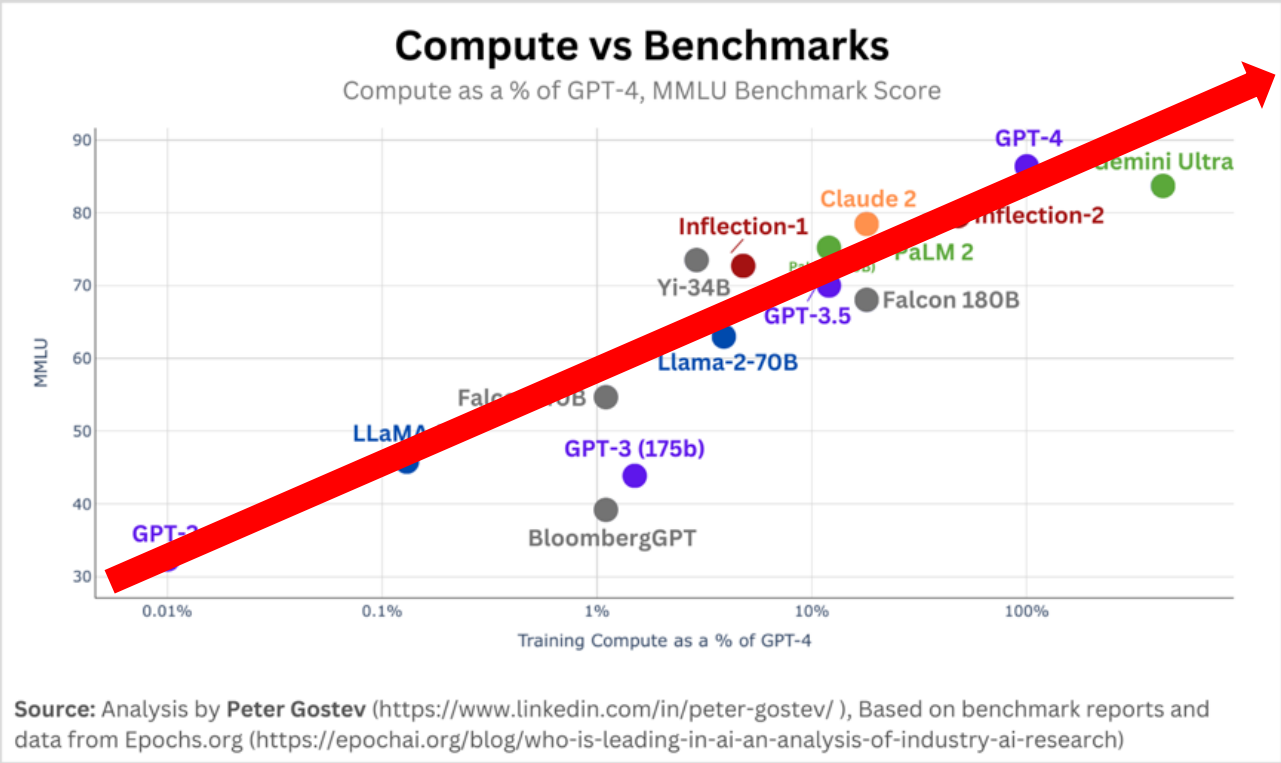
Model Name	Win Rate	Length
XwinLM 70b V0.1	95.57%	1775
GPT-4	95.28%	1365
LLaMA2 Chat 70B	92.66%	1790
UltraLM 13B V2.0 (best-of-16)	92.30%	1720
XwinLM 13b V0.1	91.76%	1894
UltraLM 13B (best-of-16)	91.54%	1980
Claude 2	91.36%	1069

UltraLM-13B-v1.0 发布时登顶斯坦福 AlpacaEval* 开源模型榜单，UltraLM-13B-v2.0（best-of-16 采样）在 AlpacaEval 榜单取得了 92.30% 的好成绩，成为 70B 以下模型最高分。

https://huggingface.co/collections/openbmb/ultra-series-65d490fedc2727f5807f4688

| 2018年以来见证大模型规模法则（Scaling Law）

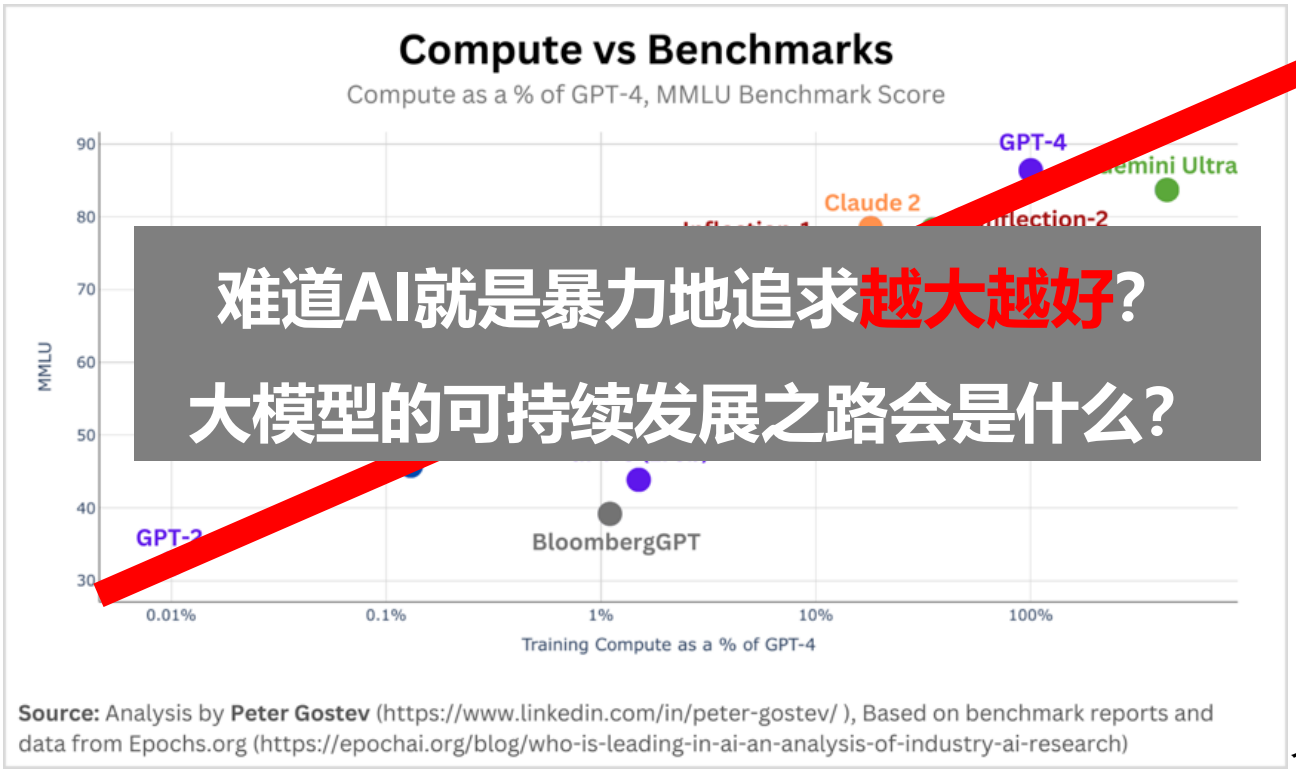
OpenAI引领验证在大数据+大算力支持下，越大模型产生越强智能，即**规模法则**



