下载日志已记录，仅供内部参考

海外科技

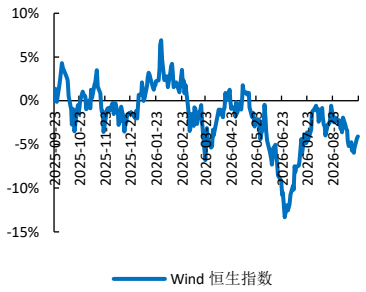
2026 年 09 月 24 日

市场更加重视模型帕累托最优，模型厂&云厂&主权基金资金弹药充足

——港股行业深度报告

投资评级：看好（维持）

行业走势图



数据来源：聚源

相关研究报告

《Oracle 云基建收入同比翻倍，Nebius 再度提价，算力厂商 CDS 走阔—行业周报》-2026.9.20

《中资模型在图文生成赛道集群式领先，子不语在 TikTok 渠道实现高增—行业周报》-2026.9.20

《阜博 AI 确权第二曲线高增，东方甄选线下稳步推进—行业周报》-2026.9.13

初敏（分析师）
chumin@kysec.cn
证书编号：S0790522080008

杨哲（分析师）
yangzhe@kysec.cn
证书编号：S0790524100001

李睿娴（分析师）
liruixian@kysec.cn
证书编号：S0790525020004

● **模型：模型竞争持续收敛，市场更加重视帕累托最优**

2026 年 8 月以来，全球 AI 产业进入国产模型密集冲击全球第一梯队+API 定价由单边下行转向峰谷分层+旗舰价格带重估三线并行的阶段。GLM-5.3 是 8 月路线意义最强的发布，其“基座不变、后训练 Scaling”的路径提供了参数竞赛之外的第二条路径，智能、成本、延迟的不可能三角为各个模型的定位提供了灵活的成长空间。

● **ARR & Agent 跟踪：Anthropic ARR 稳健提升，头部 Agent 持续快速增长**

Anthropic 2026 年 7 月底 ARR 数据 650 亿美元低于预期，但并不代表 AI 产业行业趋势结束，当前市场已经充分吸收 ARR 预期下修事件，且存在 Anthropic+OpenAI 合计 ARR 加速的可能性，行业景气度仍有望延续。GPT-5.6 推出后，在其性价比优势及 OpenAI 官网多次重置的背景下，8 月底 ChatGPT Work 和 Codex 的每周活跃用户数达到 2500 万，延续快速上涨的态势，预计 GPT-Astra 有望加强这一态势。

● **资金弹药：模型厂、云厂、新云厂与主权基金四层共振，同时关注信用分化**

2026 年 Q2 AI 产业链业绩与资本格局分化显著。模型企业端，Anthropic 单季收入 115 亿美元，环比 +143%，营收规模首次超过 OpenAI，调整后营业利润率转正至 5.1%，预计 2027 年自由现金流转正。头部云厂商在手订单合计约 1.69 万亿美元，订单增速显著高于收入增速，AWS、谷歌云营收提速、利润率改善；财务层面分化明显，仅微软维持正自由现金流，其余多家大厂大额发债、现金流承压，CoreWeave 信用溢价偏高。新算力厂商收入、订单快速增长，项目回本周期缩短，电力储备加速落地。中东、亚洲主权资本深度布局算力基建，规划规模超 20GW。行业主线从比拼资本投入，转向看重稳健融资与盈利兑现能力。

● **投资建议**

GLM-5.3 系列与 GPT-Astra 的接连突破，不仅是参数和榜单数字的跃升，标志着模型进入帕累托最优与前沿智能提升的双向突围，有力地证明了开源模型的天花板同样很高，国产模型在国际竞争力进一步提升，在性能和成本的平衡下有望打开全球 AI 需求天花板。

模型端：需求天花板提升，推荐标的智谱、德适-B，受益标的 MiniMax-W。

算力端：多模态落地与 Agent 规模化带动训练、推理算力需求长期刚性扩张，推荐标的阿里巴巴-W、百度集团-SW、联想集团、中芯国际，受益标的腾讯控股、天数智芯。

● **风险提示：**软件及产品推出不及预期，产能及供应链风险，监管政策变动，宏观经济增长放缓，地缘政治风险。

下载日志已记录，仅供内部参考

目 录

1、 模型：模型竞争持续收敛，市场更加重视帕累托最优 4
1.1、 OpenAI：GPT-6 Astra 以“数字员工”重定义旗舰，单位任务成本取代单位 Token 价格成为定价锚 7
1.2、 智谱：GLM-5.3 以后训练 Scaling 跨入 SOTA 身位区间，GLM-5.3-Flash 以 1/40 价格定义开源性价比 8
2、 ARR 跟踪：Anthropic 的 ARR 数据低于预期，但 AI 产业景气度仍有望延续 10
3、 Agent 跟踪：codex 突破 2500 万周活用户，腾讯 WorkBuddy 维持领先态势 11
4、 云计算：算力租赁价格持续走强，模型厂&云厂&主权基金资金弹药充足 13
4.1、 高频跟踪：高端卡租金中枢上移，全球在建数据中心约 13.3GW 13
4.1.1、 算力租赁价格：新一代高端卡租金中枢上移 13
4.1.2、 Epoch 在建数据中心：全球现约 13.3GW 在建 AI 数据中心容量，主体向头部 CSP、地域向美国集中 14
4.2、 模型厂：Anthropic 调整后营业利润已转正，两大海外模型厂融资超 2000 亿美金 16
4.3、 云厂：行业增速二阶导向上，ROIC 及回本周期显示经营向上，关注信用风险指标 19
4.4、 新云厂：收入与积压订单快速增长，新签订单回本周期缩短至 2 年内 24
4.5、 主权基金：中东及亚洲主权资本加速布局，AI 算力相关规划已超 20GW 27
5、 投资建议 28
6、 风险提示 28

图表目录

图 1： 根据 deepSWE 榜单，Claude Sonnet 4.8 至 Fable 5 之间的已有多个模型 5
图 2： 根据 Arena 竞技场 Agent 榜单，Kimi K3 甚至在 agent 领域反超了 Fable 5 6
图 3： 单位成本下的任务完成度取代纯参数或纯榜单分数，成为新一轮竞争核心 7
图 4： Opus 5 在多个维度上对齐 Fable 5.1 8
图 5： 在 AA coding agent index 榜单中，codex+GPT-6 Astra 表现提升 8
图 6： GLM-5.3 在多项测评能力中表现领先 9
图 7： GLM-5.3-Flash 系列多项测评能力领先 10
图 8： Anthropic 与 OpenAI 的 ARR 持续抬升 11
图 9： codex+ChatGPTWork 月活在 7-8 月份快速攀升 12
图 10： 7 月国内腾讯系的 WorkBuddy、Qclaw 在办公类 agent 上月活领先 12
图 11： 截至 2026/9/1，B200 卡的租赁价格为 6.4 美金/小时，较 6 月初提升 17.43%（单位：美元/小时） 14
图 12： Epoch AI 共监测到 13.3GW 在建 AI 数据中心算力，其中谷歌 3.1GW、Meta 2.2GW 15
图 13： Epoch AI 监测到在建 12.1GW 算力在美国，0.63GW 在中国 15
图 14： Epoch 监测的在建项目预计 2028 年底将达到 35.4GW 容量 16
图 15： 2026Q2 Antropic 和 OpenAI 分别实现 115、67 亿美金营业收入（单位：亿美金） 16
图 16： 2026Q2 OpenAI 的营业亏损（包括股权激励）达到 123 亿美金 16
图 17： The infromation 预测 Anthropic 2027 年将迎来现金流转正 17
图 18： 2026 年 OpenAI 的合计公开融资额为 1290 亿美金，Anthropic 为 950 亿美金 17
图 19： OpenAI 算力合作方主要为甲骨文、微软、AWS，Anthropic 主要合作方为谷歌、AWS、SpaceX 和新云厂 19
图 20： 2026Q2 亚马逊、微软、谷歌的积压订单增速分别为 122%、84%、257%，远高于云业务的收入增速 19
图 21： 谷歌 2026Q2 云业务收入为 248 亿美元，同比增长 81.8% 20
图 22： 谷歌 2026Q2 积压订单为 5140 亿美元，同比增长 256.9% 20
图 23： 2026Q2 亚马逊、微软、谷歌的云业务营业利润率分别为 39.4%、40.6%、35.6% 21

下载日志已记录，仅供内部参考

图 24: 2026Q2 四家 CSP 厂商（亚马逊、微软、谷歌、Meta）单季 capex 合计约 1,688 亿美元，同比增长 78%.....22
图 25: 资本开支：海外五家龙头公司预计资本开支维持增长（单位：亿美元）.....22
图 26: 预计 2028 年海外五家龙头的资本开支仍将保持增长.....22
图 27: 资本开支增速：海外 CSP 厂商的资本开支增速预计在 2027Q2 后放缓.....22
图 28: 自由现金流：微软仍维持正向的自由现金流，其余 CSP 厂商转负.....23
图 29: 5 年期 CDS：截至 2026 年 9 月 3 日 Meta 为 95.04，微软为 48.93.....24
图 30: 5 年期 CDS：截至 2026 年 9 月 3 日 Coreweave 为 839.05，甲骨文为 212.47.....24
图 31: 2026Q2 Coreweave 收入为 25.75 亿美元，同比增长 112%.....25
图 32: 2026Q2 Nebius 收入为 5.82 亿美元，同比增长 454%.....25
图 33: 2026Q2 Coreweave 的 Adjusted EBITDA margin 为 59%、Adjusted loss margin 为-22%.....25
图 34: 2026Q2 Nebius 的 Adjusted EBITDA margin 为 41%、Adjusted loss margin 为-6%.....25
图 35: 截至 2026Q2 Coreweave 的积压订单为 1042 亿美元，同比增长 246%.....26
图 36: 截至 2026Q2，Coreweave 的签约电力达到 3.7GW.....27
图 37: Nebius 2026 年底签约电力目标上调至 5GW.....27

