艺恩 endata

INDUSTRY WHITE PAPER · LLM DATA

全球大模型数据市场白皮书

The Global Data Market for Large Models

当算力见顶，数据成为 AI 时代的价值坐标。本白皮书系统呈现全球大模型数据市场的规模、价值链、资本、合规与多模态前沿。

艺恩数据 · ENDATA · NEEQ 871430

2026 年版

四个章节,读懂数据市场

数据,正成为大模型时代的稀缺生产要素

进入 2025–2026 年,随着算力竞赛逼近边际、公开互联网语料趋于枯竭,**数据**已从「可廉价获取的原料」 转变为决定模型上限的稀缺生产要素。市场的核心命题,正由「数据规模」转向「数据质量、专业度与合规性」。

20–35%
全球 AI 训练数据相关市场年复合增速区间(多家机构口径)

2026–32
Epoch AI 测算的公开人类文本语料耗尽窗口(中位约 2028)

143亿$
Meta 入股数据公司 Scale AI 金额,估值达 290 亿美元

15亿$
Anthropic 版权和解额——美国史上最大版权和解

三个结构性信号

① **峰值数据逼近** ——公开语料趋于枯竭,价值向高质量、专家级、合规与合成数据迁移; ② **资本空前涌入** ——数据与专家公司估值集体飙升,内容授权走向规模化; ③ **合规成为护城河** ——诉讼频发叠加欧盟透明度义务,合规数据获显著溢价。

本白皮书为对外发布的行业研究,不构成投资建议;前瞻性表述以「预计/预测」标识,完整来源见末页。

01

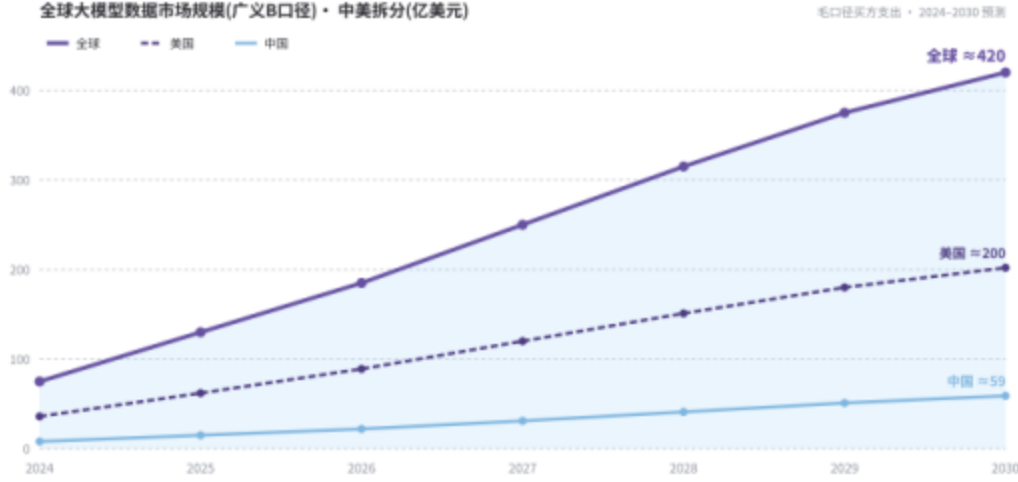
PART 01 · MARKET & INFLECTION

市场与拐点

规模口径分歧 · 增速共识 · 峰值数据理论

狭义口径,显著低估真实市场

常被引用的「AI 训练数据集」狭义口径**仅约 28–32 亿美元**(2024–25),只统计打包数据集+标注软件。但本白皮书采用**广义口径(B)= 数据集 + 采集标注 + RLHF/专家数据 + 合成数据**:自下而上测算**2024 约 60–90 亿、2025 约 100–160 亿美元**(毛口径买方支出)。



为何狭义口径失真

- **钱在服务里:**真实支出多在标注与 RLHF/专家数据**服务**,而非打包数据集。
- **三家即超全市场:**Scale(约20亿)+Surge(约14亿)+Mercor(约7.6亿)2025 毛收入合计 **约 42 亿美元**,已超「训练数据集」狭义全球值。
- **口径关系:**训练数据集 ⊂ 采集与标注 ⊂ 数据服务;狭义是子集而非全貌。

来源:狭义口径 MarketsandMarkets(28.2亿/2024)、Grand View(32亿/2025);广义区间为自下而上加总(Scale/Surge/Mercor/Turing/Appen/Innodata/海天/数据堂等 2025 营收)与 GVR「数据采集与标注」(48.9亿/2025)交叉验证。头部厂商多为毛口径(含承包商支付)。2030 为预测值。

各细分赛道,高速增长共振

DATA COLLECTION & LABELING

采集与标注

2024: 37.7 亿 →
2030: 171 亿美元

CAGR 28.4% | GVR

ANNOTATION TOOLS

标注工具

2025: 32 亿 →
2035: 343.8 亿美元

CAGR 26.8%

Precedence

SYNTHETIC DATA

合成数据(最快)

2023: 2.18 亿 →
2030: 17.88 亿美元

CAGR 35.3% | GVR

DATA-AS-A-SERVICE

DaaS(泛企业级)