Source: Analysis by **Peter Gostev** (https://www.linkedin.com/in/peter-gostev/), Based on benchmark reports and data from Epochs.org (https://epochai.org/blog/who-is-leading-in-ai-an-analysis-of-industry-ai-research)

| 2018年以来见证大模型规模法则（Scaling Law）

OpenAI引领验证在大数据+大算力支持下，越大模型产生越强智能，即**规模法则**



难道AI就是暴力地追求越大越好？

大模型的可持续发展之路会是什么？

Llama-3 405B 16000张 H100

1000B 40000张 H100

10000B 400000张 H100

100000B 4000000张 H100

理想情况的线性假设 ↓

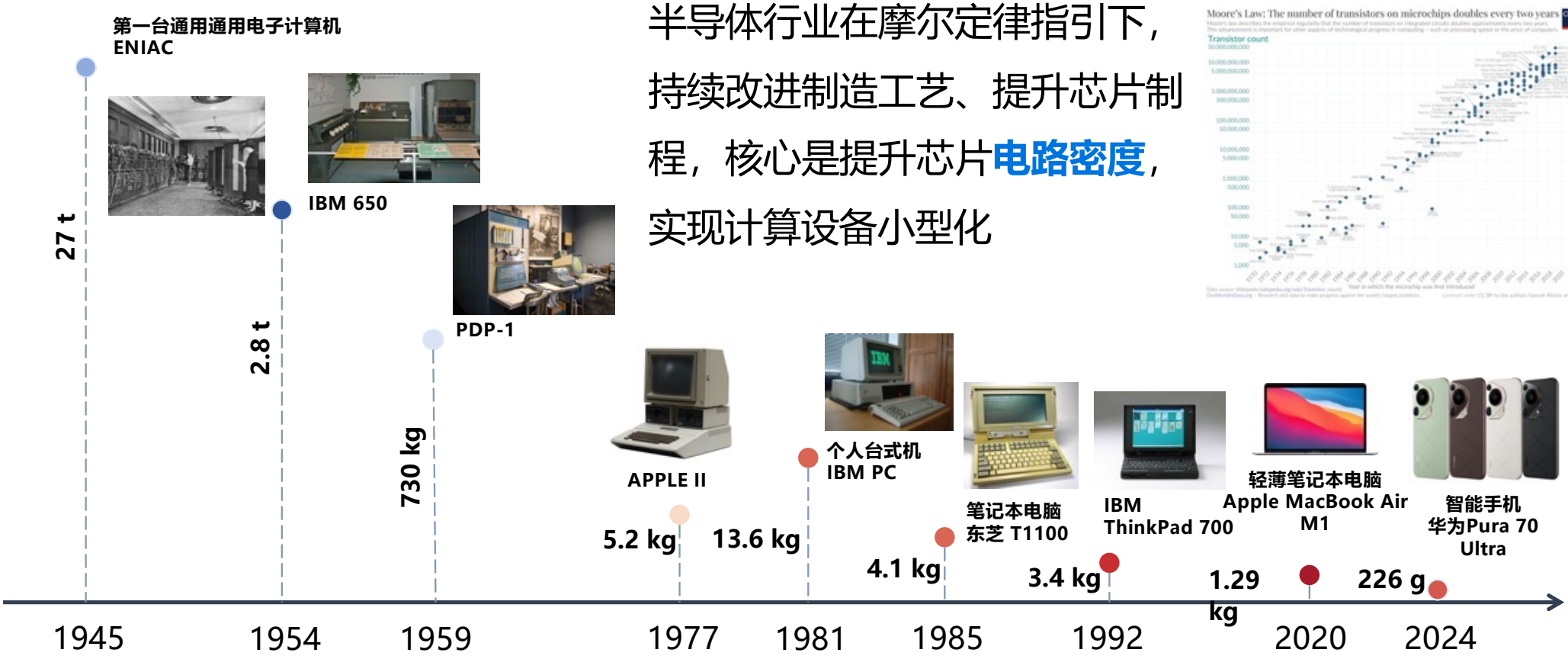
2023年全部H100 GPU产能的**接近十倍**[1]

用电功率超过美国一座**1000万人**城市[2]