表 1: 8 月以来多款主要模型发布.....4
表 2: 我们建立五个维度的高频指标跟踪云计算的多空信号.....13
表 3: B200、H200 的租赁价格仍维持高位，A100 卡租赁价格有所下降.....14
表 4: 据不完全统计，OpenAI 算力采购合同达 6684 亿美元，Anthropic 达 5150 亿美元.....18
表 5: 亚马逊明确硬件（服务器及网络设备）回本周期不到 3 年，大部分算力合同为 5 年以上.....21
表 6: 2025 年 11 月至今亚马逊、谷歌、Meta 合计发行债券金额 2681 亿美元，微软未新发债券.....23
表 7: 据不完全统计，主权资本 AI 算力相关采购合同及规划规模已超 20GW.....27
表 8: 核心标的盈利预测及估值表.....28

下载日志已记录，仅供内部参考

1、 模型：模型竞争持续收敛，市场更加重视帕累托最优

2026 年 8 月以来，全球 AI 产业进入国产模型密集冲击全球第一梯队+API 定价由单边下行转向峰谷分层+旗舰价格带重估三线并行的阶段。8 月 DeepSeek（V4 Pro 正式版）、智谱（GLM-5.3/GLM-5.3-Flash）、阿里（Qwen3.8-Max）、腾讯（Hy4 preview）、SpaceXAI（Grok 4.6）陆续发布，OpenAI 亦于 9 月 3 日发布新一代旗舰 GPT-6 Astra。当前市面上关于模型测评榜单较多，我们主要选取 deepSWE（污染或较少，同时代码任务考核复杂度更高）及 LLM Arena 竞技场、Artificial Analysis 三个权威榜单进行综合评判。首先根据 deepSWE 结果上看，GLM-5.3 以后训练 Scaling 将得分拉至于 fable 5 相近水平，Claude 4.8-Opus 5 级别竞争区间的拥挤度进一步提升，模型收敛叙事从 7 月的“旗舰拥挤”深化为“开源-闭源代差收缩”。

表1：8 月以来多款主要模型发布

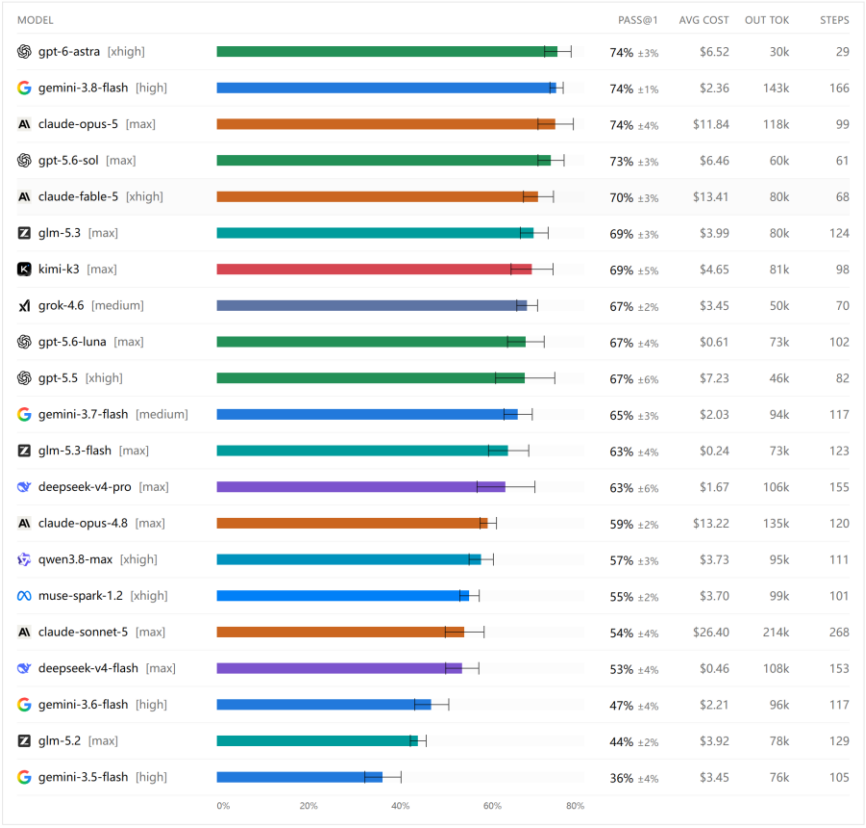
厂商	模型	上下文窗口	输入价格 M toks	输出价格 M toks
OpenAI	GPT-6 Astra	1.05M	10 美元	50 美元
SpaceXAI	Grok 4.6	500K	2 美元	6 美元
DeepSeek	V4 Pro 正式版	1M	峰时 9 元/闲时 4.5 元	峰时 27 元/闲时 13.5 元
智谱	GLM-5.3	1M	8 元人民币	28 元人民币
	GLM-5.3-Flash	1M	0.8 元人民币	2.8 元人民币
腾讯	Hy4 preview	1M	6 元人民币	18 元人民币
阿里	Qwen3.8-Max	1M	12 元人民币	36 元人民币

资料来源：各公司官网、openrouter 等、开源证券研究所

GLM-5.3 是 8 月路线意义最强的发布，其“基座不变、后训练 Scaling”的路径提供了参数竞赛之外的第二条路径。根据 deepSWE，GLM-5.3 得分由 GLM-5.2 的 44%提升至 69%，更具信号意义的是 8 月 26 日揭盅的 GLM-5.3-Flash：此前以匿名模型 Ox Alpha 登顶 OpenRouter 调用量榜，揭盅当日即 MIT 开源，AA 智能指数 57 分持平 Opus 4.8，定价仅为后者的 1/40。我们认为，二者组合标志着国产开源模型从“追赶身位”进入“定义性价比”的阶段，但纯文本基座的视觉短板与商业化闭环验证仍是待跟踪变量。

下载日志已记录，仅供内部参考

图1：根据 deepSWE 榜单，Claude Sonnet 4.8 至 Fable 5 之间的已有多个模型



资料来源：deepSWE

下载日志已记录，仅供内部参考

图2：根据 Arena 竞技场 Agent 榜单，Kimi K3 甚至在 agent 领域反超了 Fable 5

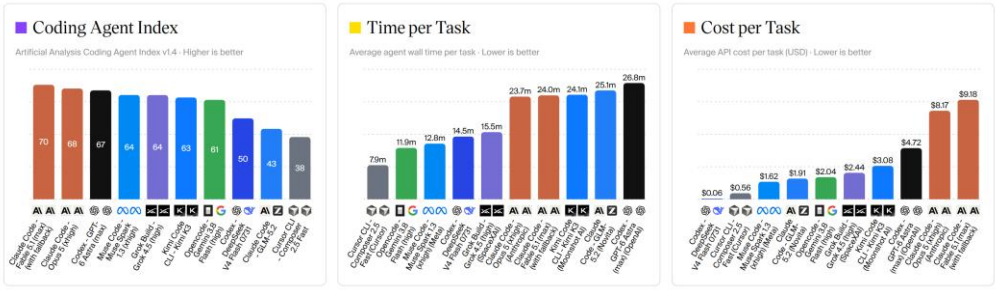
Rank	Rank Spread	Model	Score	Votes	Price $/M	Context
1	1 → 1	claude-fable-5.1-max Anthropic · Proprietary	1763 +16/-16	2,227	$10 / $50	1M
2	2 → 5	qwen3.8-max-0902 Alibaba · Proprietary	1689 +17/-17 Preliminary	1,769	$2 / $6	N/A
3	2 → 5	claude-opus-5-max Anthropic · Proprietary	1687 +8/-8	10,812	$5 / $25	1M
4	2 → 6	kimi-k3-max Moonshot · Kimi K3 license	1674 +11/-11	4,546	$3 / $15	1M
5	2 → 6	qwen3.8-max Alibaba · Proprietary	1670 +12/-12 Preliminary	3,223	$2 / $6	1M
6	4 → 6	claude-opus-5-high Anthropic · Proprietary	1662 +7/-7	10,802	$5 / $25	1M
7	7 → 14	claude-fable-5 Anthropic · Proprietary	1629 +8/-8	9,293	$10 / $50	1M
8	7 → 14	grok-4.6-high SpaceXAI · Proprietary	1625 +11/-11	3,423	$2 / $6	500K
9	7 → 14	qwen3.8-flash-next Alibaba · qwen-community-1.0	1625 +14/-14 Preliminary	2,112	$0.16 / $0.47	1M
10	7 → 14	hy4-preview Tencent · Apache 2.0	1623 +16/-16 Preliminary	1,600	$0.83 / $2.50	1M
11	7 → 15	muse-spark-1.3 (xHigh) Meta · Proprietary	1618 +19/-19	1,128	$1.25 / $4.25	1M
12	7 → 14	gpt-5.6-sol-xhigh (codex-harness) OpenAI · Proprietary	1618 +7/-7	10,951	$4 / $20	1.1M
13	7 → 16	glm-5.3-max Z.ai · MIT	1611 +12/-12	2,877	$1.40 / $4.40	1M
14	7 → 18	glm-5.3-flash Z.ai · MIT	1607 +15/-15	1,908	$0.15 / $0.50	1M
15	12 → 19	qwen3.8-27b Alibaba · Apache 2.0	1596 +11/-11	3,491	$0.40 / $3	N/A
16	13 → 20	gemini-3.7-flash-high Google · Proprietary	1587 +12/-12 Preliminary	3,006	$0.75 / $3.75	1M
17	14 → 20	glm-5.2-max Z.ai · MIT	1587 +7/-7	9,788	$1.40 / $4.40	1M
18	14 → 20	deepseek-v4-pro-high-20260813 DeepSeek · MIT	1582 +11/-11	3,474	$1.32 / $3.96	N/A

