

产业深度研究

AI 产业深度分析

产业链全景 · 代表性公司 · 应用现状与前景

核心判断：价值当前高度沉淀于算力与云两层；
未来 3-5 年将部分向「数据 + 分发 + 切换成本」型应用迁移。

王子秋

数据截至 2026 年年中 | 2026 年 7 月编制

应用层

编程 · 办公 · 消费 · 垂直行业

模型层

闭源双雄 · 开源生态

云与基础设施层

超大规模云 · 新云 · 电力

算力硬件层

GPU · 先进封装 · HBM

三个核心判断



01

利润高度集中于算力与云

7,250 亿

四大云厂商 2026 年合计资本开支

英伟达单季数据中心收入 752 亿美元
(+92%)，毛利率约 75%；模型层收入爆发但仍深度亏损。



02

Token 价格每年约 10 倍下降

\$20 → \$0.40

GPT-4 级 / 百万 token (2022 末→2026 初)

利好应用层扩张、压缩模型层毛利；推理量随智能体爆发，「量补价」反而巩固算力层。



03

最大风险：Capex 与 ROI 背离

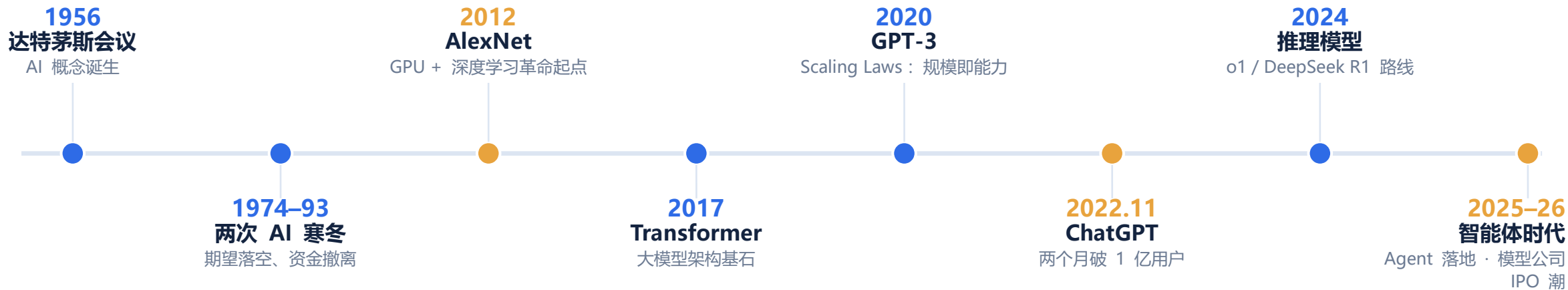
5%

MIT：仅 5% 企业试点产生百万美元级回报

「英伟达—OpenAI—Oracle—新云」循环融资结构积累系统性风险；但云在手订单超 1.4 万亿美元，需求目前仍真实。

基准判断：非 2000 年式纯泡沫，但 capex- 收入背离是真实系统性风险 —— 结局取决于企业 ROI 兑现速度与信贷环境。

七十年三次浪潮：从符号主义到智能体时代



符号主义与传统机器学习

1956 – 2012

深度学习时代

2012 – 2022

生成式 AI 与智能体时代

2022 – 今

* 轴距为示意, 非等比时间刻度

四层价值链：利润当前集中于底部两层



英伟达：产业最确定的价值捕获者

816 亿

FY2027Q1 总营收（美元，同比 +85%）

752 亿

单季数据中心收入（+92%，其中网络 148 亿）

≈75%

毛利率 —— 定价权的直接体现

≈80%

AI 加速器市场份额；超大规模客户约占 DC 收入一半

从游戏 GPU 到 AI 基石

- 2006 推出 CUDA，二十年构筑软件生态护城河
- 2012 AlexNet 后全面转向 AI；2023 起成为生成式 AI 最大赢家
- Rubin 架构 2026 下半年推进；新批 800 亿美元回购授权