人类当今计算集群并行数量**上限的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

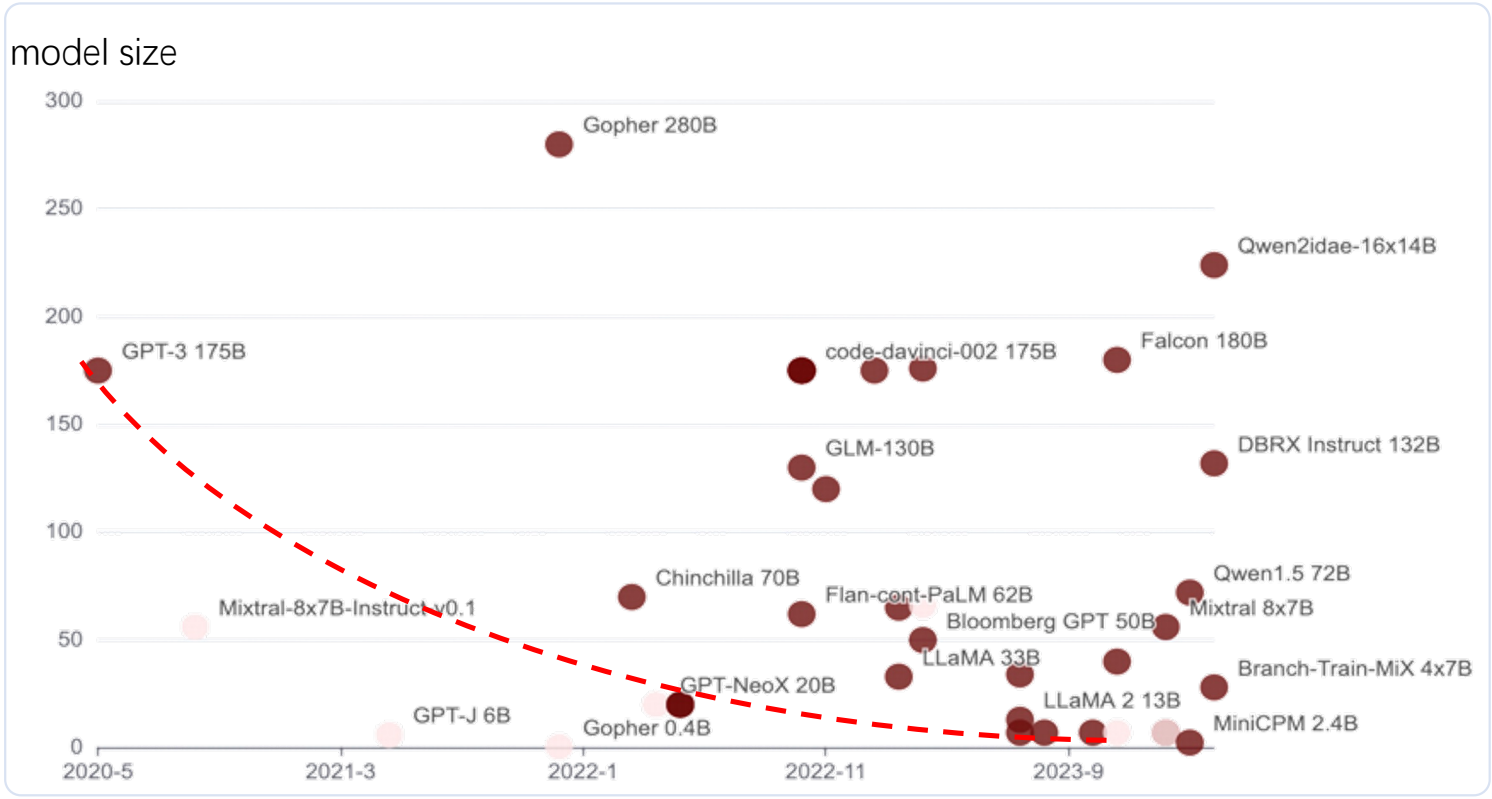
半导体发展趋势：摩尔定律



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**，实现计算设备小型化

大模型发展趋势：知识密度定律

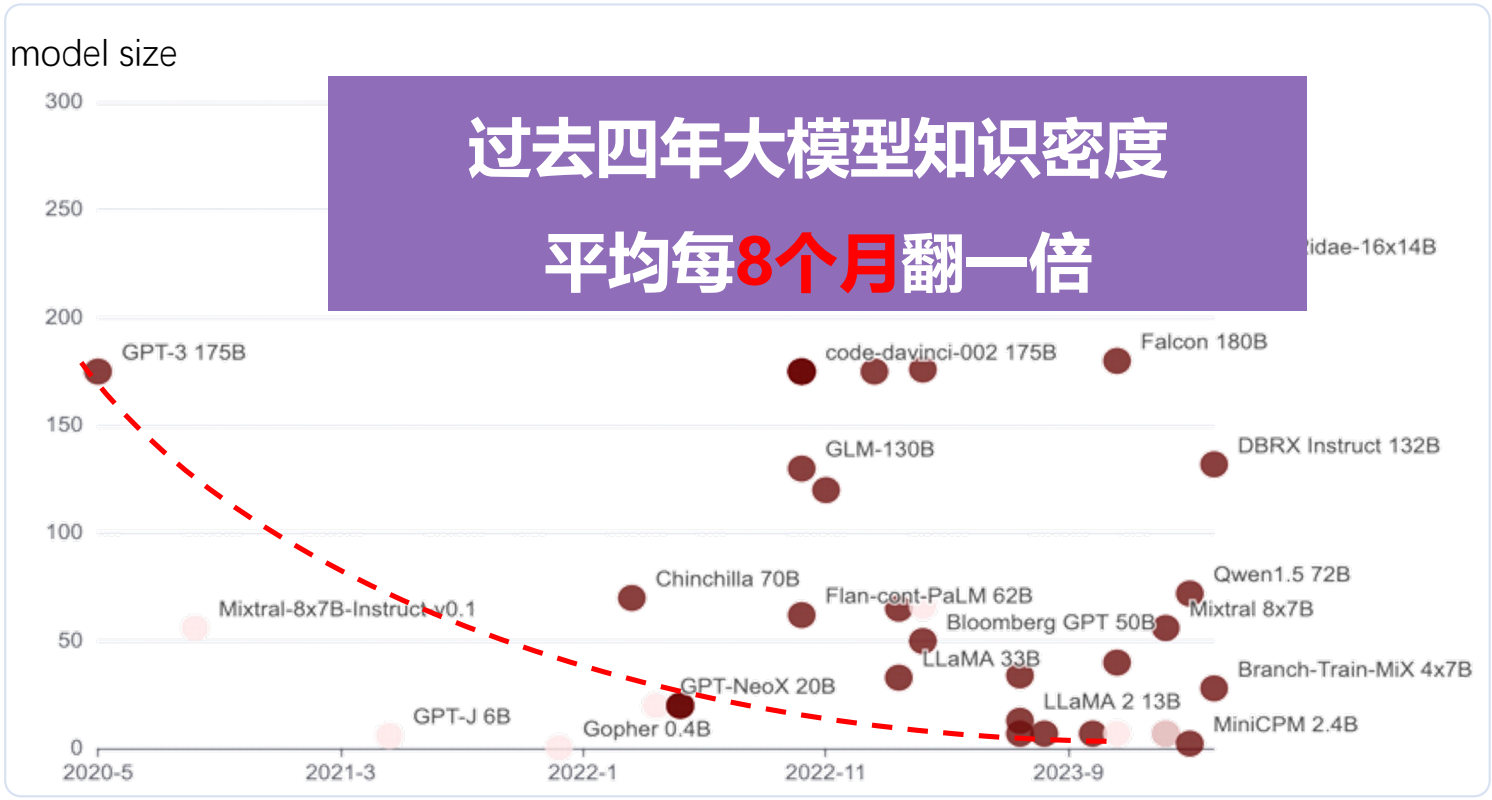
随数据-算力-算法协同发展，模型**知识密度**持续增强：20年GPT-3 **175B**能力24年**2B**参数量即可达到



MiniCPM 2.4B为团队2024年2月发布语言大模型

大模型发展趋势：知识密度定律

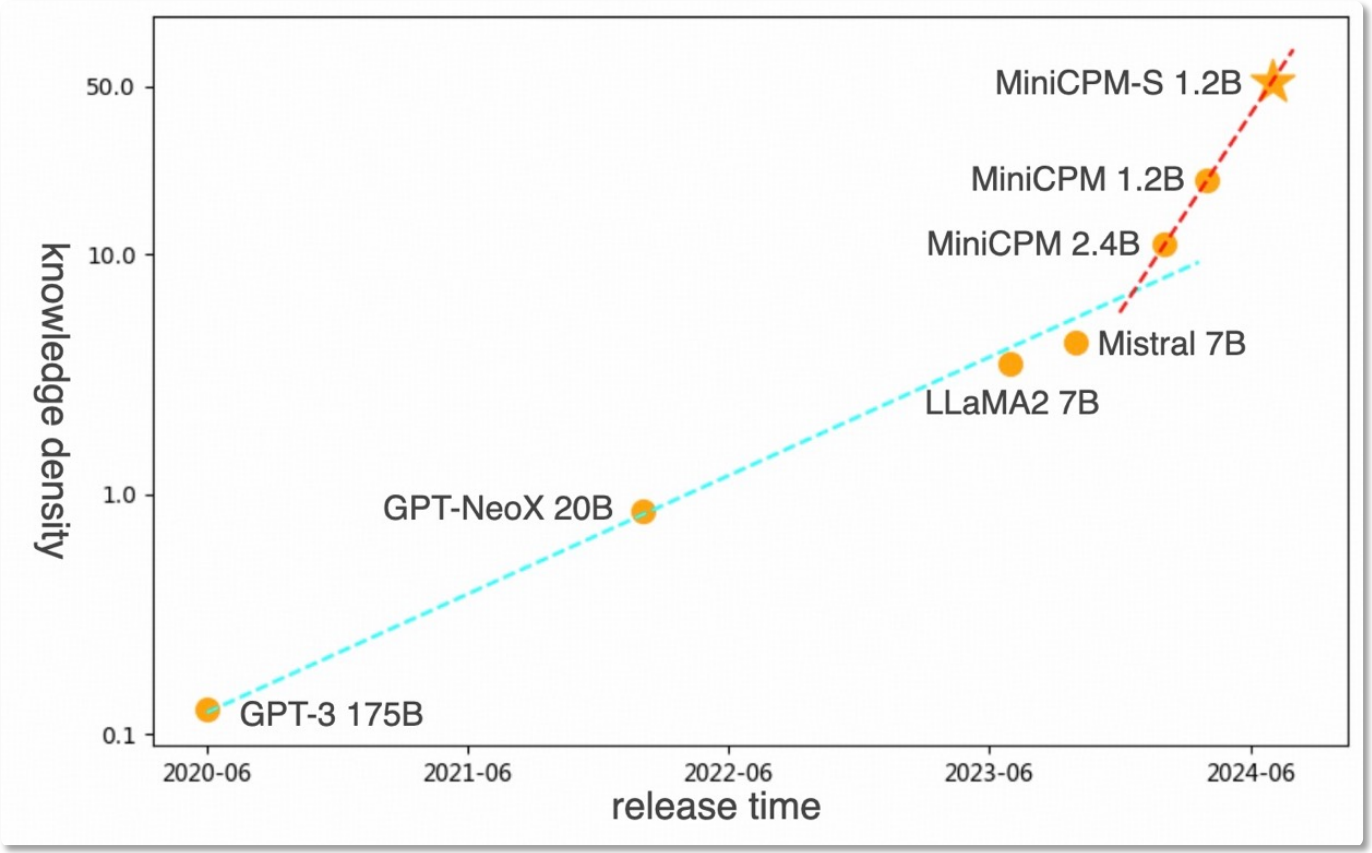
随数据-算力-算法协同发展，模型**知识密度**持续增强：20年GPT-3 **175B**能力24年**2B**参数量即可达到