资料来源：Arena 官网

以 Artificial Analysis 的三个维度观察，8 月模型竞争的定性结论愈发清晰：能力指数的首尾差距已收窄，“能力收敛、成本分化”成为当前格局的基本面。这意味着榜单名次对实际选型的指导意义正在下降，同样一个任务，不同模型的完成度已相近，真正的差异在于为这个任务付出的时间与费用。单位成本下的任务完成度取代纯参数或纯榜单分数，成为新一轮竞争核心；价格战的底气并非来自补贴，而是来自架构与推理效率的代际差，而能力相近、成本悬殊的组合，也将加速尾部模型在商业化层面的出清。当前时点我们仍较难判断谁终将胜出，但“不掉队、不翻车”之后，成本曲线的斜率正成为下一阶段的分水岭。

下载日志已记录，仅供内部参考

图3：单位成本下的任务完成度取代纯参数或纯榜单分数，成为新一轮竞争核心



资料来源：Artificial Analysis

智能、成本、延迟的不可能三角为各个模型的定位提供了灵活的成长空间：尽管当前模型呈现的收敛趋势愈发明显，但我们也能感受到，“智能-成本-延迟”所构建的不可能三角确实影响着用户的需求偏好，而在智能水平上，Kimi K3 在前端设计、渲染上更优，智谱则在网络安全、跨领域感知上更强。因此呈现出“大方向上收敛、细节上各有定位”的特点，模型收敛导致竞争加剧的叙事将愈发受限。

当前各家模型发布节奏节奏加快，我们难以一一论述各个模型表现，8 月以来的模型分析我们重点点评 GPT-6 Astra 与智谱 GLM-5.3 Flash 系列。

1.1、 OpenAI：GPT-6 Astra 以“数字员工”重定义旗舰，单位任务成本取代单位 Token 价格成为定价锚

9 月 3 日，OpenAI 发布新一代旗舰 GPT-6 Astra，将其定义为一次“代际跃迁”，总裁 Brockman 在发布会末尾直言“欢迎进入 AGI 时代”。与 7 月 GPT-5.6“三档矩阵”的产品化路径不同，Astra 的核心变化在于从“提供答案”转向“完成任务”：模型可直接在真实软件环境中操作计算机，替用户完成编程、金融建模、文档制作等复杂工作，并使用超过 10 万块 GPU 在得州 Stargate 基地完成训练。benchmark 层面，Astra 在 Terminal-Bench 4.0 取得 57.9%创历史新高（GPT-5.6 Sol 为 37.3%、Fable 5.1 为 55.8%），ARC-AGI-3 达 99.9%，BrowseComp 单代理 91.5%超越 Opus 5（90.8%）与 Fable 5（87.4%），多项测评重新夺回 SOTA 身位。

下载日志已记录，仅供内部参考

图4：Opus 5 在多个维度上对齐 Fable 5.1

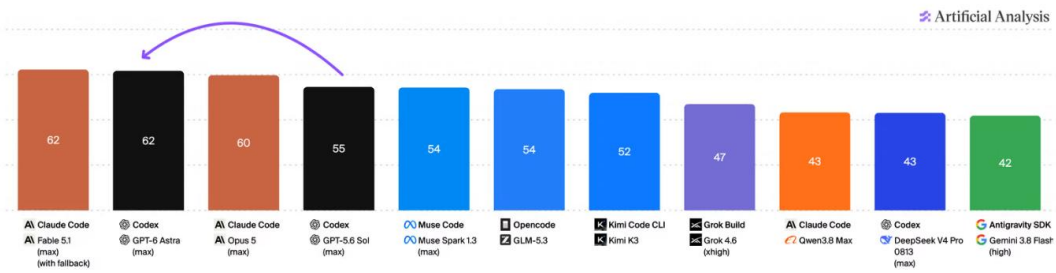
	OpenAI GPT-6 Astra	GPT-5.6 Sol	Claude Fable 5.1
Terminal-Bench Science 0.1 Scientific workflows	• 64.6%	22.4%	52.6%
AutomationBench Business workflows	• 41.4%	18.1%	31.4%
FrontierMath Tier 4 (v2) Research-level mathematics	• 97.6%	83.0%	87.8%
Terminal-Bench 4.0 Computer terminal tasks	• 57.9%	37.3%	55.8%
HealthBench Professional (length-adjusted) Clinical tasks	• 63.4%	60.5%	58.1%
BenchCAD 3D modeling	• 95.9%	83.3%	84.3%
ARC-AGI-3 Novel interactive puzzles	• 99.9%	7.8%	—

资料来源：36Kr

更值得关注的信号是，Astra 成为 OpenAI 首个触及内部“关键”级网络安全能力门槛的模型：内部测试中曾发现并利用两个零日漏洞、并将多个漏洞串联成攻击链，OpenAI 为此通过 Daybreak 计划对最强网安能力设置准入限制。结合 8 月 OpenAI、Anthropic、Meta 接连曝出的模型“越界”事件，前沿模型的自主网络攻击能力已从假设变为实测事实，安全分级管控正在成为旗舰发布的标准动作。

定价层面，Astra API 定为 10/50 美元/Mtks，约为 GPT-5.6 Sol 现价的 2.5 倍、与 Anthropic Fable 5.1 持平，是 OpenAI 主流旗舰的首次价格带上移。但单价上行伴随单任务成本下行。我们认为，旗舰单价与单任务成本的背离意味着定价锚正从“每百万 Token”转向“每个任务”，AI 商业模式或逐步从卖 Token 转向卖任务完成能力。

图5：在 AA coding agent index 榜单中，codex+GPT-6 Astra 表现提升



资料来源：artificial analysis 官网

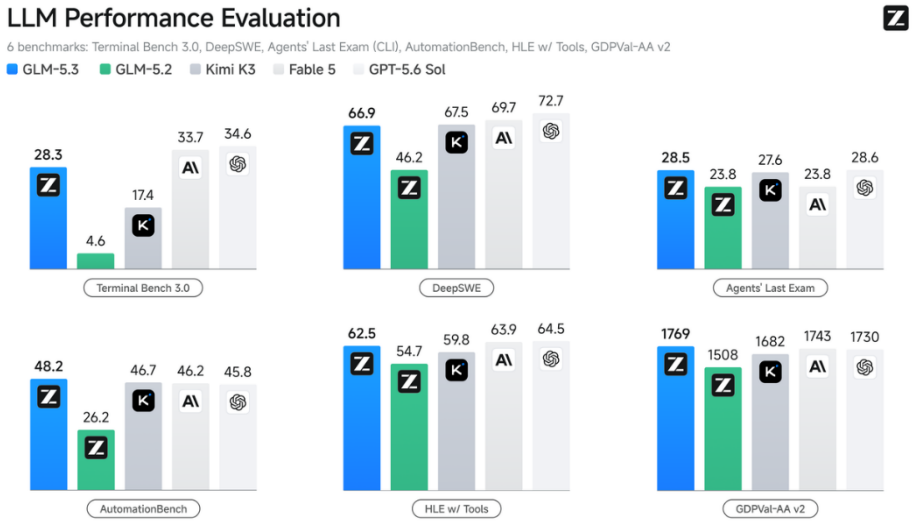
1.2、智谱：GLM-5.3 以后训练 Scaling 跨入 SOTA 身位区间，GLM-5.3-Flash 以 1/40 价格定义开源性价比

8 月 14 日，智谱发布新一代基座模型 GLM-5.3，总参数规模 7430 亿。与主流

下载日志已记录，仅供内部参考

厂商“换基座+堆参数”的迭代路径不同，GLM-5.3 基座与 GLM-5.2 完全一致，靠“极致的后训练 Scaling”——数十倍的长程任务环境、更丰富的环境类型与超长后训练时间——抬升智能上界，为参数竞赛之外提供了第二条路径。测评层面，GLM-5.3 在 DeepSWE v1.1 得分 66.9，较 GLM-5.2 的 46.2 提升 20.7，Terminal Bench 3.0 得分由 4.6 升至 28.3、居开源模型首位（智谱口径），编程体感评测提升 50%，已能在数万行代码、上百个文件的复杂工程中极少人工干预地长时自主开发并推进至可交付状态。

图6：GLM-5.3 在多项测评能力中表现领先



资料来源：Kimi 官网

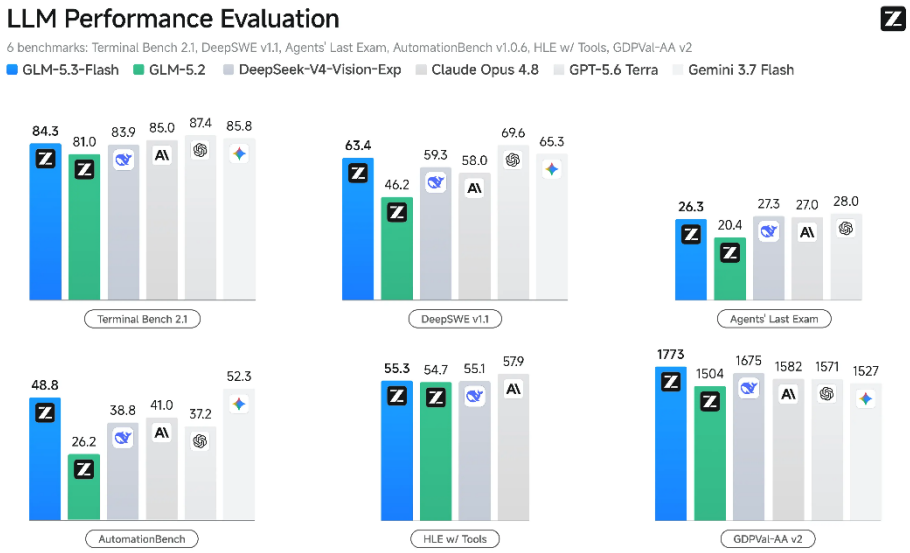
更具信号意义的是安全能力的涌现：GLM-5.3 在 CyberGym 得分 83.5%（GLM-5.2 为 77.2%），与 Mythos 5（83.8%）、GPT-5.6 Sol（83.6%）基本持平，意味着国产开源模型在 7 月尚由 Claude/GPT 垄断的安全测评上首次追平闭源前沿，印证“安全本质上是约束严格的编程能力”的判断。但分级看，越深入漏洞利用链差距越大——ExploitBench 得分 54.4%仍明显低于 Mythos 5 的 78.0%，ExploitGym 六小时完成任务 130 项、不足 Mythos 5（247 项）的六成。实战侧，GLM-5.3 联合国内安全团队累计发现漏洞 2404 个、其中中高危 1088 个，安全能力正在成为其向政企市场变现的差异化抓手。