2023: 143.6 亿 →
2030: 768 亿美元

CAGR 28.1%

GVR · 全行业

大厂入局合成数据

NVIDIA 于 2025 年以约 **3.2 亿美元**收购合成数据公司 Gretel.ai,标志头部算力厂商正式将合成数据纳入战略版图。

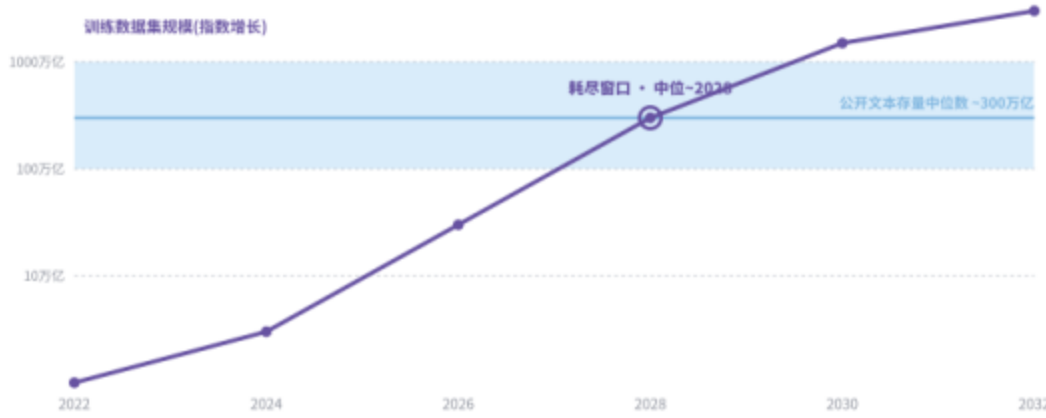
增长的底层逻辑

Stanford HAI《2025 AI Index》:训练计算量每 5 个月翻倍,**数据集规模每 8 个月翻倍**——数据需求的指数级膨胀,是各赛道高增速的根本驱动。

来源:GVR、Precedence、TBRC、Stanford HAI。注:DaaS 为**泛企业级数据服务**(覆盖全行业数据交付),口径远大于 P05「AI 训练数据集」的狭义市场,二者非同口径、不可直接相加比较。

公开语料枯竭,价值范式转移

Epoch AI(经 ICML 2024 同行评审)测算:可用人类公开文本存量约 300 万亿 token, 若当前趋势持续,训练数据集规模将在 **2026–2032 年间**与之持平,中位数预测约 2028 年。



来源:Epoch AI《Will we run out of data?》(ICML 2024);Stanford HAI《2025 AI Index》。

「数据墙」三重证据

- **存量有限分层。** CommonCrawl 约 130 万亿、索引网络约 510 万亿 token,高质量部分远小于此。
- **过度训练加速耗尽。** Llama 3 过度训练约 10 倍;若转向 100 倍,数据触顶更早。
- **多轮训练放大 3–15 倍**有效存量,但难以根本解决枯竭。

从「更多数据」到「更对的数据」

行业领袖判断

「人类知识的累积总和,已基本在 AI 训练中被耗尽——大体上去年就发生了。」

— Elon Musk,2025/1(via The Guardian)

真正的问题

「如果训练模型的最佳方式是生成一千万亿 token 合成数据再喂回去,那会很奇怪」——核心是**「如何从更少的数据中学到更多」**。

— Sam Altman,2024/6(述要,via The Decoder)

四条出路

- **多模态扩容。** 引入图像/视频/音频可使训练数据约增 3 倍(Epoch)。
- **合成数据。** 以模型生成数据反哺训练,成为缓解数据墙主路径。
- **数据效率与策展。** 更少但更优质数据获更强能力,「数据中心化 AI」兴起。
- **高质量/专家数据。** 通用语料见顶,稀缺的专业、垂直、合规语料价值凸显。

本章要点

公开人类文本趋于枯竭(中位约 2028),迫使行业从「数据规模」转向「质量、专业度与多模态扩容」——这是后续所有市场变化的根本动因。

02

PART 02 · VALUE CHAIN & CAPITAL

价值链与资本

八层结构 · 质量溢价 · 估值狂飙与内容授权

八层结构,价值层层递进

①	**预训练语料(web/书籍/代码)**	规模化基础,边际价值随枯竭下降
②	**SFT 指令微调数据**	对齐任务格式,塑造可用性
③	**RLHF / 偏好数据**	对齐人类偏好,决定「好用」程度
④	**RLAIF / AI 反馈数据**	降低人工成本的规模化对齐
⑤	**专家 / 领域数据(PhD 级)**	突破专业能力天花板,溢价最高的人工数据
⑥	**评测 / 基准数据**	能力度量与迭代方向的标尺
⑦	**合成数据**	缓解数据墙、增速最快的新供给
⑧	**多模态数据(图像/视频/4D)**	最稀缺、溢价最高的层级