护城河

- CUDA 生态：开发者与框架的二十年锁定
- NVLink / 网络：系统级（而非单芯片）竞争
- CoWoS 与 HBM 的产能优先分配权（锁定台积电 CoWoS 约 60%）



风险清单

- 中国数据中心收入归零（H20 出口管制，曾计提 45 亿美元）
- 循环融资敞口：向 OpenAI 承诺至多 1,000 亿美元投资
- 定制 ASIC 替代：2026 ASIC 出货增速 44.6% vs GPU 16.1%
- 估值高度依赖「完美增长」持续兑现

2026 年反转：Anthropic 收入与估值双双反超

Anthropic

运行率（2026 年中）

≈470 亿美元

估值（2026.5，报道口径）≈ 9,650 亿美元

收入结构 企业 / API 约 80%；Claude Code 运行率 ≈ 25 亿美元

单位经济 月活人均收入 ≈ \$192；付费转化率 ≈ 13%

客户 8 家财富 10 强在 Claude 上运行关键负载

里程碑 2026.5 Series H 融资 650 亿美元；6.1 递交保密 IPO

VS

OpenAI

运行率（2026 年中）

≈250 亿美元

估值（2026.3，报道口径）≈ 8,520 亿美元

收入结构 ChatGPT ≈170 亿 + API ≈65 亿 + Sora/ 授权 ≈ 15 亿

用户规模 周活 ≈ 9 亿、月活破 10 亿 —— 消费分发之王

盈利状况 Q1 2026 运营利润率 ≈ -122%（每 \$1 收入亏 \$1.22）

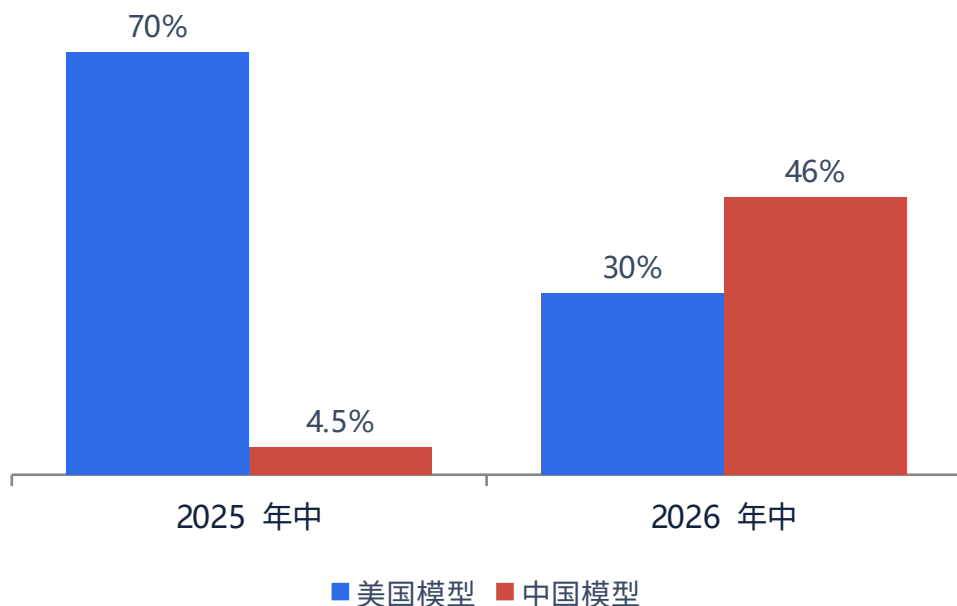
里程碑 2026.6.8 递交保密 IPO，目标估值超 1 万亿美元

两家均深度亏损、均于 2026 年 6 月递交保密 IPO —— 商业模式的终极检验从一级市场移交二级市场。

来源：CNBC (2026.5.28)；Sactra；The Information —— 私有公司数据为报道口径，非审计

中国路线：极致效率 + 开源攻占全球份额

OpenRouter 企业 token 份额 (%)