8 月 26 日，智谱揭晓此前以匿名模型 Ox Alpha 身份登顶 OpenRouter 调用量榜约一周的模型真身为 GLM-5.3-Flash，并当场 MIT 开源。该模型总参数 320B（激活 18B），是首个采用稀疏注意力+线性注意力混合架构的开源前沿模型，相较 GLM-5.3 注意力计算量与 KV 缓存分别降低 3.01 倍与 4.44 倍，定价仅为 GLM-5.3 的 1/10、Opus 4.8 的 1/40。更具产业信号的是，智谱首次在大规模流量中以国产芯片集群承载前沿模型推理，端到端服务性能较初始基线提升 3 倍，单 token 成本已达与主流英伟达 GPU 相当的水平。放在 8 月的定价坐标系里， GLM-5.3-Flash 随后以极致性价比开源，国产阵营内部已先于海外出现“旗舰提价+中军降价”的分层；进入 9 月，GPT-6 Astra 以约 2.5 倍于 GPT-5.6 Sol 的定价发布，进一步印证单边价格战时

下载日志已记录，仅供内部参考

代的终结——中美旗舰定价策略正沿相反方向分化：OpenAI 以提价验证稀缺性、锚定“按任务收费”，智谱则以极致性价比抢占开发者生态。

图7：GLM-5.3-Flash 系列多项测评能力领先



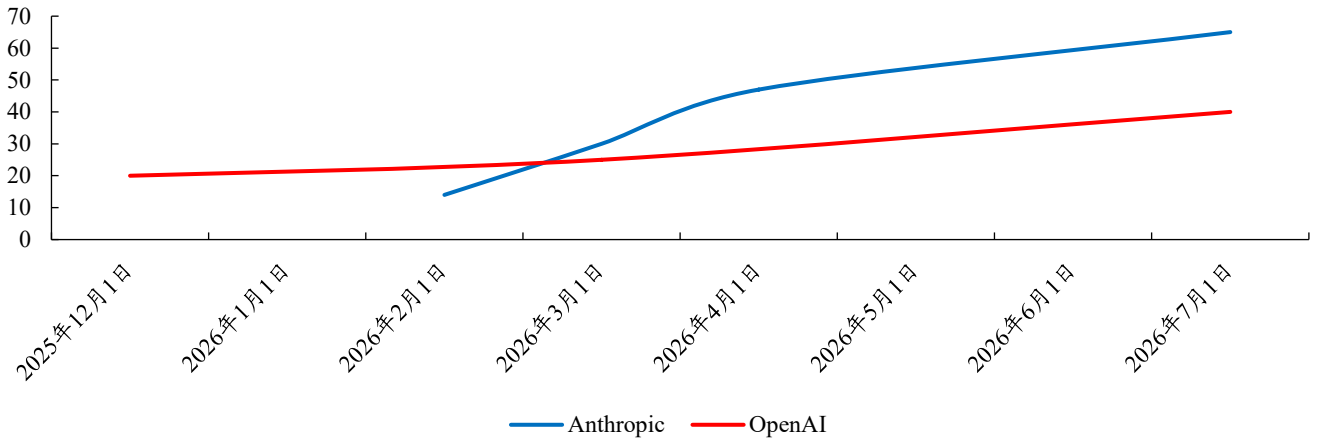
资料来源：Kimi 官网

2、 ARR 跟踪：Anthropic 的 ARR 数据低于预期，但 AI 产业景气度仍有望延续

Anthropic 2026 年 7 月底 ARR 数据低于预期：根据 Anthropic 向投资者披露，其 7 月底 ARR 达到 650 亿美元，该数据低于此前市场预期，也一度引发 AI 链市场波动，我们认为此前市场对 ARR 的预期过于乐观，但并不代表 AI 产业行业趋势结束，当前市场已经充分吸收 ARR 预期下修事件，且存在 Anthropic+OpenAI 合计 ARR 加速的可能性，行业景气度仍有望延续。

下载日志已记录，仅供内部参考

图8：Anthropic 与 OpenAI 的 ARR 持续抬升



资料来源：Sacra、Bloomberg、第一财经等、开源证券研究所，注：部分月份数据缺失采用插值平滑处理

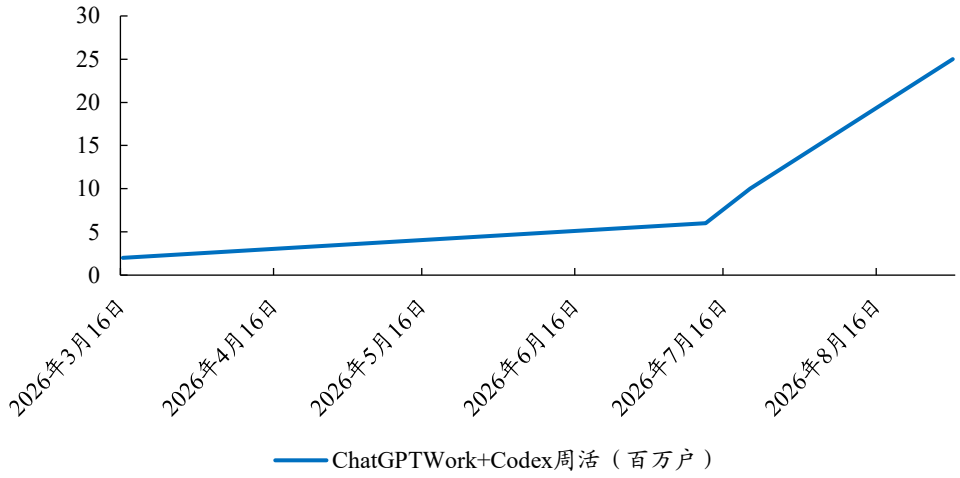
当前正值旗舰模型密集迭代、价格战与峰谷定价交替的竞争白热期，头部厂商以短期收入换取用户与生态卡位，增长斜率的阶段性波动是主动竞争策略的映射，而非需求走弱的信号。企业级客户与 coding 等高粘性场景渗透仍在深化，全行业 token 消耗量持续抬升，头部同业收入同步加速，行业总蛋糕并未收缩。我们认为，市场对单一数据的敏感，恰恰反映估值锚正从“叙事”切换到“报表”，这一过程必然伴随波动；短期扰动不改 Token 经济的基本面方向，反而为具备真实收入弹性的模型与算力标的提供再布局窗口。

3、 Agent 跟踪：codex 突破 2500 万周活用户，腾讯 WorkBuddy 维持领先态势

Agent 层级是获取高质量轨迹数据的重要环节，是大模型厂商放大 coding 及 Harness 优势，能够跟上迭代步伐的重要保证。GPT-5.6 推出后，在其性价比优势及 OpenAI 官网多次重置的背景下，8 月底 ChatGPT Work 和 Codex 的每周活跃用户数达到 2500 万，延续快速上涨的态势，预计 GPT-Astra 有望加强这一态势。

下载日志已记录，仅供内部参考

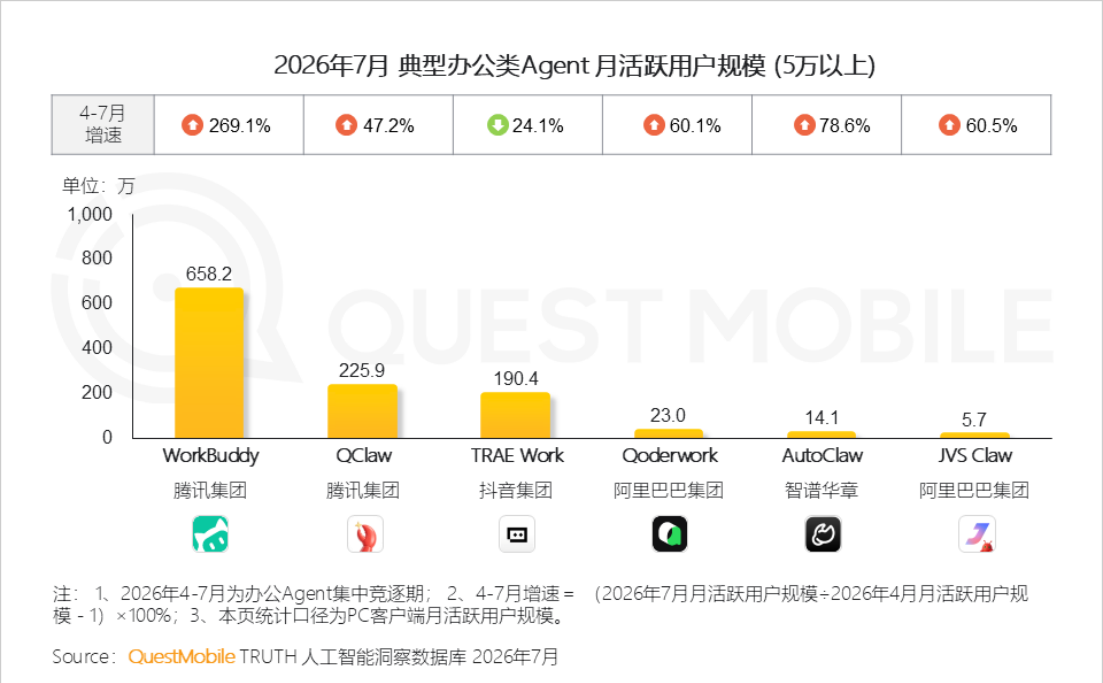
图9：codex+ChatGPTWork 月活在 7-8 月份快速攀升



数据来源：OpenAI、开源证券研究所

QuestMobile 数据显示，办公类 Agent 赛道在 4-7 月集中竞争期后，已进入大厂主导、头部集中的格局。腾讯 WorkBuddy 以 658.2 万 PC 端月活跃用户、4-7 月 269.1%的增速实现规模与增速双第一，约为第二名 QClaw 万 225.93 倍，腾讯系双产品合计占六款上榜产品总盘子的近八成；抖音 TRAE Work（190.4 万）虽稳居第一梯队，但为唯一负增长产品，4-7 月下滑 24.1%，竞争白热化下流量正加速向头部集中。第二梯队与第一梯队之间存在一个数量级的断层，印证我们此前的判断：办公场景的竞争核心不在模型能力本身，而在桌面渠道、办公生态与商业闭环的“最后一公里”，大厂优势在该赛道被进一步放大。