MiniCPM 2.4B为团队2024年2月发布语言大模型

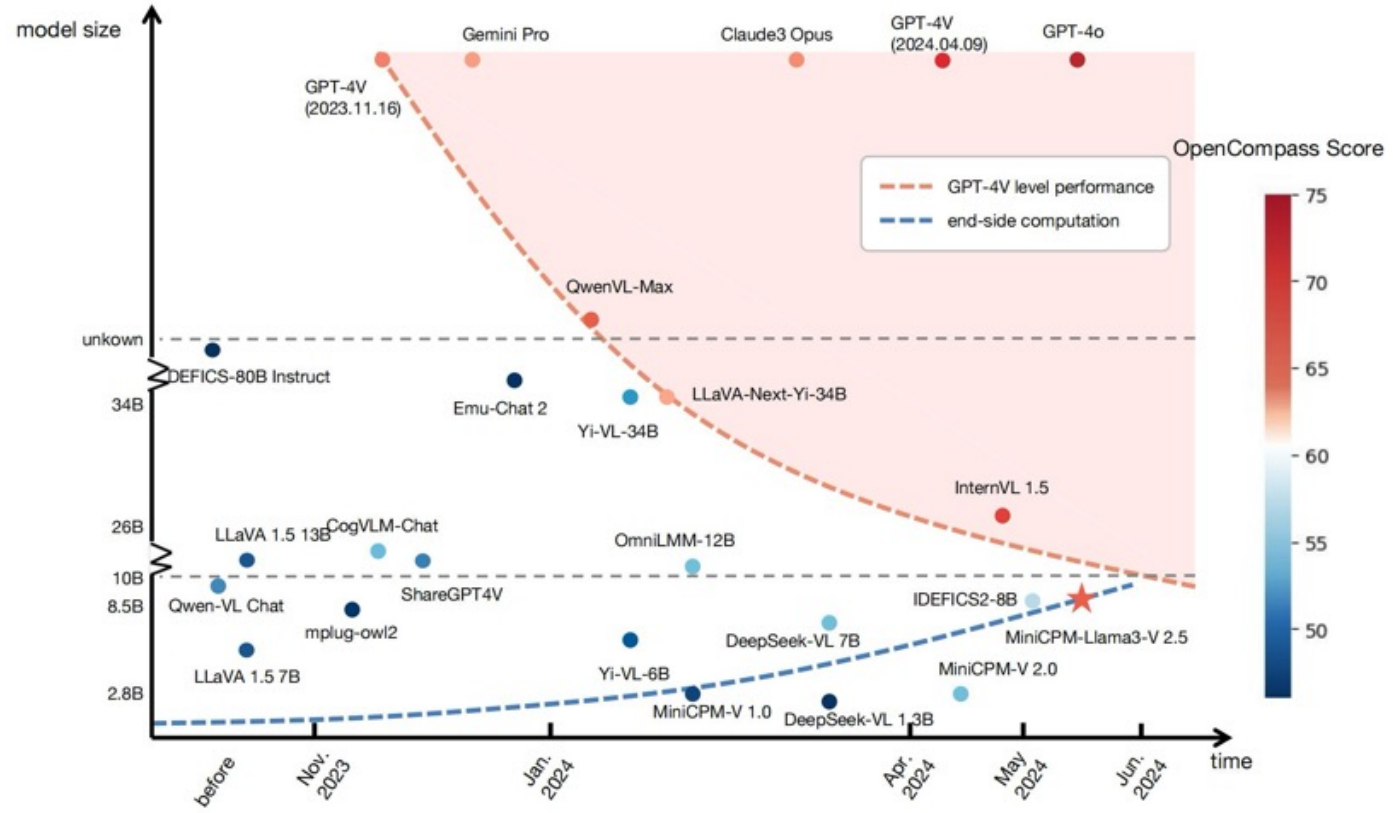
知识密度定义

知识密度 = 模型能力 / (参与计算)参数规模



模型知识密度增强揭示端侧智能潜力

终端算力（摩尔定律）和模型知识密度（知识密度定律）持续增强，两条曲线交汇呈现**端侧智能**巨大潜力



MiniCPM-V 2.5为团队2024年5月发布多模态模型

端侧算力总量巨大亟待激活

7100+ EOPS

存量手机终端算力总规模

(存量手机终端近 10 亿)

2022年全国数据中心算力 **12+倍**

阿里云张北超级计算中心 **147+个**

英伟达 H100 芯片近 **100万片**

智能终端年出货量

智能手机 **2.7亿台**

个人电脑 **4,000万台**

智能汽车 **2,000万台**

智能家居 **3.3亿台**

智能穿戴 **1.2亿台**

注：1）算力规模统一这算为INT8算力进行类比，1TOPS代表处理器每秒钟可以进行一万亿次操作，1EOPS等于100万TOPS
2）出货量数据来自IDC、Canalys、中商产业研究院、高工机器人产业研究所预测

模型产品 | MiniCPM 旗舰端侧模型，做世界最好的终端智能

面壁智能 ModelBest

MiniCPM 1.0

MiniCPM

- 全球领先的轻量高性能端侧大模型
- 2B规模性能越级超越Mistral-7B

MiniCPM-V

- 首次端侧部署多模态
- 部分能力超越10B、17B量级模型

MiniCPM 2.0

MiniCPM-V 2.0

- OCR综合性榜单开源模型最佳表现
- 通用场景文字理解比肩Gemini Pro

MiniCPM-V 2.5

- 最强端侧多模态，综合性能超多模态巨无霸Gemini Pro 、GPT-4V

MiniCPM-V 2.6

- 取得单图、多图、视频理解方面「三位一体」SOTA成绩（20B以下）；
- 在端侧形成全面对标GPT-4V局势

MiniCPM 3.0

MiniCPM-V 3.0

- **表现超越GPT-3.5-T**urbo-0125和Phi-3.5-mini-instruct；
- 比肩 Llama3.1-8B-Instruct、Qwen2-7B-Instruct、GLM-4-9B-Chat 等多个 7B-9B 参数量的模型。
- 支持工具调用（Function Calling）和代码解释器（Code Interpreter）
- 出色的中英文指令遵循能力：**效果超越 GLM-4-9B-Chat、Qwen2-7B-Instruct**。
- 长文本能力：提出 LLM x MapReduce，理论可处理的**上下文长度达到 +∞**。
- RAG能力：发布了 MiniCPM RAG 套件，**在中文、中英跨语言检索测试中取得 SOTA 表现**；

MiniCPM 3.0 | 再次刷新端侧模型最强战力

面壁智能 ModelBest

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							
MMLU	67.2	70.5	72.4	72.6	69.4	69.2	68.4
BBH	70.2	64.9	76.3	65.2	67.8	70.3	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	71.6	70.1	19.2	73.3	75.4	48.4
综合能力							
平均分	66.3	65.3	65.0	57.9	60.8	61.0	57.2

提前近4个月，面壁智能实现了最初发布时立下的Flag：

2024年内
让GPT-3.5水平的模型在端侧跑起来！

仅4B参数，MiniCPM3.0在自然语言理解、知识、代码、数学等多项能力上对 GPT-3.5实现赶超，并越过Qwen2-7B，Phi-3.5，GLM4-9B，LLaMa3-8B等一众中外知名模型的表现脱颖而出。