* 46% 为单周峰值; 美国模型份额一年内 70% → 30%



DeepSeek

R1 以 ≈ 560 万美元训练成本对标 o1 (MoE 671B/ 激活 37B) ; 2025.1.27 触发英伟达单日 -16.9%、抹去 5,890 亿美元市值



阿里通义 Qwen

开源生态之王: 衍生模型 17 万 + ; 国内企业 token 消耗份额 32.1% 居首 (2025 下半年)



字节豆包

分发取胜: 国内份额 21.3% ; 除夕峰值处理 633 亿 token / 分钟



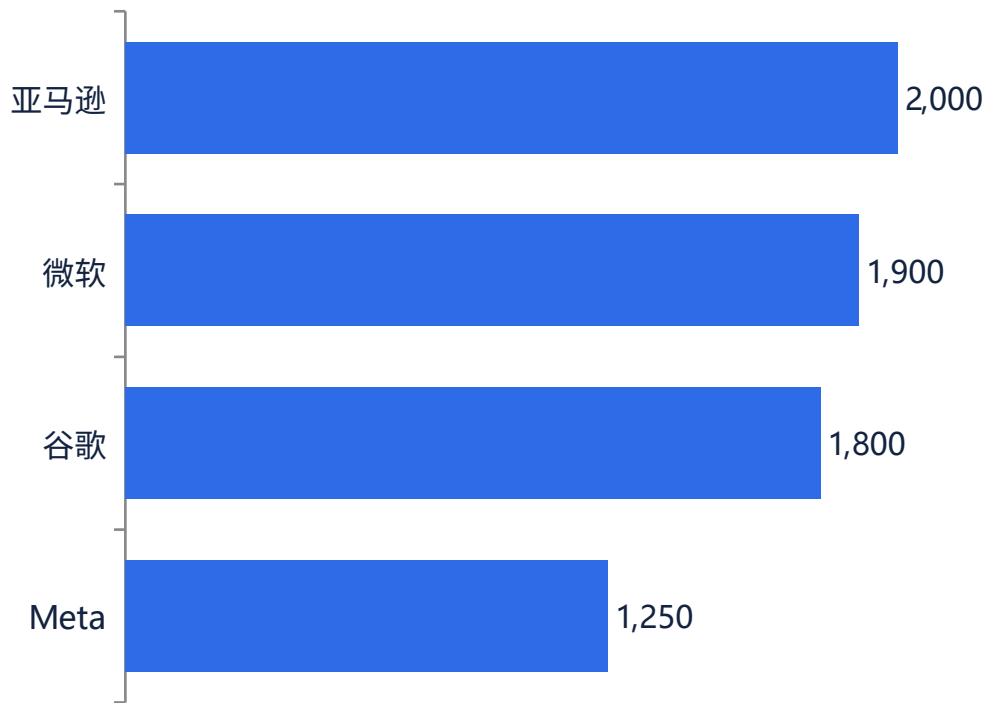
Kimi / 智谱

Kimi K2.5 成「最快十角兽」(估值 100-120 亿美元) ; GLM-5 开源、编程需求过载限售

含义: 开源 + 低价 commoditize 基础层, 直接压低全球 token 价格、冲击闭源商业模式。

2026 资本开支军备赛：四巨头合计约 7,250 亿美元

2026 年资本开支指引 (亿美元)



* 谷歌 / Meta 取指引区间中值 (1,750–1,850 / 1,150–1,350)

在手订单 (RPO / Backlog) —— 需求可见度

Oracle	6,380 亿	OpenAI 约占 54% (3,000 亿 / 5 年合约)
微软	6,270 亿	商业 RPO , 同比 +99%
谷歌云	4,600 亿	较上季约翻倍, 「近期算力受限」
AWS	3,640 亿	年化运行率 1,500 亿美元