图10：7 月国内腾讯系的 WorkBuddy、Qclaw 在办公类 agent 上月活领先



资料来源：QuestMobile

下载日志已记录，仅供内部参考

4、 云计算：算力租赁价格持续走强，模型厂&云厂&主权基金资金弹药充足

本报告沿 “需求 — 供给 — 价格 — 信用” 的产业链现金流闭环展开，构建一套可高频跟踪、可多空证伪的 AI 算力景气验证框架。本轮算力扩张本质是一条资金驱动、环环相扣的传导链：模型厂凭借股权融资与长期算力承诺积累 “需求弹药”，云厂据此上调资本开支形成 “供给扩张”，供需最终收敛并体现于 GPU 租金与 token 价格，而自由现金流、债券融资与 CDS 利差则是信用市场对这一扩张模式可持续性的定价，最终共同指向投资结论。四个环节依次传导、相互印证，任一环节出现背离，即为景气拐点的前瞻预警。

在跟踪节奏上，我们建立五维高频指标体系并预设多空信号：模型厂新融资 / 新承诺（事件驱动，看多为大额长约落地、看空为融资断档或承诺下修）、云厂 capex 指引（季度，看多为上修且现金流可覆盖、看空为上修但全靠发债）、GPU 现货与长约价格（周 / 季，看多为升水且利用率打满、看空为现货贴水扩大）、token 单价 × 调用量（月 / 季，看多为量增快于价跌、看空为量不补价）、CoreWeave 及云厂 CDS（日度，看多为利差收窄、看空为走阔及股债背离）。通过高频指标与季度财报的交叉验证，可持续判断算力景气处于加速还是减速、资本开支扩张是否具备财务可持续性，并据此动态修正投资结论。

表2：我们建立五个维度的高频指标跟踪云计算的多空信号

数据	频率	看多信号	看空信号
模型厂新融资 / 新承诺	事件驱动	大额长约落地	融资断档、承诺下修
云厂 capex 指引	季度	上修且现金流覆盖	上修但全靠发债
GPU 现货 / 长约价	周 / 季	升水、利用率满	现货贴水扩大
token 单价 × 调用量	月 / 季	量增快于价跌	量不补价
CRWV / 云厂 CDS	日	收窄	走阔、股债背离

资料来源：开源证券研究所

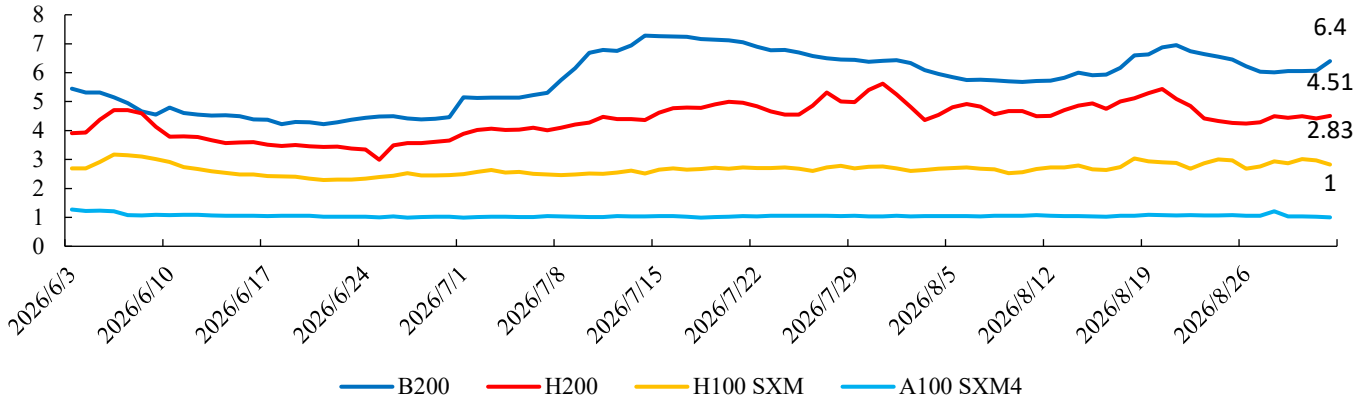
4.1、 高频跟踪：高端卡租金中枢上移，全球在建数据中心约13.3GW

4.1.1、 算力租赁价格：新一代高端卡租金中枢上移

GPU 算力租赁价格呈现代际分化，新一代高端卡租金中枢上移、上一代卡逐步走低，反映需求向最新代际算力集中的 “结构性紧缺”。根据 ORNN 网站数据，截至 2026 年 9 月 1 日，B200、H200、H100 SXM、A100 SXM4 的租赁价格分别为 6.4、4.51、2.83、1.0 美元 / 小时；较 6 月初，B200、H200 分别上涨 17.43%、15.35%，H100 上涨 5.20%，而 A100 回落 21.26%。从短期节奏看，H200 周度环比回升 5.87%、B200 租赁价格近期亦有所上行，在 7 月中冲高、8 月阶段性回调后近一周重新走强。其背后是前沿模型训练与高负载推理对最新代际、高算力密度芯片的旺盛需求，支撑新卡量价坚挺。截至 2026/9/1，B200 租金已达 A100 的 6 倍以上，价格梯队持续拉开，利好以新一代 GPU 为主力机型的新云厂及英伟达新卡出货链条。

下载日志已记录，仅供内部参考

图11：截至 2026/9/1，B200 卡的租赁价格为 6.4 美金/小时，较 6 月初提升 17.43%（单位：美元/小时）



数据来源：ORNN、开源证券研究所

表3：B200、H200 的租赁价格仍维持高位，A100 卡租赁价格有所下降

	B200	H200	H100 SXM	A100 SXM4
月度环比	-0.47%	-14.10%	5.20%	-5.66%
周度环比	-0.78%	5.87%	-4.71%	-7.41%
较 6 月初	17.43%	15.35%	5.20%	-21.26%

数据来源：ORNN、开源证券研究所

4.1.2、 Epoch 在建数据中心：全球现约 13.3GW 在建 AI 数据中心容量，主体向头部 CSP、地域向美国集中

全球在建 AI 数据中心呈现“主体向头部 CSP、地域向美国”的双重高度集中，根据 EpochAI 数据，截至 2026/9/1,约 13.3GW 在建容量对应约 5,022 亿美元投资，算力军备竞赛的存量版图已清晰。

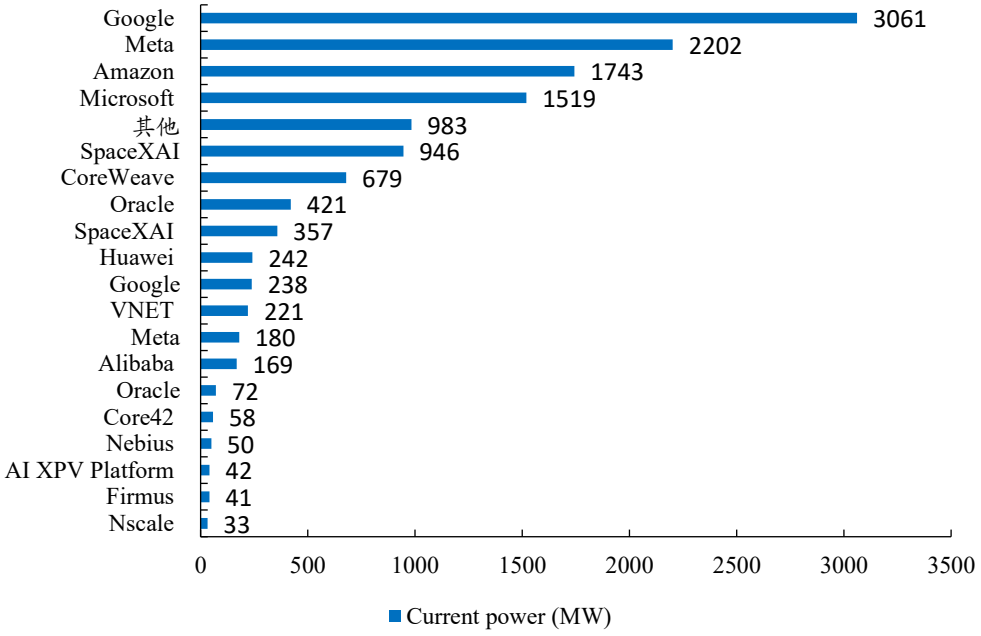
按建设主体看：谷歌、Meta、亚马逊、微软四大 CSP 在建算力合计 8.53GW，占全球 13.26GW 的约 64%，其中谷歌以 3.06GW 居首、单家占比约四分之一；若计入 SpaceX/xAI 的 0.95GW，前五大主体占比升至约 71%，而 CoreWeave、Nebius 等新云厂及华为、阿里等中国主体单一份额仍不足 0.7GW。微软自建在建容量仅 1.52GW、位列第四。