注:价值排序为相对定性概括,实际溢价随任务、领域与稀缺度浮动。

核心规律

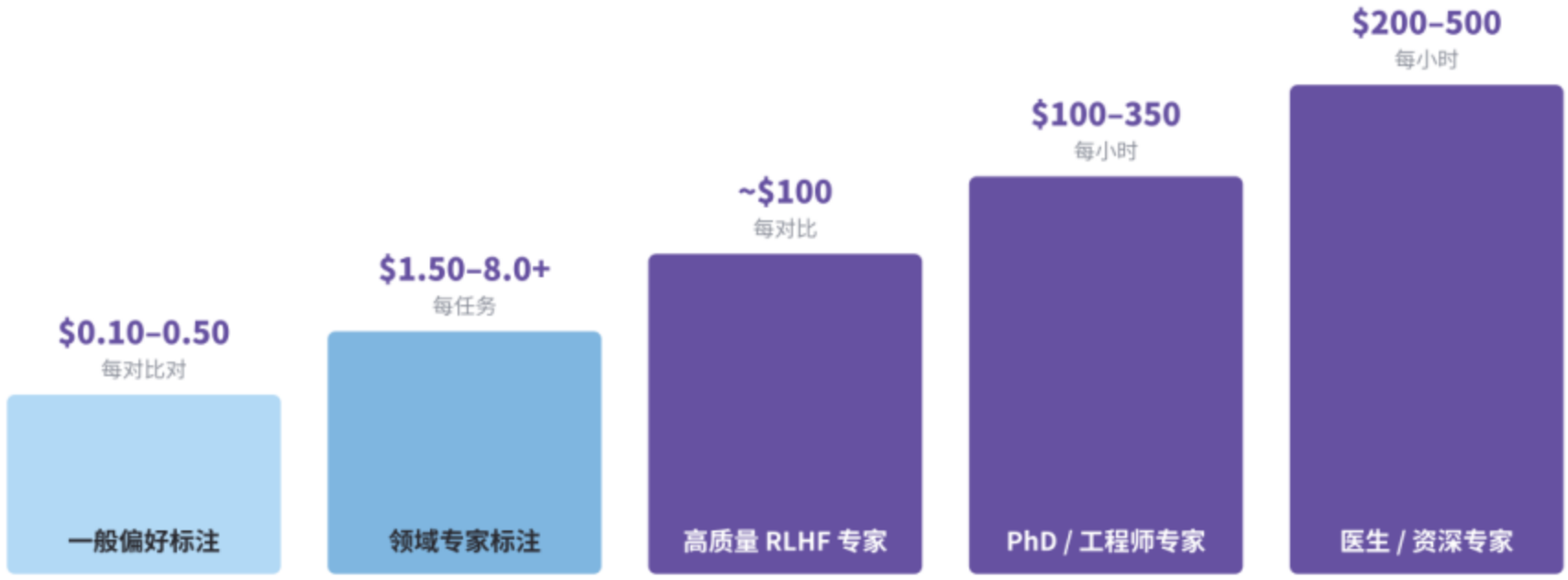
越靠近**「专家级、多模态、可验证」**的一端,单位价值越高、可复制性越低。 当通用网络语料见顶,价值链上半部(⑤–⑧)的稀缺溢价持续抬升。

70%+

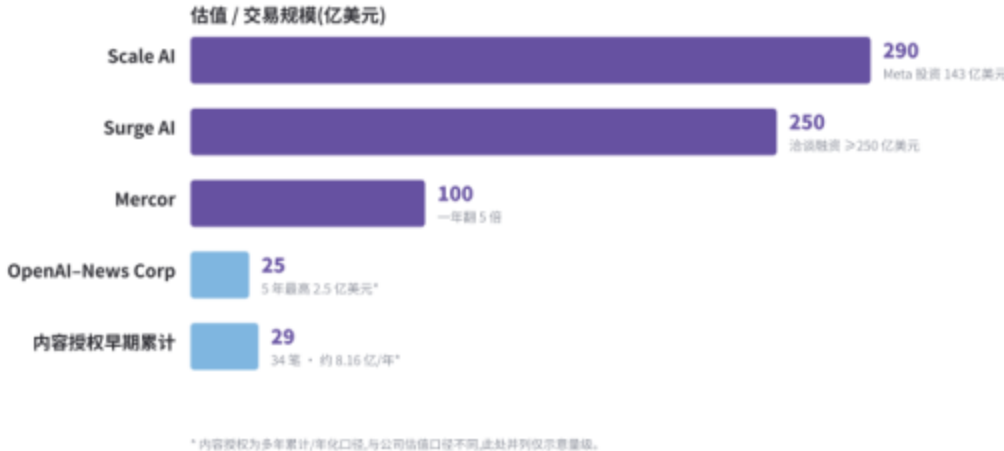
业界观察:模型性能提升中归因于**数据质量**(而非架构)的比例 (Technavio,数据中心化 AI)

同一份标注,价差可达数十倍

标注单价随专业度上升 · 通才→专家可达数十倍价差



估值狂飙,资本以真金确认数据稀缺



* 内容授权为多年累计/年化口径,与公司估值口径不同,此处并列仅示意量级。

头部交易

- **Scale AI:** 2025/6 Meta 投资 143 亿美元获 49% 无投票权股份,估值 290 亿。
- **Surge AI:** 2025/7 洽谈以 ≥250 亿美元估值融资;ARR 至 8 月达 14 亿美元。
- **Mercor:** 2025/10 估值 100 亿,较 2 月翻 5 倍;管理 3 万+ 专家。

连锁反应:中立性即资产

Meta 入股 Scale AI 后,因数据机密性顾虑,Google、OpenAI、xAI 等削减或暂停与 Scale 的合作,为 Surge、Mercor 让出空间——印证数据供应行业「中立性」本身即核心资产。

来源:Bloomberg、CNBC、TechCrunch、Sacra、PitchBook。

从「抓取」到「付费授权」

当公开语料枯竭、版权诉讼频发,模型厂商转向授权协议。据 Media & the Machine 追踪,早期 34 笔交易总承诺约 **29.2 亿美元**(约 8.16 亿/年)。