苹果：端侧路线的观察对象

Apple Intelligence 进展相对滞后；AI 手机 / PC 换机周期能否启动，是端侧叙事的关键验证点。微软 AI 年化收入 $\approx$ 370 亿美元、Copilot 席位 2,000 万 + (渗透仅 3–4%) 。

瓶颈迁移：从芯片，到封装、存储与电力



先进封装 CoWoS

- 台积电月产能 12.5 万片 (+79%) 仍满产
- 英伟达锁定约 60% 产能; 2nm 制程已订满
- 封装 = 比晶圆更紧的硬约束



HBM 高带宽存储

- 2026 年三家产能全部售罄; 市场 546 亿美元 (+58%)
- 份额: SK 海力士 $\approx 53\%$ · 三星 $\approx 35\%$ · 美光 $\approx 11\%$
- HBM4 于 2026.2 量产, 成本占比持续抬升



电力：最终约束

- 数据中心用电 460 TWh (2024) $\rightarrow \approx 1,300$ TWh (2035E)
- PJM 并网排队超 2,600 GW、等待 5 年以上
- 电力已取代芯片成为扩张的第一瓶颈

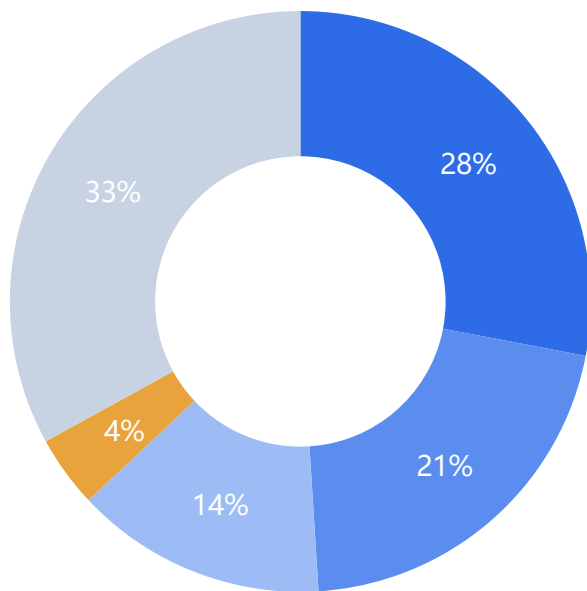


核电回归 (含 SMR)

- 四大云厂商签约核电近 10 GW
- 微软 20 年 PPA 重启三里岛 (835MW, 2028 目标)
- 亚马逊 X-energy · 谷歌 Kairos · Meta 1-4GW RFP

云：三巨头格局稳定，新云靠 AI 订单爆发

全球云基础设施份额（2026Q1, Synergy）



■ AWS ■ Azure ■ Google Cloud
■ Oracle ■ 其他

三巨头合计 63%



CoreWeave —— 新云样本

2025 营收 51 亿美元（+170%）；Q1 2026 backlog 达 994 亿美元；微软约占收入 67% —— 高增长与高集中并存。



Oracle —— AI 订单的极端集中

OCI 增速 +93%（FY26Q4）；RPO 6,380 亿美元，其中 OpenAI 约占 54%（3,000 亿 / 5 年合约）—— 单一客户风险显性化。



中国云：国产算力 + 开源协同

阿里云 / 腾讯云 / 华为云依托昇腾等国产算力与 Qwen/GLM 开源生态，走「性价比 + 主权云」路线。

利润池地图：通缩的 Token，昂贵的算力

利润之王

算力硬件层

英伟达毛利 $\approx 75\%$ ；HBM/CoWoS 全线售罄

现金牛

云与基础设施层

高利润但重资本；backlog 超 1.4 万亿美元

烧钱换地盘

模型层

收入爆发但深亏（OpenAI 运营利润率 $\approx -122\%$ ）

分化

应用层

毛利被 API 侵蚀：Cursor 毛利仅 30–50%，每 \$1 收入付 \$0.40–0.70 给模型

Token 价格通缩

\$20 $\rightarrow$ \$0.40

GPT-4 级能力 / 百万 token（2022 末 $\rightarrow$ 2026 初）

- 每年约 10 倍下降；含效率提升最高 ≈ 200 倍 / 年（Epoch / a16z）
- 推理已占 AI 计算约 2/3；智能体单任务 token 量 $\times 5\text{--}50$
- 结果：「量补价」—— 通缩反而巩固算力层