首次上端

实时视频理解

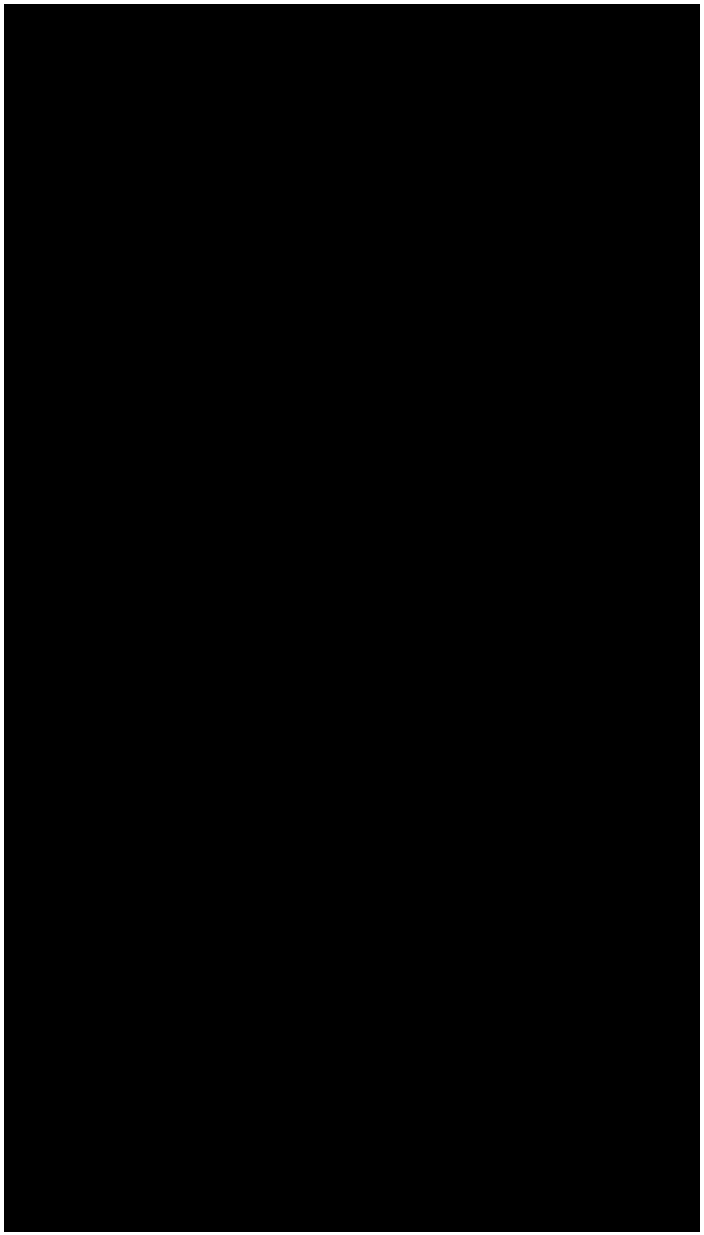
实时看见与理解真实世界
开启具身智能等 AGI 无限可能

帮我看看这里是什么样子？

瞧，她正要画些什么？

太长不看，这段视频讲了什么？

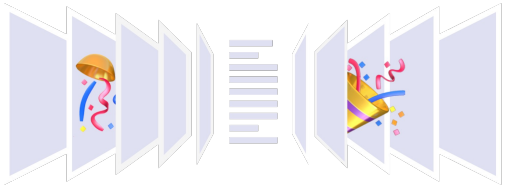
北京的早高峰是什么体验？



首次上端

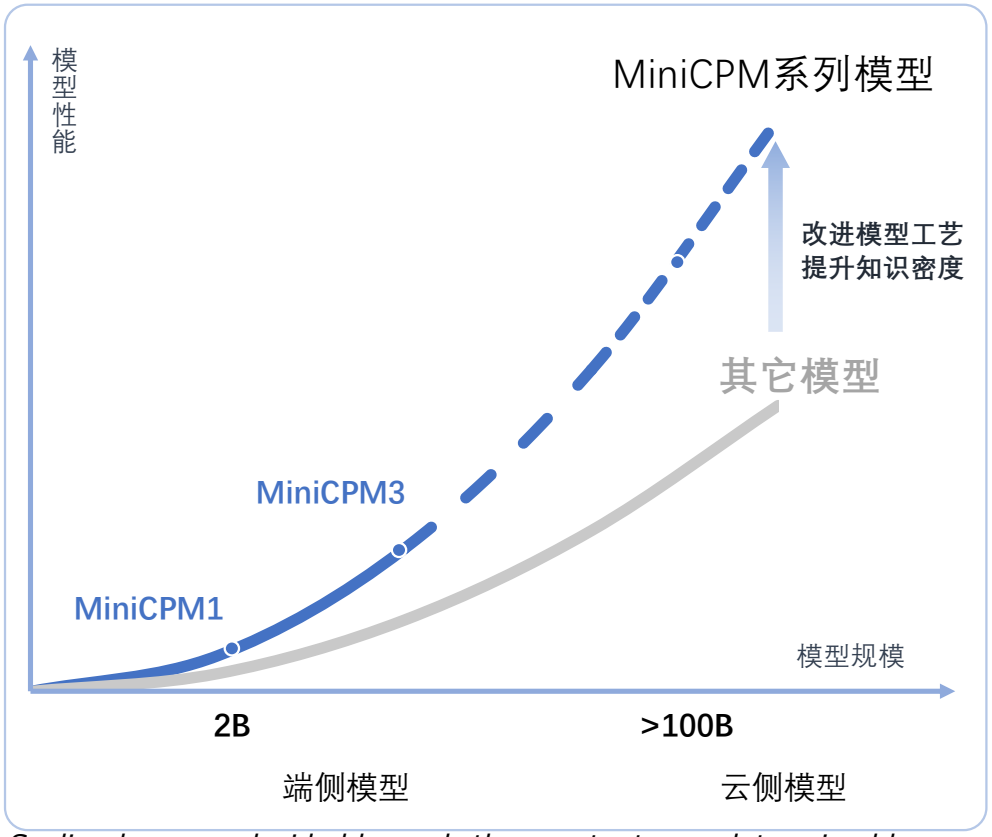
流畅的多图联合理解

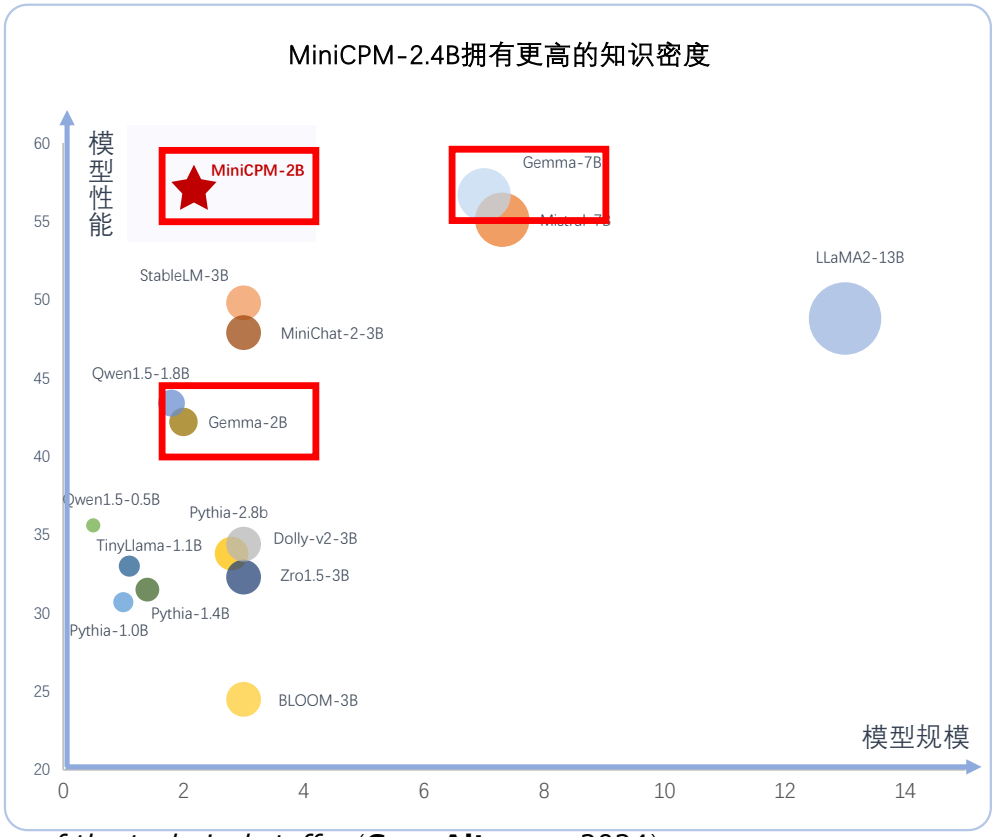
多张图片一口气处理
一直识图一直爽



超聪明的复杂推理能力
难图梗图不在话下

更高模型制造工艺带来更高模型知识密度





Scaling laws are decided by god; the constants are determined by members of the technical staff. （**Sam Altman**，2024）

"Scaling Law 是上帝决定的，参数是靠科学家追寻的 。" ———— 山姆·阿特曼

**注：右图根据公开测评集的测评结果绘制，具体评测集包括 MMLU、CMMLU、CEVAL、GSM8k、MATH、HumanEval、MBPP、BBH、HelloSwag*

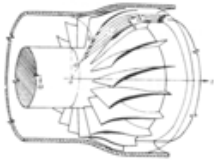
人工智能科学化：推动大模型高质量发展

数据驱动的大模型技术方向大致确定，架构-算法-数据的技术路线高速迭代，需要围绕如何**极致提升“知识密度”**，探索大模型科学化建设方案

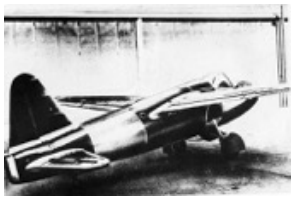
科学化引领高质量发展

科学化发展引领高质量发展

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）

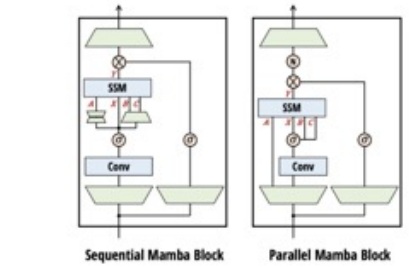
三叉戟客机

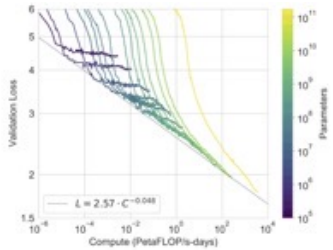
A-7E

F4K

歼8（国产）

大模型科学化问题





模型架构

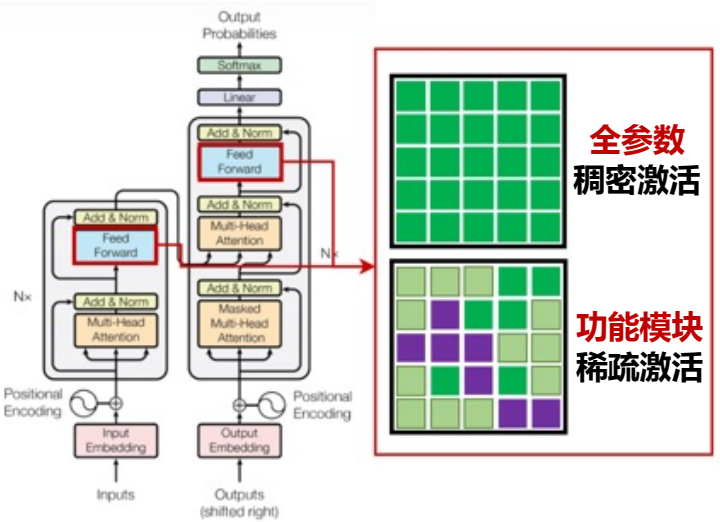
数据治理

数据->知识
成长规律

人工智能科学化-架构高效：模型功能分区

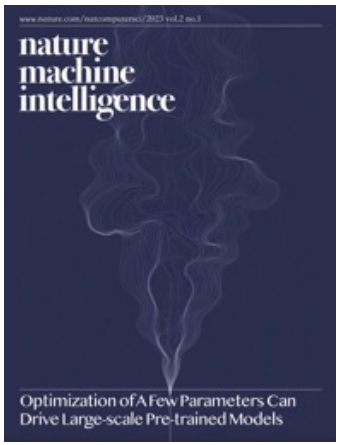
大模型自发涌现类脑的**功能模块分区**，表明大模型参数的**空间可解耦特性**，动态功能分区训练可以实现稀疏激活推理成本降低80%以上，外部接入功能模块仅微调5%参数即可适配具体任务