OpenAI – News Corp(2024/5)	5 年最高 2.5 亿美元 · 出版史最大
Reddit – Google	约 6,000 万美元 / 年
Reddit – OpenAI	估约 7,000 万美元 / 年
Shutterstock – 多家	2024 AI 授权营收约 1.04–1.38 亿
News Corp – Meta(2026/3)	5,000 万/年 × 3 年 = 1.5 亿
Disney – OpenAI(2025/12)	授权角色 + 入股 OpenAI 10 亿

来源:Media & the Machine、Yahoo Finance、Bloomberg。部分金额为推算。

Reddit:数据资产价值重估

据 Reddit 2025 Q2 股东信引用 Profound(分析 40 亿+ 次 AI 引用):在截至 2025/6/30 三个月中,Reddit 占全部 AI 引用 **3.11%**,为 Wikipedia(1.35%)两倍多——成为 AI 模型第一大被引来源,推动「固定费→使用量→动态定价」演进。

一张图看懂全球数据玩家版图

美国前沿数据公司估值已达**软件级**,中国玩家则多为「盈利但体量小」的上市/挂牌企业。

美国 · 前沿数据公司

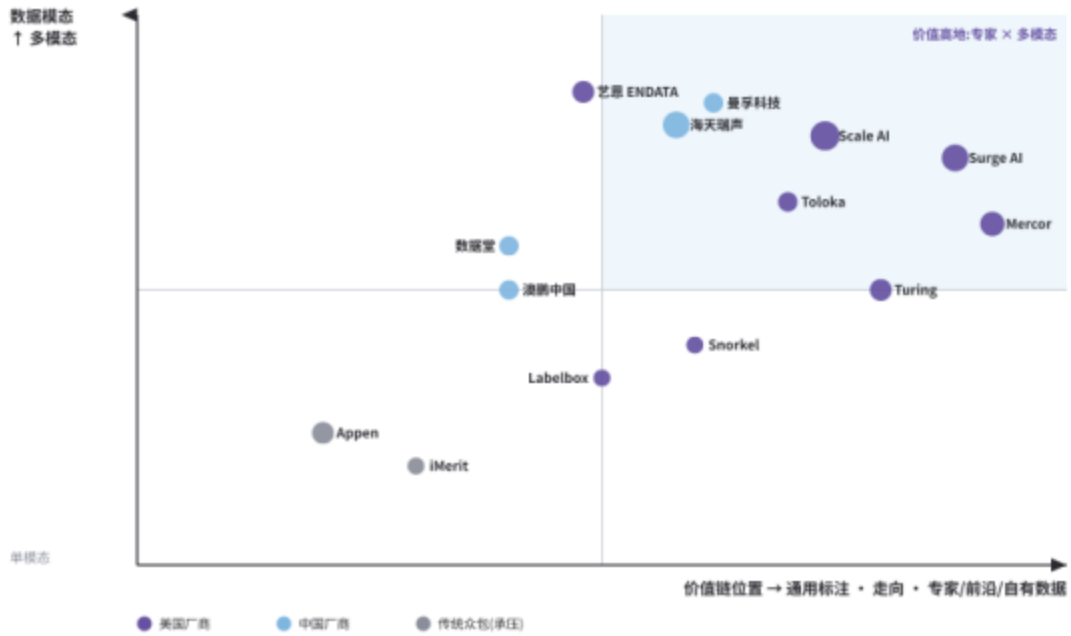
Scale AI	2025/6 Meta 投 143 亿$ · 通用标注+RLHF+政企	≈290 亿$
Surge AI	2025/7 融资洽谈中(未完成) · 高端 RLHF · 自举	≈250 亿$(洽谈)
Mercor	2025/10 C 轮 3.5 亿$ 已完成 · 专家人才市场/RL	≈100 亿$
Turing	2025/3 E 轮 1.11 亿$ · 工程/编程专家数据	≈22 亿$
Snorkel AI	2025/5 D 轮 1 亿$ · 程序化数据+评测	≈13 亿$

中国 · 数据/标注厂商

海天瑞声	2025 营收 ¥3.77 亿(+59%) · 语音+多模态	A股 688787
数据堂	2025 营收 ¥3.62 亿(+49.2%) · 拟北交所辅导	新三板 831428
澳鹏中国 Appen	2025 中国营收 ≈¥7.3 亿(+74.8%) · 多模态标注	ASX:APX 中国区
曼孚科技	2026/4 数亿元(五源资本领投) · AI数据生成/标注	Pre-C 轮
艺恩 ENDATA	2025 营收 +49.86% · 垂类视频多模态	新三板 871430

来源:各公司 2025 年报/公告(NEEQ/STAR/ASX)、Bloomberg、36氪。

价值前沿,正向「专家 × 多模态」迁移



三大梯队

- **价值高地(右上):** Surge、Mercor、Scale、Turing、Toloka——以专家人工判断与多模态数据获取软件级溢价。
- **平台/工具(中部):** Labelbox、Snorkel——程序化标注与评测工具。
- **传统众包(左下·承压):** Appen、iMerit——通用单模态标注被自动化与合成数据挤压。