基准判断：未来 3–5 年利润向应用层渐进迁移，但只有握有专有数据、分发渠道与切换成本的应用能真正捕获；算力仍是全周期最确定的赢家。

波特五力：四层竞争格局一览



壁垒极高

算力硬件层

- 近垄断：CUDA 生态 + 产能优先分配权
- 替代威胁上升：云自研 ASIC 增速 44.6%
- 买方（云厂商）议价力随自研增强



寡头稳固

云与基础设施层

- 资本壁垒极高，三巨头 63% 份额
- 对上游英伟达议价弱、对下游客户强
- 新云以「专用 AI 云」切入缝隙市场



竞争最烈

模型层

- 开源替代 + 价格战，护城河最脆弱
- 能力上移不断吞噬下游套壳应用
- 被迫向应用 / 智能体延伸找利润



两极分化

应用层

- 进入壁垒低：套壳易、模仿快
- 分发 + 数据 + 切换成本造就局部垄断
- 价值向 workflow 深度集成者集中

应用渗透分层：编程一骑绝尘，企业 ROI 仍是短板



企业采用的现实：广而不深

88%

组织已常态化使用 AI (McKinsey, 2025.11)

≈ 7%

实现完全规模化; 高绩效者仅 ≈ 6%

5%

试点产生百万美元级回报 (MIT NANDA) —— 尽管已投入 300-400 亿美元

失败根因在数据与集成，非模型：外部供应商主导的部署成功率约 67%，企业自建仅约 1/3。

未来 3-5 年的五条主线

01



智能体 Agent

从对话到自主执行任务；单任务 token 量 $\times 5-50$ ，形成算力需求飞轮 —— 本轮周期的主叙事。

02



多模态与视频生成

Sora · 可灵 · Gemini Omni；能力快速成熟，货币化仍处早期。

03



具身智能与机器人

2025 年融资超 340 亿美元；Figure 估值 390 亿、宇树申报上交所 IPO、PI 推 π 模型 —— 试点密集，仍未到「ChatGPT 时刻」。

04



AI for Science

AlphaFold 路线延伸至药物发现与材料科学；长周期、高壁垒、与算力层深度绑定。

05



端侧 AI 与换机周期

AI 手机 / AI PC 能否触发换机潮，决定苹果、高通与整个消费电子链的 AI 叙事成色。

谁受益、谁承压：影响评估矩阵

评估四要素： workflows可自动化程度 · 专有数据壁垒 · 分发优势 · 客户切换成本



受益方

- 算力与电力产业链：英伟达 · SK 海力士 · Vertiv · 核电 /SMR
- 云厂商：AWS · Azure · GCP · Oracle · 新云
- 「数据 + 分发」型应用：微软 · Palantir · 企业侧 Anthropic 生态

共同点：位于供给瓶颈，或握住不可替代的客户资产



转型中

- Adobe 与传统软件大厂：看 AI 化速度与定价权保卫战
- 按席位定价的 SaaS：面临「按结果定价」的模式重构
- IT 服务商：从人力堆叠转向 AI 交付，利润率承压

关键变量：自我革命快于被革命



承压方

- 劳动密集型 IT 服务与外包、内容代工
- Chegg：股价较 2021 年峰值约 -99%；2025 两轮裁员共 636 人（第二轮 45%）
- Stack Overflow：月提问量较峰值下滑逾 90%