模型内部动态功能分区



模型外部功能模块接入

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

人工智能科学化-数据高效：高质量数据治理

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

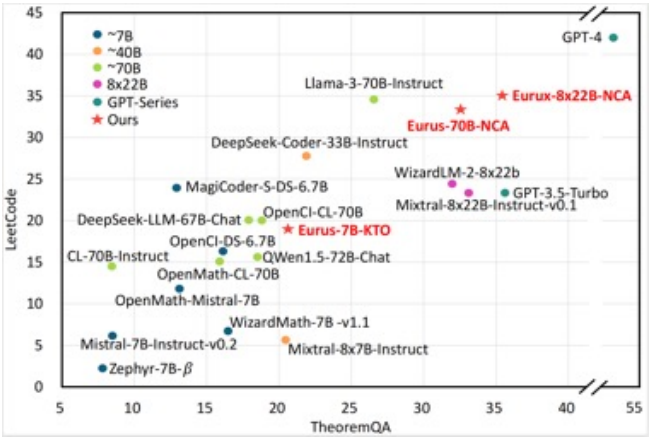
数百人 专业数据清洗团队

十万亿 Tokens 级 高质量数据集

中文 英文 图文 代码

清洗率 90%+

十万亿Tokens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

RLAIF-V

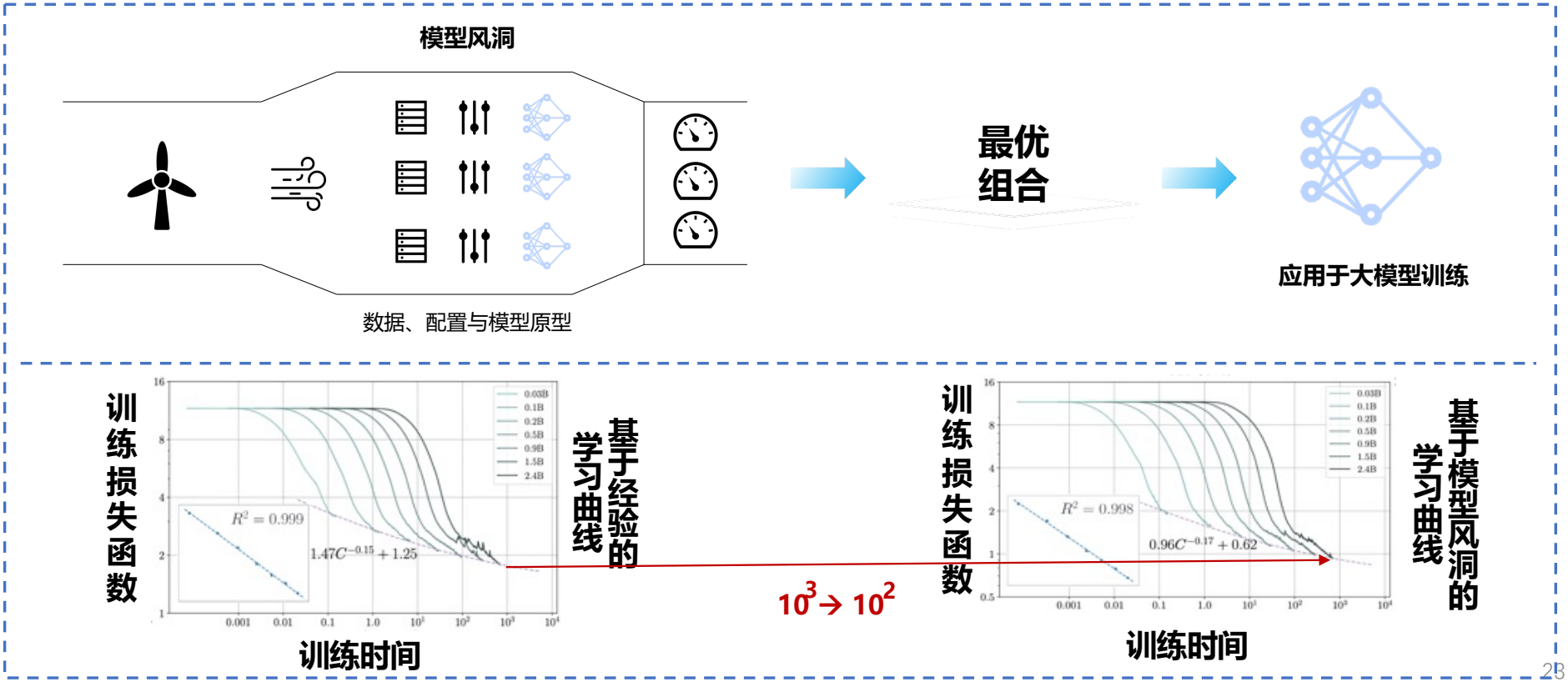
MiniCPM-V 2.5 官方多模态对齐数据

曾登上HF Datasets Trending榜单

https://huggingface.co/models?dataset=dataset:HuggingFaceH4%2Fultrachat_200k&sort=trending

人工智能科学化-成长高效：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



人工智能科学化-成长高效：模型风洞技术

超参稳定的模型规模扩增

同一超参掌握所有模型

最优学习率

在任意规模取得最优loss

最优Batch Size

收敛速度与资源消耗最优平衡

固定模型倍增上限

随时可退火，阶段最优增长倍数

持续训练友好

在退火阶段加入高质量数据可以获得更优的能力，也支持持续训练

最优学习率调度器

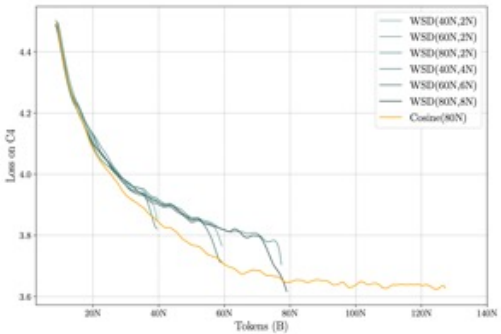
Warmup-Stable-Decay（WSD）调度器。全新学习率调度策略，取得最佳Decay步数

Warmup-Stable-Decay（WSD）

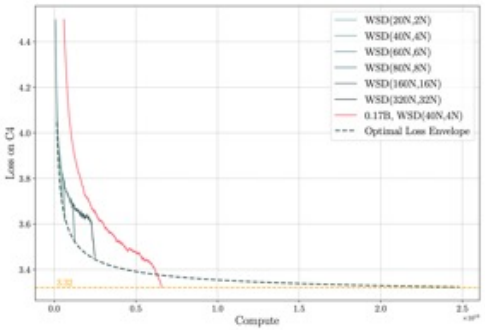
最优学习率调度器

$$WSD(T;s) = \begin{cases} \frac{s}{W}\eta, & s < W \\ \eta, & W < s < T \\ f(s-T)\eta, & T < s < S \end{cases}$$