中国厂商卡位

海天瑞声(语音/多模态·已上市)、澳鹏中国/曼孚科技(自动驾驶·标注)、艺恩(垂类数据)正集中卡位**多模态与垂直**象限,与全球价值迁移方向一致。

四条赛道,四个范式样本

① 通用标注 / RLHF

Scale AI

营收 3 年 ×11

2021 年 8,000 万→2024 年 8.7 亿美元;Meta 143 亿入股引发中立性裂变,客户外流反成对手红利。

② 专家 / 前沿数据

Surge AI / Mercor

ARR 14 亿 / 4.5 亿$

Surge 零融资自举至超 10 亿营收;Mercor 3 万专家、日付 150 万$、时薪 ~85$——瞄准 5 万亿$ 知识劳动市场。

③ 内容授权

Reddit

AI 引用 #1

18 个月市值至 ~390 亿$;Google 6,000 万/年 + OpenAI 7,000 万/年;引用量达 Wikipedia 3 倍,首创动态定价。

④ 垂直 / 具身多模态

智元 AgiBot World

百万真机片段

开源百万级真机数据集,长程任务规模约为 Google Open X-Embodiment 的 10 倍、场景覆盖 100 倍;中国具身数据供应链成型。

范式启示

价值正从「规模化通用标注」流向「专家判断、独家授权与具身多模态」。谁掌握稀缺合规数据,谁就掌握定价权。

来源:Inc.、CNBC、Bloomberg、澎湃新闻、智元机器人。Surge 营收为第三方追踪,Scale 2025 营收为预测。

03

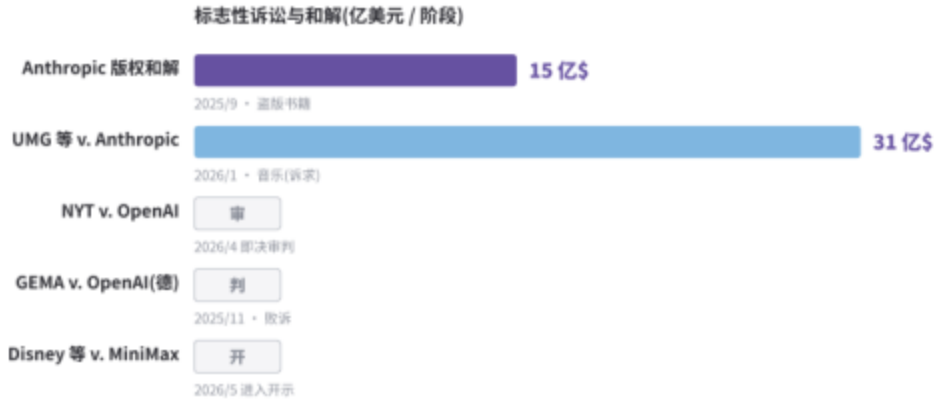
PART 03 · COMPLIANCE & REGULATION

合规与监管

版权诉讼 · 出海风险 · 欧盟 AI 法案

从诉讼频发,到「合规即护城河」

截至 2025 年 10 月,全球追踪到的 AI 版权诉讼达 **51–166 起**。法院核心分野正在形成:「合法获取」可能构成合理使用,「盗版内容」则明确不被宽宥——这直接抬高合规数据溢价。



核心启示

美国已有 3 位法官就 AI 训练合理使用裁决(2 支持训练方、1 反对),均强调训练「高度转化性」,但严格区分内容**获取来源是否合法**。合规由此从「成本项」转为「定价项」。

来源:AI Lawsuit Tracker、McKool Smith、OpenAI 及各法院文书。

海外视频版权风险 · MiniMax / 海螺案

Disney 等 v. MiniMax

- **原告:** 迪士尼、环球、华纳兄弟探索等 12 家(2025/9/16,加州中区联邦法院)。
- **指控:** 大规模盗用版权角色;复诉引述海螺自标榜「口袋里的好莱坞工作室」。
- **诉求:** 每件作品最高 15 万美元法定赔偿 + 禁令。
- **最新:** 2026/5/23 法官驳回撤案动议,进入证据开示(Reuters)。

提示:此为程序性裁决,尚未认定侵权,无赔偿/和解;15 万美元为法定上限索赔额。

中国 AIGC 版权裁决 · 态度已确立

- **北京互联网法院李某案(2023/11):** 首例 AI 生成图片版权案,认定可版权性,判赔 5,000 元。
- **广州互联网法院奥特曼案(2024/2):** 全球首例 GenAI 输出侵权裁决,判赔 1 万元,要求关键词过滤。
- **杭州互联网法院 LoRA 案(2024):** 认定平台帮助侵权判赔 3 万元;区分训练与生成阶段责任。

来源:Reuters、Variety、Wolters Kluwer、KWM 等。

训练数据透明度,成为硬约束

欧盟《AI 法案》2024/8/1 生效,GPAI 义务 2025/8/2 适用,2026/8/2 全面适用。其对训练数据的透明度要求,正把「合规」从自愿变为**法定义务**。

ARTICLE 53(1)(d)

训练内容摘要披露

须按 AI Office 模板公开训练内容「充分详细摘要」(含受版权数据),披露数据类型、来源及公开数据集前 10% 域名。

CODE OF PRACTICE

GPAI 行为准则

2025/7/10 发布,含透明度、版权、安全三章;系统性风险门槛 10^{25} FLOP(全球约 5–15 家公司适用)。

TRANSITION

过渡安排

2025/8 前已上市模型,有至 2027/8/2 宽限期;模板自 2025/7/24 起强制使用。

合规如何转化为定价能力

诉讼频发、监管趋严下,可审计、可溯源的**合规授权数据**获结构性溢价。内容/肖像/音乐版权的多层授权链条与可披露来源证明,已成为高端供应商区别于「爬虫式」供给的核心壁垒——合规转化为**议价筹码**。