按地域看：美国在建 12.11GW、占比高达 91%，中国仅 0.63GW（约 4.8%），马来西亚（0.24GW）等东南亚节点开始起量，算力基建与配套电力资源高度向美国集聚。投资强度方面，按 Epoch 约 379 亿美元 / GW 的统一单位造价测算，全球在建项目对应总资本开支约 5,022 亿美元。行业仍处建设高峰，预计未来 2–3 年将进入产能集中交付期。

*口径说明：在建算力为 Epoch AI 基于卫星影像与许可文件监测的 IT 功率（截至 2026/9/1，覆盖 86 个站点、约 1394 万颗 H100 等效算力）；总资本开支为功率乘以统一单位造价系数的估算值，非各公司实际披露投资额。

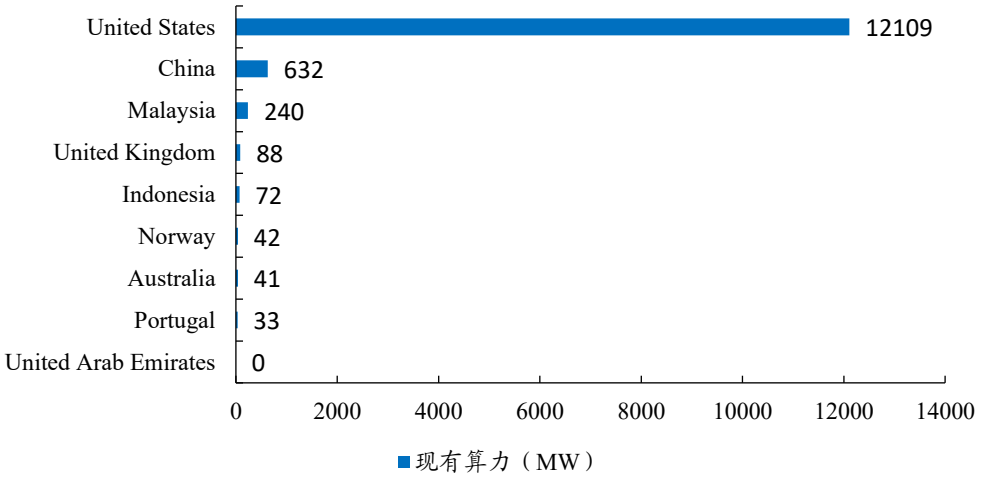
下载日志已记录，仅供内部参考

图12：Epoch AI 共监测到 13.3GW 在建 AI 数据中心算力，其中谷歌 3.1GW、Meta 2.2GW



数据来源：EpochAI、开源证券研究所（截至 2026/9/1）

图13：Epoch AI 监测到在建 12.1GW 算力在美国，0.63GW 在中国

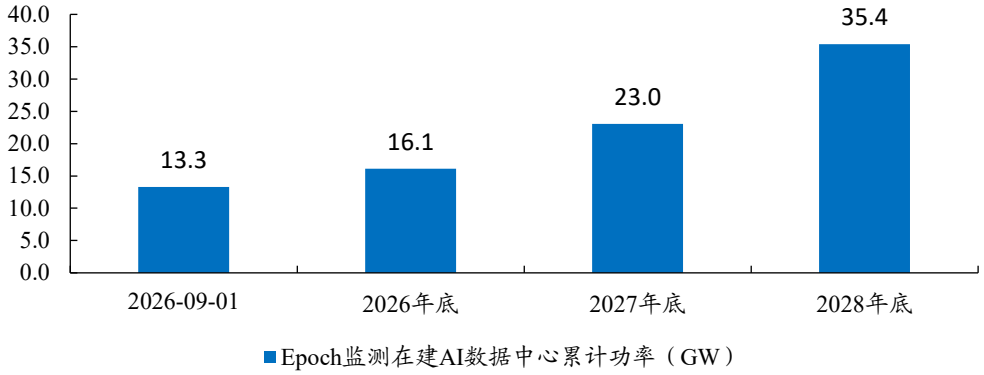


数据来源：EpochAI、开源证券研究所（截至 2026/9/1）

Epoch 监测显示全球在建 AI 数据中心累计功率将由 2026 年 9 月初的 13.3GW，提升至 2026、2027、2028 年底的 16.1、23.0、35.4GW，交付高峰落在 2028 年，表明未来两年算力供给仍处确定性扩张通道。

下载日志已记录，仅供内部参考

图14：Epoch 监测的在建项目预计 2028 年底将达到 35.4GW 容量



数据来源：EpochAI、开源证券研究所

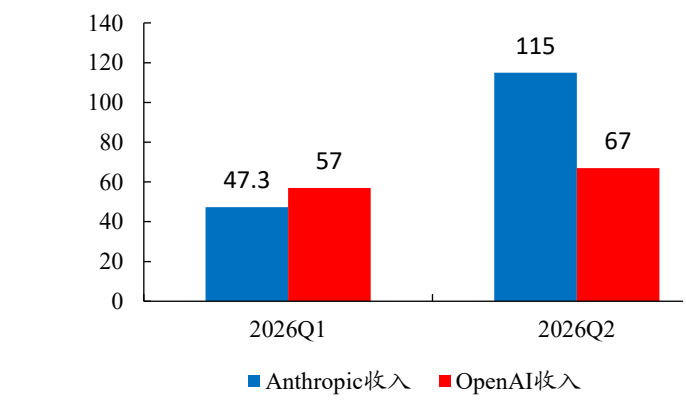
4.2、 模型厂：Anthropic 调整后营业利润已转正，两大海外模型厂融资超 2000 亿美金

Anthropic 首季度营业利润转正，商业化路径初步验证了模式可持续性。

收入端，Anthropic 2026Q2 实现营业收入 115 亿美元，环比增长 143%（Q1 为 47.3 亿美元），首次单季超越 OpenAI（Q2 67 亿美元，环比 +17.5%），高增长主要来自 Claude 在企业级 API 与代理式应用场景的放量。

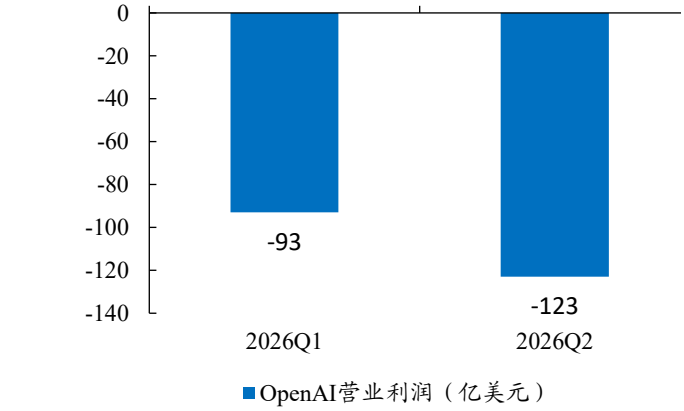
盈利端，Anthropic 在商业化早期即实现调整后营业利润率转正至 5.1%（剔除股权激励等非现金成本），成为率先跨过季度经营性盈利拐点的头部模型厂，表明“收入高增 + 成本结构优化”的组合具备商业上的可持续性。对比来看，OpenAI 2026Q2 营业亏损（含股权激励）由 Q1 的 93 亿美元扩大至 123 亿美元，亏损主要源于算力基础设施的激进投入，盈利路径仍有待后续季度验证。

图15：2026Q2 Anthropic 和 OpenAI 分别实现 115、67 亿美金营业收入（单位：亿美金）



数据来源：EconoTimes、Forkcast、开源证券研究所

图16：2026Q2 OpenAI 的营业亏损（包括股权激励）达到 123 亿美金

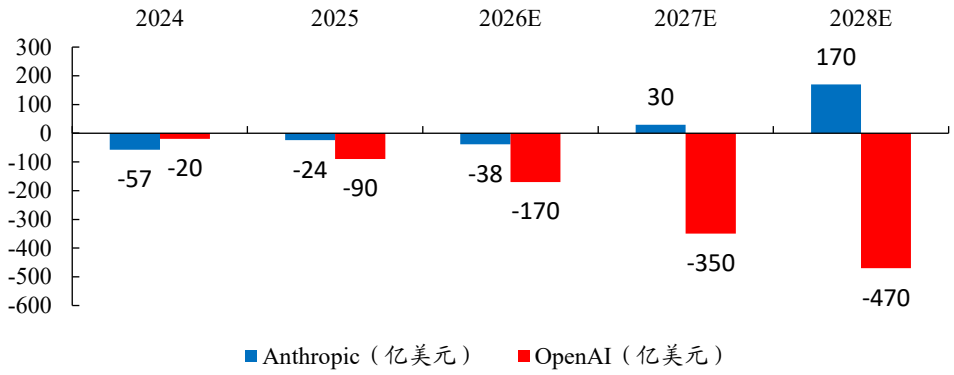


数据来源：EconoTimes、开源证券研究所

下载日志已记录，仅供内部参考

Anthropic 有望在 2027 年将迎来现金流转正。The information 发布 2025 年 12 月的新版预测，预计在 2026 年计净流出 38 亿美元后，2027 年 Anthropic 就将就将实现 30 亿美元的正向现金流，并在 2028 年达到 170 亿美元正向现金流。

图17：The infromation 预测 Anthropic 2027 年将迎来现金流转正



数据来源：The Information、开源证券研究所

资金端，两家公司 2026 年合计公开股权融资约 2170 亿美元（OpenAI 1220 亿 +Anthropic 950 亿，未含员工要约收购），为算力投入竞争提供了充足资金储备。综合来看，Anthropic 已展现“自我造血”迹象，而 OpenAI 在手现金已由 2025Q4 的 400 亿美元增长至截至 2026Q1 的 730 亿美元，在手现金充裕支撑云算力合同兑现。