共同点： workflows被模型直接替代，且无数据反哺闭环

判别式：被 AI 增强（有数据闭环、握分发）还是被 AI 替代（纯执行环节、可被直答绕过）。

五个必须正面回答的问题



① 泡沫论 vs 早期论

- 正方：云 backlog 超 1.4 万亿美元、企业现金流为 1999 年三倍
- 反方：循环融资（英伟达→OpenAI 至多千亿·微软持股 27%·OpenAI 持 AMD 10%）+ MIT「仅 5% 有回报」
- 判断：非 2000 式纯泡沫，但背离是系统性风险



② 电力与产能瓶颈

- PJM 并网排队超 2,600 GW、等待 5 年以上
- 电力已取代芯片成为扩张的最终约束
- 核电 / SMR 是 2028 后才兑现的远水



③ 中美芯片管制

- 英伟达中国数据中心收入已归零
- 昇腾 910C 目标 60 万颗、政府采购去英伟达化
- 中国以「横向扩展 + 国产替代」应对，受 HBM 约束



④ 开源对闭源的冲击

- 中国开源模型企业 token 份额单周峰值 46%（一年前 4.5%）
- 基础层被 commoditize，价格战常态化
- 闭源被迫向应用、企业信任与智能体转移价值



⑤ 监管走向

- 欧盟：GPAI 规则 2025.8 已生效；高风险条款推迟至 2027-28
- 美国：2025.12 行政令推动联邦框架优先、挑战州法
- 中国：AI 内容标识生效；已备案 796 个生成式服务

2026–2030：乐观 / 基准 / 悲观三情景



乐观

智能体兑现

- Agent 规模落地，企业 ROI 转正、可衡量回报比例显著上升
- Capex 被真实收入消化；AI 资本开支向数万亿美元级演进
- 利润健康地向应用层迁移，形成新一代平台公司

含义：应用层与全栈平台受益最大



基准

增长换挡

- 需求延续但增速放缓；利润仍集中于算力与云
- 部分高杠杆新云与同质化初创出清
- 监管温和收紧；开源与闭源长期并存

含义：「卖铲人」与数据 / 分发型应用最具韧性



悲观

背离兑现

- 信贷收紧触发循环融资逆转，capex 急刹
- 至少一家大型基础设施玩家爆雷（Fink 式警告成真）
- 估值大幅回调 —— 但底层需求不归零，区别于 2000 年

含义：高杠杆基础设施首当其冲

切换信号：2027 云厂商 capex 若继续加速（分析师预测破 1 万亿美元）→ 乐观 / 基准；若集体下修 → 悲观概率显著上升。

价值沉淀路径与跟踪清单

短期 • 0-12 月

确定性最高的是「卖铲人」：算力（英伟达 / ASIC 链）、HBM、电力与散热 —— 需求可见度最高的环节。

中期 • 1-3 年

应用层价值归于「专有数据 + 分发 + 高切换成本」型应用与平台；通用套壳与被搜索直答吞噬的内容型公司最易受冲击。

长期 • 3-5 年

盯住 ROI 兑现：利润真迁移则价值向应用层转移；capex- 收入持续背离则高杠杆基础设施首当其冲。

核心结论：本轮周期中，价值最先、最厚地沉淀在算力硬件与云两层；应用层的赢家将由数据、分发与切换成本决定。



高频跟踪指标

- 四大云厂商季度 capex 与 2027 指引（是否破 1 万亿）
- 英伟达数据中心收入（及中国口径变化）
- HBM / CoWoS 产能与利用率
- 主流模型 API 价格曲线（token 通缩速度）
- 旗舰模型基准进展（SWE-Bench 等）
- 企业 AI 支出与 ROI 调查；云 backlog 转化率

改变判断的阈值：capex 集体下修 = 风险兑现；企业可衡量 ROI 比例显著上升 = 迁移加速；API 价格跌破自托管开源盈亏线 = 闭源承压加剧。