来源:European Commission、Jones Day、Skadden、WilmerHale 等。

04

PART 04 · GLOBAL · US-CHINA · OUTLOOK

全球格局 · 中美双核 · 未来

多模态前沿 · 中美生态 · 趋势判断

最稀缺、溢价最高的层级

当文本语料见顶,竞争前沿转向视频与多模态。视频生成与世界模型对「高质量、合法授权、富标注」语料的渴求,使这一层级成为整个数据市场中**最稀缺、单位价值最高**的部分。

视频生成:竞争白热化

OpenAI Sora 2、Google Veo 3、快手可灵、Vidu、海螺、Runway 同台竞技。Google Veo 的 4K 真实感与原生音频优势,直接来自 **YouTube** 训练数据——印证独家高质量视频语料的决定性价值。

「黄金数据」稀缺

可灵团队论文明确指出:视频生成严重依赖**同时具备高视觉质量(VQ)与高运动质量(MQ)的「黄金数据」**;此类数据集稀少且获取昂贵,是规模化的主要限制。

4D / 多视角空间数据与世界模型

世界模型(NVIDIA 定义:理解真实世界动态、含物理与空间属性的生成式 AI)面临**「配对多视角数据严重稀缺」**。具身数据「稀缺、采集困难、高维」,被视为机器人达到「GPT 时刻」的关键瓶颈。

中国进展

合成灵巧抓取数据集 DexGraspNet 2.0 达 10 亿规模;头部具身公司部署百台机器人,日产真机数据上万条(信通院)。路径正从「真机采集」走向「合成引擎 + 真机精调」。

来源:NVIDIA、Kling(arXiv)、信通院等研究。

「数据要素」国家战略驱动的独特生态

120亿元
信通院测 2024 中国数据标注产业规模,核心企业超 600 家

7
已建成国家数据标注基地,赋能 121 个国产大模型(商务部)

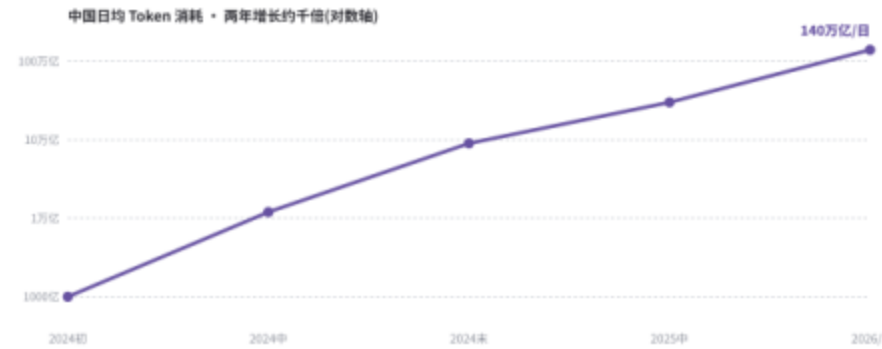
140万亿
日均 Token 消耗(2026/3,国家数据局),较 2024 初增千倍

7.5万亿
2030 年我国数据产业规模预测(国家数据局,元)

三大政策支柱

- **「数据要素 ×」三年行动(2024–2026):** 聚焦 12 行业,2026 年底打造 300+ 示范场景,数据产业年均增速超 20%。
- **数据资源入表:** 财政部《暂行规定》2024/1/1 施行,数据从费用化转为资产负债表内显性化。
- **数据标注产业专项:** 国家发改委 2024/12 实施意见,目标 2027 年产业规模大幅跃升。

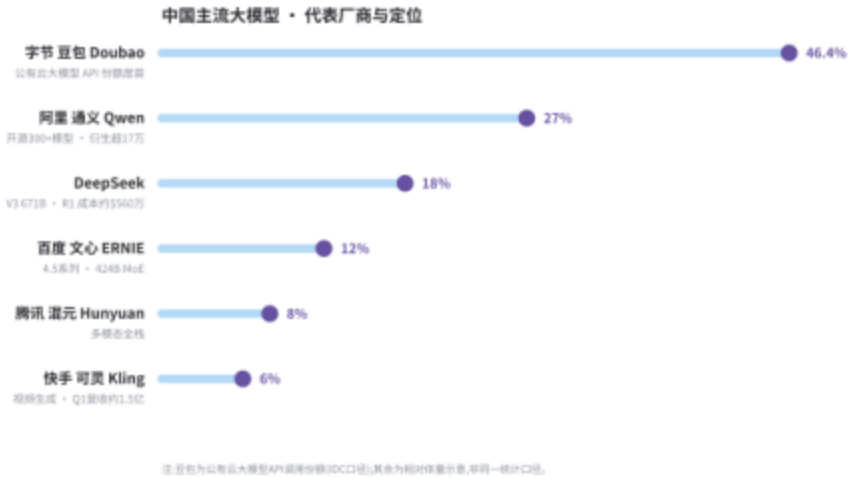
Token 消耗爆发



来源:国家发改委、财政部、国家数据局、商务部、信通院公开文件。

大模型数量全球第一

中国已发布大模型 **1509 个**、数量居全球首位(占全球约 40%),AI 企业超 **5,300 家**(中国信通院,2025)。在开源社区 HuggingFace 趋势榜上,2025 年 7 月前 10 名**开放权重**模型中国占 9 席(智谱 GLM、阿里 Qwen 等);GPT、Claude、Gemini 等闭源模型不在该开源榜内。