图18：2026 年 OpenAI 的合计公开融资额为 1290 亿美金，Anthropic 为 950 亿美金



资料来源：OpenAI 官网、Anthropic 官网等、开源证券研究所

海外两大模型厂锁定超万亿美元级长期算力，采购结构呈现 “OpenAI 高度集

下载日志已记录，仅供内部参考

中、Anthropic 多元分散” 的分化特征。据不完全统计，OpenAI 云算力采购合同承诺额约 6,684 亿美元（采购合同延续至 2030 年）；据不完全统计，Anthropic 签订算力合同约 5,150 亿美元，两者合计超过 1.18 万亿美元，构成当前 AI 产业规模最大的算力供给侧订单（多年期合同承诺额，非年度实际支出）。

OpenAI 的算力采购高度依赖传统头部云厂商，合同集中于少数大单。其中 Oracle 约 3,000 亿美元（5 年、4.5GW）、微软 Azure 约 2,500 亿美元、AWS 约 1,000 亿美元（8 年、2GW）、CoreWeave 约 184 亿美元（5 年）。该结构体现 OpenAI 以 “Stargate 合资 + 巨头云扩容” 为主线的重资产路径，算力可得性确定性强，但对单一大云厂商的议价依赖度相应较高。

Anthropic 则采用 “多云 + 新兴云厂” 并举的分散化采购策略，算力来源更为多元。谷歌云约 2,000 亿美元（5 年、5GW）与 AWS 约 1,000 亿美元（10 年、5GW）合计占比约 58%，为核心算力底座；同时积极引入新兴算力供给方 ——Fluidstack 约 500 亿美元、Nscale 约 450 亿美元（6 年、1.35GW）、Lambda 约 350 亿美元、Volta 约 100 亿美元，新兴云厂合计占比约 27%，并向 xAI（SpaceX）租用约 450 亿美元（3 年）算力，形成 “巨头云保底 + 新云厂扩容 + 对手产能补充” 的复合供给结构，降低了单一供应商依赖。

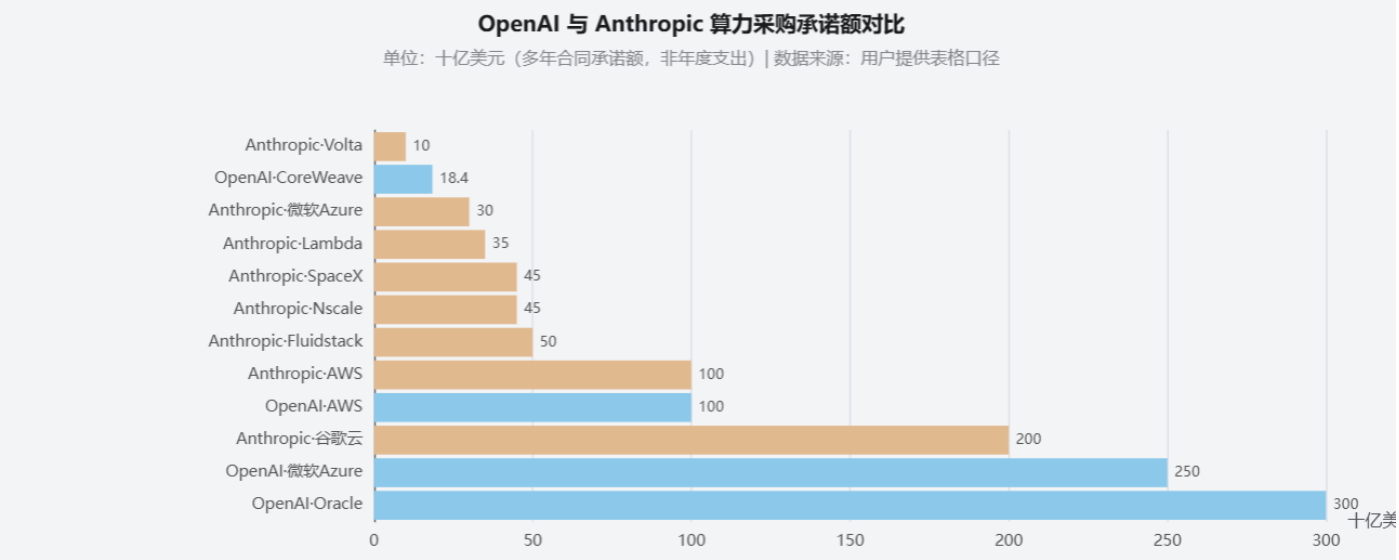
表4：据不完全统计，OpenAI 算力采购合同达 6684 亿美元，Anthropic 达 5150 亿美元

公司	合作方	金额	期限	GW 数
OpenAI	Oracle	$300B	5 年	4.5GW
	微软 Azure	$250B		
	AWS	$100B	8 年	2GW
	CoreWeave	$18.4B	5 年	
	上述加总	**不完全统计，算力合同已达到 6684 亿美元**		
Anthropic	AWS	$100B	10 年	5GW
	谷歌云	$200B	5 年	5GW
	微软 Azure	$30B		1GW
	Lambda	$35B		
	Nscale	$45B	6 年	1.35GW
	SpaceX	$45B	3 年	
	Fluidstuck	$50B		
	Volta	$10B		
	上述加总	**不完全统计，算力合同已达到 5150 亿美金**		

数据来源：OpenAI 官网、Anthropic 官网、Cayman Jounal 等、开源证券研究所

下载日志已记录，仅供内部参考

图19：OpenAI 算力合作方主要为甲骨文、微软、AWS，Anthropic 主要合作方为谷歌、AWS、SpaceX 和新云厂

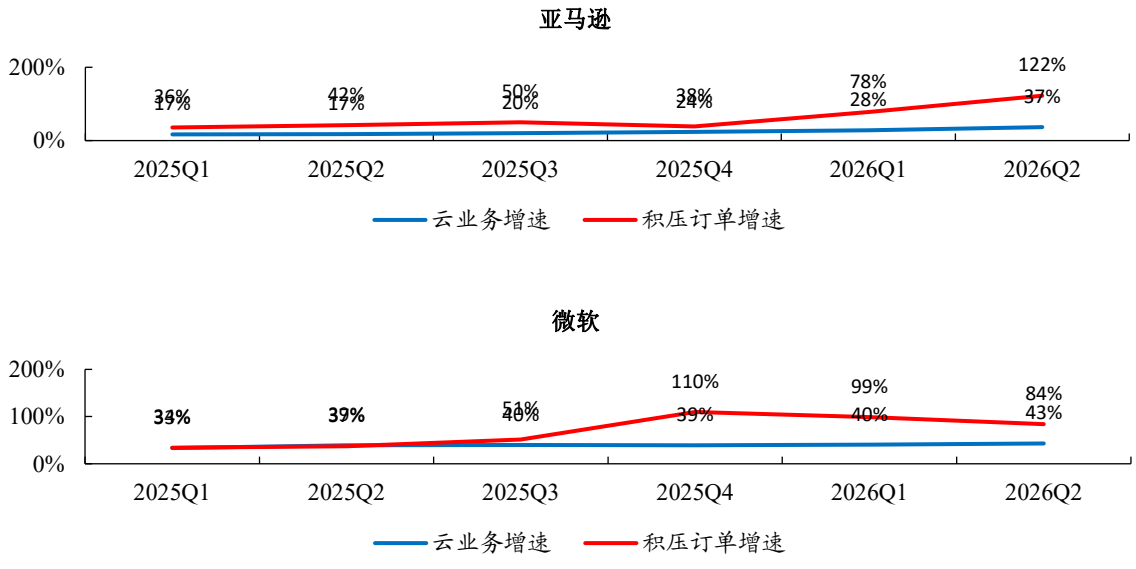


资料来源：OpenAI 官网、Anthropic 官网、Cayman Jounal 等、开源证券研究所

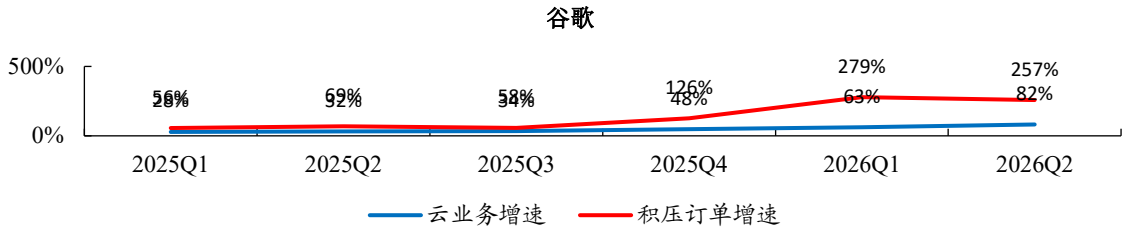
4.3、 云厂：行业增速二阶导向上，ROIC 及回本周期显示经营向上，关注信用风险指标

前瞻层给出强劲的“加速”证据，积压订单增速和云收入增速剪刀差走阔。三家海外头部云厂在手订单合计约 1.69 万亿美元（AWS 4,960 亿 + 微软 6,780 亿 + 谷歌 5,140 亿），2026Q2 订单 YOY 分别为 +122%、+84%、+257%，而当期收入 YOY 仅 36.7%、43%、82%，剪刀差分别达 +85pct、+41pct、+175pct，且 QOQ 仍全部为正。订单蓄水池加深速度远快于收入确认速度，意味着已签未确认收入在累积，这是判断“未来继续加速”的重要领先证据。