WSD LRS包含三个阶段:预热阶段（结束步长用W表示）、稳定训练阶段（结束步长用T表示）和剩余衰减阶段



在WSD LRS的衰减阶段，模型训练损失突然减小



连续训练一个0.036B的模型可以达到0.17B模型的性能

端侧智能广泛应用场景

穿戴设备

- 健康监测分析
- 手势与行为识别
- 运动训练辅助
- 个性化提醒和服务

手机场景

- 离线智能助手
- 图像识别和对话
- 个性化内容推荐

PC场景

- 智能写作辅助
- 个性化搜索和推荐
- 内容摘要和总结
- 图像,视频编辑增强

智能家居

- 家居设备智能控制
- 记忆习惯自动调节
- 智能对话陪伴系统

汽车场景

- 车载语音助理
- 车辆状态监控
- 情感陪伴聊天

具身智能

- 感知理解
- 人机交互，自主操作
- 决策规划
- 陪伴护理

金融场景

- 本地化智能客服
- 预防金融风险
- 行业资料一键速读
- 敏感隐私保护
- 量化投资交易

政务场景

- 保密文件摘要
- 优化政务流程
- 规范公文写作

智能汽车典型场景

面壁智能 ModelBest

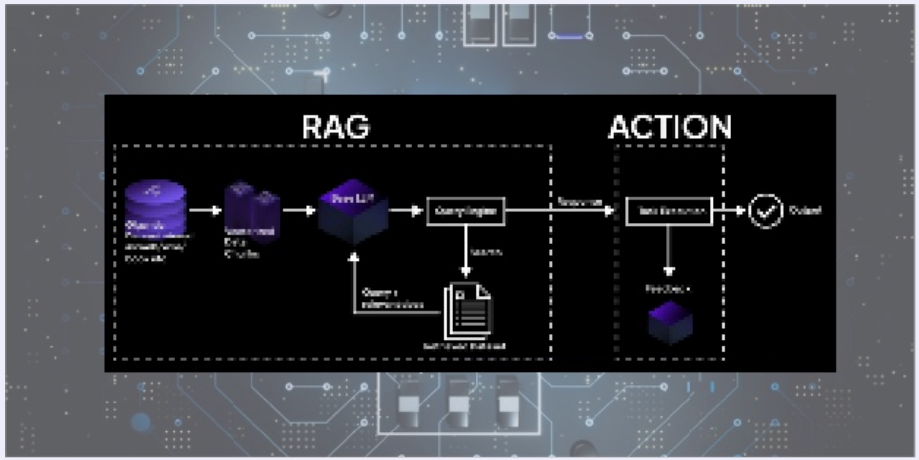
识别道路要素

辅助导航

识别周围的停车场、超市等关键设施，并通过导航系统引导驾驶员前往目标地点。这一功能可以防止导航设备的偏差，提高精确度。

识别道路状况

识别交通拥堵、施工、事故等信息，弥补传统导航系统在时效性上的不足，提供更及时、更准确的行车建议。



车书查询

向车载系统提出关于车辆使用手册、保养指南、常见问题等方面的问题，系统会根据用户的问题，从车辆的数字手册中快速检索并给出精准的答案。

维修指导

依据车辆的状态、故障码等信息，系统会根据用户的问题，提供维修指导、配件信息、注意事项等。

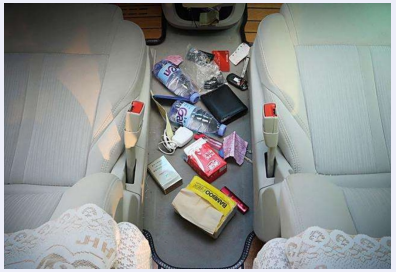
智能汽车典型场景

面壁智能 ModelBest

车内人员、氛围综合识别

分析颗粒度提升

姿态识别种类 **100+**



物品识别种类增加

日常物品全覆盖 **5000+**

场景理解推理能力

基于表情、动作的综合理解

车内车外场景综合识别

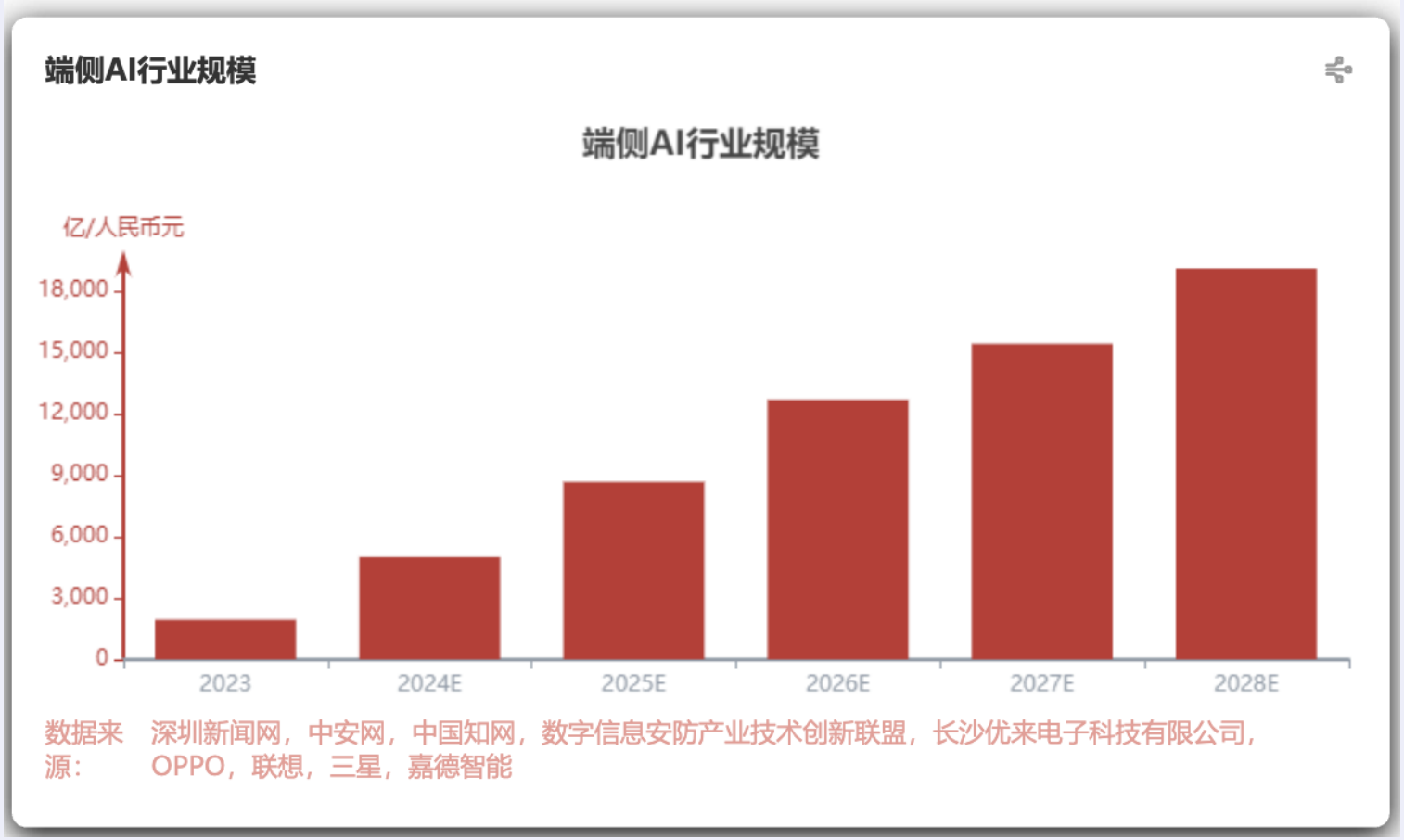
准确识别当前的路况，如草地、泥泞、沙地等，为驾驶员提供实时的路况信息。

识别标志性建筑物和景点，实时讲解

识别停车场、超市等，辅助导航

端侧模型的市场规模

面壁智能 ModelBest

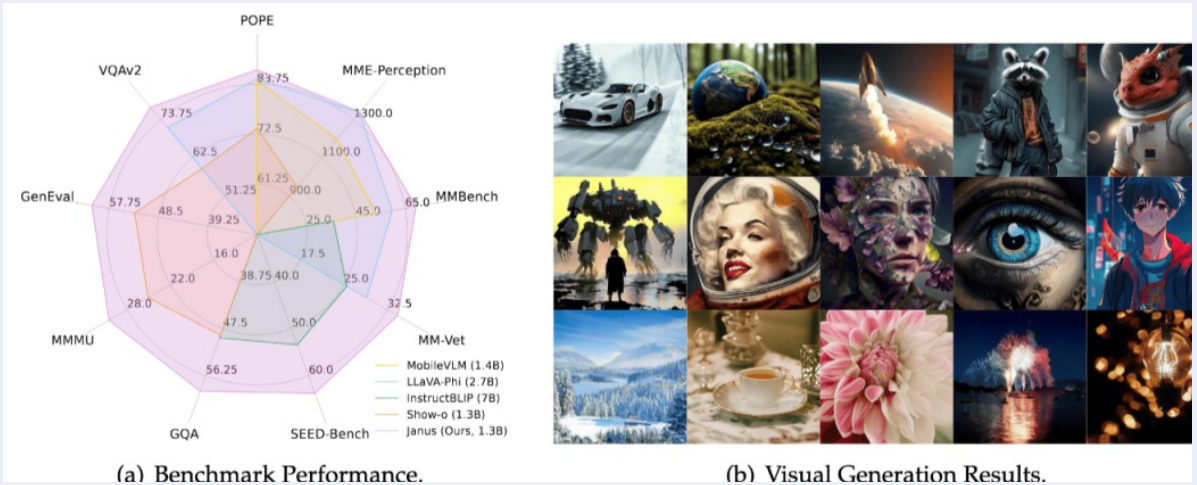


- 预计2024年—2028年，端侧AI行业市场规模由5,000.44亿人民币元增长至19,071.30亿人民币元，期间年复合增长率39.75%

国际前沿端侧模型发展

面壁智能 ModelBest

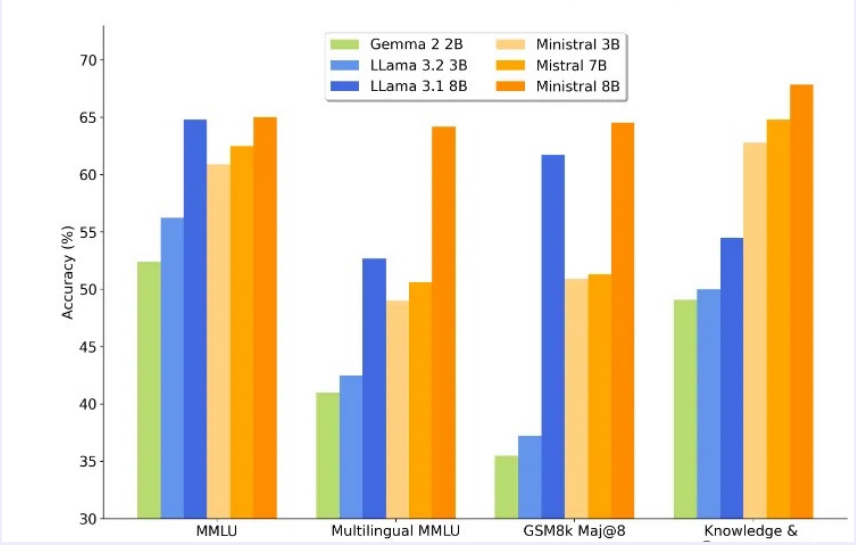