注:豆包为公有云大模型API调用份额(IDC口径),其余为相对体量示意,非同一统计口径。

市场体量

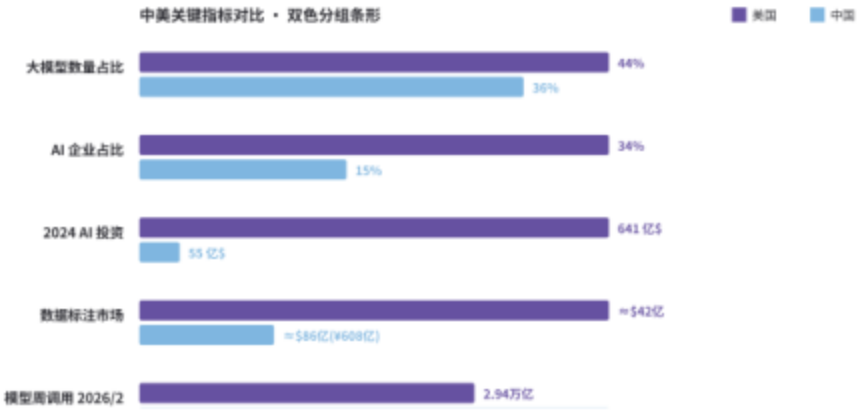
- 据信通院测算,中国 AI 大模型相关市场规模快速扩张,垂类应用加速落地。
- DeepSeek V3 671B 参数,末次训练成本约 **557.6 万美元**(不含前期研发)。
- Omdia 预测中国生成式 AI 市场到 2029 年达 **98 亿美元**。

竞争特征

区别于美国「少数寡头」格局,中国呈**「多头并进」**——通用、视频、多模态多赛道

两套逻辑,两种数据生态

中美数据市场由**两套不同逻辑**驱动:美国前沿实验室拉动、专家数据溢价、软件级估值;中国「数据要素」国家战略+垂类模型驱动。



美国 · 前沿驱动 前沿实验室拉动 + 专家数据溢价 + VC 催化;头部数据公司估值高出中国约 **1–2 个数量级**。

中国 · 战略统筹 「数据要素」国家战略 + 7 大标注基地 + 数据资源入表;要素流通市场 2024 约 **¥1,662 亿**。

来源:Stanford HAI、信通院、iResearch、OpenRouter、GVR(完整来源见末页)。

全球趋势:从「规模」到「质量与专业化」

① 合成数据走向主导(前瞻)

Gartner 预测合成数据占比将**于 2030 年全面超越真实数据**。

② 专家/前沿人类数据崛起

RL 环境、agentic data、可验证奖励成新焦点,专家数据公司估值飙升印证。

③ 数据飞轮成护城河

「交互→数据→模型改进」闭环成最难复制壁垒。

④ 从规模转向质量专业化

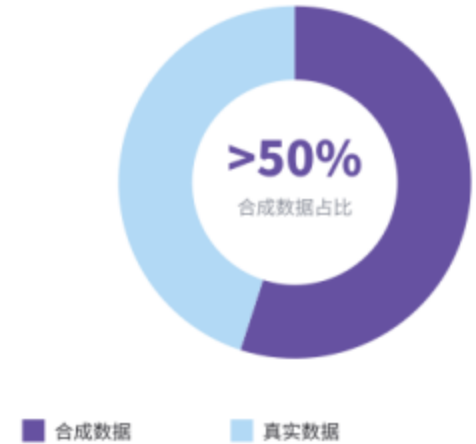
垂直领域数据溢价上升:要的不是更大模型,而是更专业数据。

⑤ 具身 AI 与世界模型成增长极

世界模型既消耗数据,也生成未来训练的合成数据。

⑥ 中美双核 · 两套路径

美国前沿实验室+专家数据溢价;中国数据要素战略+垂类落地。



Gartner:2030 合成数据将超真实数据

结语

胜负手正从「更多算力」转向「更优质、更合规的数据」;公开语料枯竭非终点,而是价值化的起点。

来源:Gartner、Omdia 及前述各章。「预测/前瞻」为前瞻性推演。

05

PART 05 · ENDATA PRODUCTS

艺恩核心产品

数据集业务 · Video Feeds · ENBASE 数据魔方

高质量、合规、垂直的数据弹药库

服务头部通用大模型、垂直多模态模型、AI Infra 与 MaaS 平台。提供数据集授权 + 全链路定制:清洗 · 结构化 · 标注 · 向量化 · 合规审计 · 场景化定制。

PRE-TRAINING

预训练数据集

TB 级多语种垂类语料

120+ 语言方言

ALIGNMENT

SFT / RLHF

高质量指令对 · 多轮 · CoT 与偏好对齐

偏好对齐

MULTIMODAL

多模态对齐

视频 · 图像 · 文本三模态自然对齐语料

V-T / I-T / A-V

CUSTOM

定制化数据

按训练目标的端到端处理与交付

私有化 · 合规

VIDEO FEEDS · 面向视频原生 AI 与具身智能

以「影视综+社媒+电商」累积全球级视频资产,配套元数据 Schema 与多任务标签,为视频生成、理解、世界模型与 VLA 训练提供持续、合规的数据流。 **2.3B+ 视频片段沉淀 · 800TB+ 日均交付带宽 · 120+ 任务族覆盖。**