- DeepSeek 推出了 Janus 1.3B，这是一款开源的、统一的多模态理解与生成模型，能够同时处理文本和图像。通过将视觉编码解耦为独立路径，同时利用单一的统一 Transformer 架构进行处理，解决了以往方法的局限性。



(a) Benchmark Performance.

(b) Visual Generation Results.

- Mistral 公司于 10 月 17 日发布了两款新的边缘计算小模型：Ministral 3B 和 Ministral 8B，主要适用于设备端计算和边缘计算场景。两款模型在 10B 以下规模的知识、常识、推理、函数调用和效率方面表现出色，支持 128k 上下文。



端云协同将成为AGI时代基本形态

以大模型为核心的类人智能引擎

终端智能

无所不在

- 低成本
- 高效率
- 低时延
- 高可靠
- 个性化
- 隐私化

面壁智能

云端智能

无所不能

- 深度领域专家
- 复杂大型工程
- 前沿科研探索

OpenAI

以大模型智能体为载托的新时代基础设施

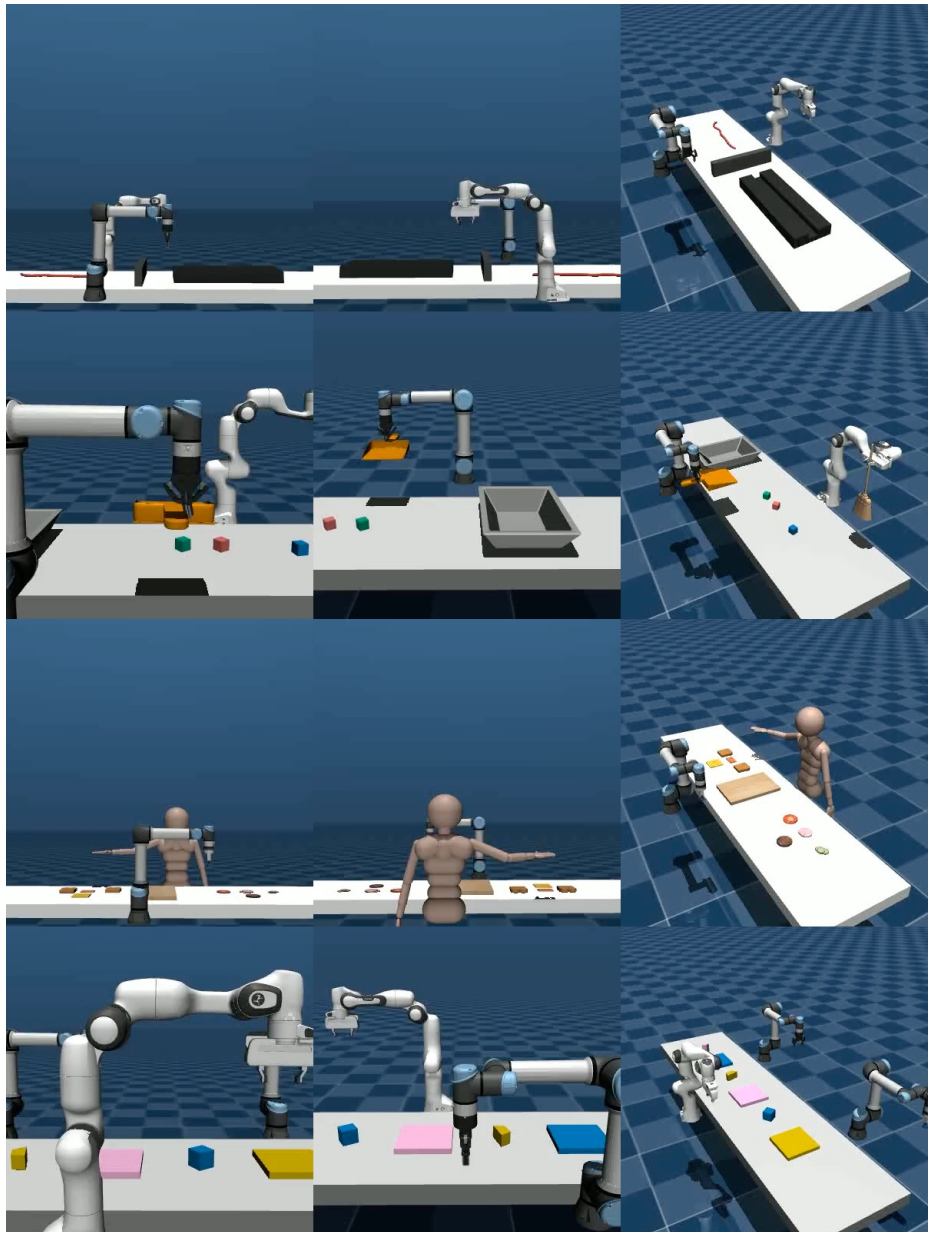
《中国AIGC应用全景报告》报告提出，2024中国AIGC（生成式人工智能）应用市场规模将达200亿元，2030年达万亿规模

展望：智能体互联网

- 制定智能体接入与沟通协议，可让**异质智能体**沟通协作
- 示例：**异质具身智能体**高效沟通协作完成 RocoBench 任务

Model	Metric	Cabinet	Sweep	Sandwich	Sort	Rope
Central Plan (oracle)	Success	0.90	**1.00**	0.96	0.70	0.50
	#Step	4.0	8.4	8.8	8.6	2.3
Roco Dialog	Success	0.75	0.70	0.70	0.70	**0.70**
	#Step	4.7	7.9	9.1	5.4	2.4
IoA	Success	**1.00**	0.80	**1.00**	**1.00**	**0.70**
	#Step	4.6	8.5	8.9	5.8	2.6

三个任务上取得100%准确率，
超过为这个任务专门设计的多智能体框架



展望：智能体互联网

大模型与终端结合为**智能体（Agents）**互联形成智能体网络，将迎来AI第二次涌现

Internet

信息高速，全球连接

全球信息高速公路，
将每一位网民紧密相连

Internet of Things

设备互联，生活便捷

物体间的沟通桥梁，
使每一件设备灵活互联

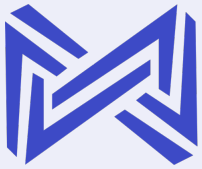
Internet of Agents

数字孪生，万物智联

交织万物的网络纽带，
绘制人机共生的未来蓝图

OpenBMB

面壁小钢炮
MiniCPM



面壁智能
ModelBest

将大模型放到用户最近的地方

AGI FOR LIVES 智周万物