ENBASE 数据魔方,AI 数据副驾驶

不再只是查数工具——enbase 是客户决策层桌面上的 AI 数据副驾驶。自然语言发起复杂查询,同步返回结构化数据、关联视频片段、趋势图谱与生成式洞察摘要。

01 · AI 检索

对话式发现

自然语言对话式发现,重构数据接入方式。

02 · 生成式分析

洞察自动生成

洞察摘要 · 趋势归因 · 对比解读自动生成。

03 · 三模态调取

一站式调用

视频·图像·文本一站式调用与关联检索。

与客户同行的三条承诺

QUALITY · 只交付「对的数据」

多维质量评估 · 合规过滤 · 人工复核,共同构成交付质量底线。

TRANSPARENCY · 透明可追溯

每条数据有清晰来源链路与授权状态,可审计、可溯源、可验证。

PARTNERSHIP · 长期伙伴

共同规划数据路线图,持续迭代、持续优化、共同进步。

可核实、区间呈现、区分事实与预测

本白皮书基于公开可核实的权威研究机构报告、监管文件、主流财经媒体与学术论文综合编撰。所有市场规模数据均标注来源与年份;口径分歧者以区间呈现;前瞻性预测均以「预计/预测」标识。本白皮书为行业研究,不构成投资建议。

01 Epoch AI — Will we run out of data?(ICML 2024)

02 Stanford HAI — 2025 AI Index Report

03 MarketsandMarkets / Grand View Research / Straits Research / FBI / GMI

04 Precedence Research / The Business Research Company / Technavio

05 Bloomberg / CNBC / TechCrunch / Reuters

06 Sacra / PitchBook

07 Media & the Machine / Search Engine Land / Yahoo Finance

08 AI Lawsuit Tracker / Sustainable Tech Partner / McKool Smith

09 European Commission / Jones Day / Skadden / WilmerHale

10 Wolters Kluwer / Linklaters / KWM / Digital Watch

11 信通院 / 国家发改委 / 财政部 / 国家数据局 / 商务部

12 信通院 / 头豹研究院 / 沙利文 / 人民日报 / IDC

13 arXiv(Kling、TesserAct、Embody4D 等) / Gartner / Omdia

让数据,成为 AI 时代的*价值坐标*

无论你是大模型厂商寻找「对的数据」,还是出海品牌寻找「懂中国也懂全球」的数据伙伴——艺恩都已准备好下一程。

BUSINESS · 商务合作

cs@endata.com.cn

TEL · 电话

+86-010-85899985

SERVICE · 服务热线

400-052-9966

北京艺恩世纪数据科技股份有限公司 · NEEQ 871430 · 北京 / 杭州 / 上海 / 海外 CA

www.endata.com.cn

全球大模型数据市场白皮书

2026 年版 · 第一版

艺恩数据
ENDATA

产品矩阵 · 2026

PRODUCT PORTFOLIO · 三条产品线

三条产品线
一个数据底座

艺恩数据以**数据集业务 / AIDATA / enbase 数据魔方**三条产品线，覆盖大模型训练数据、AI 专项数据集与营销决策 SaaS 的全场景需求。

数据集业务 | AIDATA 专线 | enbase SaaS | 全链路合规

— DATASETS · 数据集业务

高质量 · 合规 · 垂直
数据弹药库

服务大模型厂商与 AI Infra 平台。**4,000+ 成品数据集**，三模态全覆盖。

4000+	3	4	100%
成品 SKU	模态覆盖	垂直领域	授权合规

视频数据集

VIDEO DATASETS

剧集 · 短视频 · 直播全品类，VLA Ready，支持 RLDS / LeRobot v3 格式。

VLA Ready　RLDS　LeRobot

文本数据集

TEXT DATASETS

对话 SFT / 偏好 RLHF / RAG 知识库，覆盖多场景，支持 30+ 语言。

SFT　RLHF　RAG　多语言

图像数据集

IMAGE DATASETS

海报 · KOL · SKU · 电商多模态对齐，美妆 / 服饰 / 食饮 / 3C 全品类。

多模态对齐　高精度标注

版权合规体系

COPYRIGHT CHAIN

三层版权审计，每份数据集附完整授权链路文件，满足 ISO 27701 要求。

ISO 27701　三层合规

AIDATA · AI 数据专线

面向 LLM / 具身智能
出海 AI 的专属数据

为大模型厂商、具身智能公司、跨境 AI 平台提供**高质量合规内容数据集**，覆盖海外电商与社媒数据。

艺恩 AIDATA 专线

以**海外电商 + 社媒数据集**为核心产品，对接国际大模型厂商与出海 AI 公司的训练数据需求，涵盖跨境电商及海外主流社媒平台，30+ 语言覆盖。

- **LLM 预训练语料**
  TB 级中文/多语言语料，去重 · 去噪 · 质检全流程
- **VLA 具身智能数据**
  动作序列 + 语言指令对齐，RLDS / LeRobot 格式
- **海外电商数据集**
  跨境平台多语种语料，支持 SFT / RAG 场景
- **社媒数据集**
  海外主流平台内容，30+ 语言，含情感标注

SAAS · ENBASE 数据魔方

从"查数工具"升级为 AI 数据副驾驶

enbase v3.0 以自然语言为入口，整合三模态数据调取与 AI 分析，是决策层的**一站式数据工作台**。

enbase 数据魔方 v3.0 · AI Co-pilot

把数据资产 封装为决策入口

无需 SQL，自然语言直接返回多维数据洞察；视频 / 图像 / 文本三模态一键调取；内置品牌声量实时监测与竞品追踪。

AI 自然语言检索
输入问题直接返回洞察

三模态调取
视频 · 图像 · 文本

品牌声量追踪
实时监测竞品动态

API 开放接入
私有化部署支持

国家高新技术企业 · 北京市专精特新 · ISO 20000 · 27001 · 27701 · AI 数据标注能力评估 · 数据安全管理认证