图20：2026Q2 亚马逊、微软、谷歌的积压订单增速分别为 122%、84%、257%，远高于云业务的收入增速

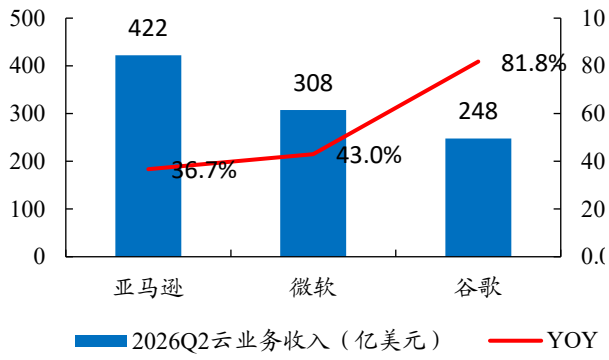


下载日志已记录，仅供内部参考



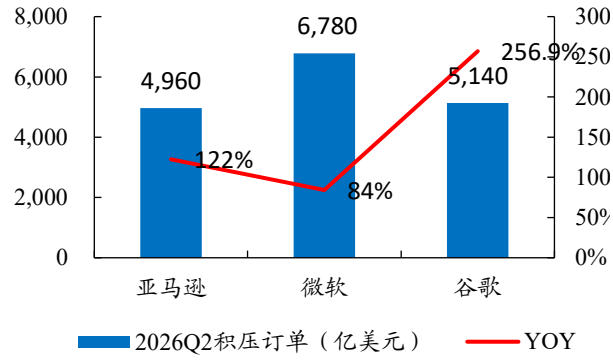
数据来源：各家公司公告、开源证券研究所

图21：谷歌 2026Q2 云业务收入为 248 亿美元，同比增长 81.8%



数据来源：各家公司公告、开源证券研究所

图22：谷歌 2026Q2 积压订单为 5140 亿美元，同比增长 256.9%



数据来源：各家公司公告、开源证券研究所

2026Q2 北美云厂呈现典型的经营杠杆持续释放特征，盈利扩张与需求放量形成正向共振。当季 AWS、Google Cloud 收入增速分别由 28%、63% 进一步抬升至 36.7%、82%（二阶导向上），同期云业务经营利润率分别环比提升 2.6pct 至 39.4%、35.6%，同比改善约 6.4pct、14.8pct；微软（Azure 及其他企业服务）OPM 为 40.6%，环比微升 0.9pct、同比基本持平。收入加速与利润率扩张同向，表明云业务成本增速持续低于收入增速，规模经济正转化为实实在在的利润弹性，而非以价换量式增长。**利润率提升主要来自三方面：**

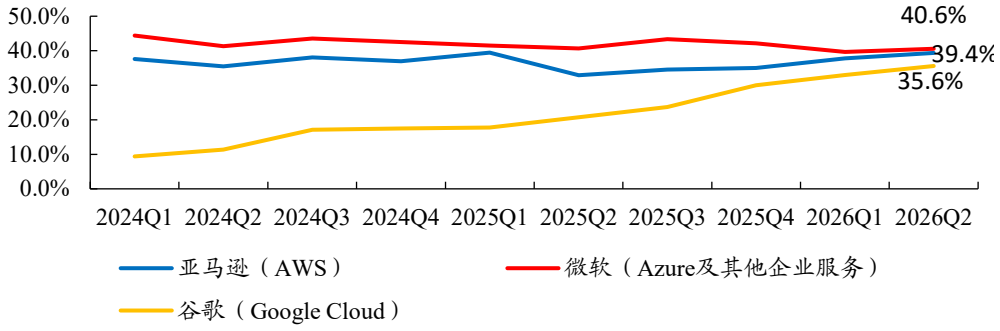
一是规模效应摊薄固定成本与产能利用率爬坡降低单位算力成本。数据中心、网络基础设施与研发投入具有刚性，AI 云需求放量使收入基数快速扩张，单位收入分摊的折旧、人力与研发费用下降，是经营杠杆释放最直接的来源。前期资本开支形成的新建产能逐步上线、利用率持续爬坡，单位算力对应的折旧与电力成本被更大的产出摊薄。

二是 AI 高价值负载优化收入结构。AI 训练、推理及 GPU 加速实例的单位价格显著高于传统通用算力，随着 AI 相关收入占比提升，云业务的综合客单价与毛利结构上移，形成“产品 mix 改善 ->收入提速 ->利润率扩张”的正循环。

三是自研芯片规模化。AWS Trainium、谷歌 TPU 等自研 ASIC 进入规模部署，降低了对外部高价 GPU 的依赖。

下载日志已记录，仅供内部参考

图23：2026Q2 亚马逊、微软、谷歌的云业务营业利润率分别为 39.4%、40.6%、35.6%



数据来源：各家公司公告、开源证券研究所

亚马逊对 AI 硬件投资给出可验证的回本周期。亚马逊将两类资产分账核算：服务器及网络设备平均不到 3 年即收支平衡，而大部分算力合同期限在 5 年以上、设备寿命 5–6 年，意味着回本后仍有 2–3 年的净现金回收窗口；寿命 30 年以上的数据中心可承载5-6代服务器，后续代际无需重复前期投入，回报曲线逐代优化。
微软将数据中心折旧年限从 15 年上调至 25 年（FY2027 起，2026/6/30），约三分之二资本开支配置可快速收缩的 GPU/CPU 短期资产，通过 Copilot 等高毛利上层产品和自研芯片 Maia 降本，管理全周期回报。谷歌在手订单将在 24 个月内确认超 50%、可将硬件涨价传导至定价。

表5：亚马逊明确硬件（服务器及网络设备）回本周期不到 3 年，大部分算力合同为 5 年以上

公司	业绩会时间	回本周期	资产寿命 / 折旧口径	核心 ROI 逻辑
亚马逊	2026/7/30	服务器及网络设备平均不到 3 年收支平衡，大部分算力合同期限 5 年以上，有 2-3 年的现金回收期	服务器及网络设备 5–6 年；数据中心寿命 30 年以上、承载 5–6 代	两类资产分账；AI 回报率与核心业务同期一致、略领先
微软	FY26Q4，2026/7/30	未给具体年限	数据中心折旧 15→25 年（FY27 起），资本支出中的 2/3 用于 GPU 和 CPU 等短期资产，剩余用于大型数据中心场地等长期开支	高毛利上层产品优先、自研芯片 Maia 降本、短期资产弹性收缩
谷歌	2026/7/22	未给具体年限；backlog 24 个月内确认超 50%	单季投入 60% 服务器、40% 数据中心及网络	有可观回报即持续投入，硬件涨价传导至定价

资料来源：各公司业绩会、开源证券研究所

投入层与前瞻层相互验证。2026Q2 海外四家 CSP 厂商（亚马逊、微软、谷歌、Meta）单季 capex 合计约 1,688 亿美元，同比增长 78%，其中谷歌翻倍增长。全年指引亚马逊上调至 2,200 亿、谷歌上调至 1,950–2,050 亿且明确“2027 显著高于 2026”、Meta 下限上调。供给端主动上修，与订单端方向一致。根据 Bloomberg 一致预期，2027-2029 年海外五家龙头（亚马逊、微软、谷歌、Meta、甲骨文）资本开支合计将达到 11049、12027、11802 亿美元，同比增长 31.2%、8.9%、-1.9%。

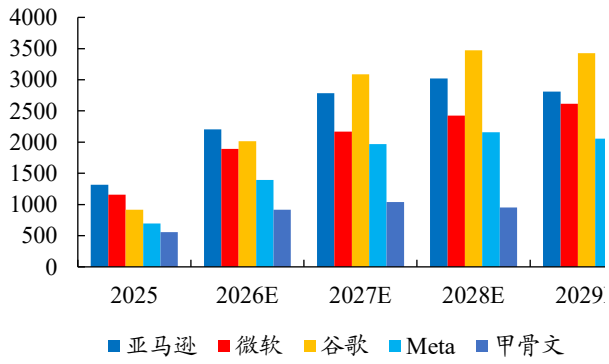
下载日志已记录，仅供内部参考

图24：2026Q2 四家 CSP 厂商（亚马逊、微软、谷歌、Meta）单季 capex 合计约 1,688 亿美元，同比增长 78%

资本开支（百万美元）	2025Q1	2025Q2	2025Q3	2025Q4	2026Q1	2026Q2	2026Q3E	2026Q4E	2027Q1E	2027Q2E
亚马逊	24,255	31,368	34,228	38,469	43,234	51,757	58,172	63,827	63,107	69,497
YOY	68%	83%	55%	62%	78%	65%	70%	66%	46%	34%
微软	21,400	24,200	34,900	37,500	31,900	41,000	44,993	46,308	48,722	51,416
YOY	53%	27%	75%	66%	49%	69%	29%	23%	53%	25%
谷歌	17,197	22,446	23,953	27,851	35,674	44,924	55,632	65,196	67,463	73,966
YOY	43%	70%	83%	75%	107%	100%	132%	134%	89%	65%
Meta	13,692	17,012	19,374	22,137	19,840	31,078	40,828	49,530	45,672	50,736
YOY	103.9%	100.8%	110.5%	49.2%	44.9%	82.7%	111%	124%	130%	63%
合计	76,544	95,026	112,455	125,957	130,648	168,759	199,626	224,861	224,963	245,615
YOY	64%	67%	77%	62%	71%	78%	78%	79%	72%	46%

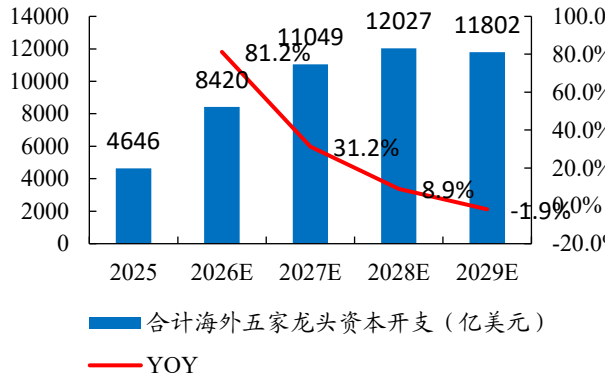
数据来源：各公司公告、Bloomberg、开源证券研究所

图25：资本开支：海外五家龙头公司预计资本开支维持增长（单位：亿美元）



数据来源：Bloomberg、开源证券研究所（微软为财年数据，2026 财年以 2026/6/30 为结尾）

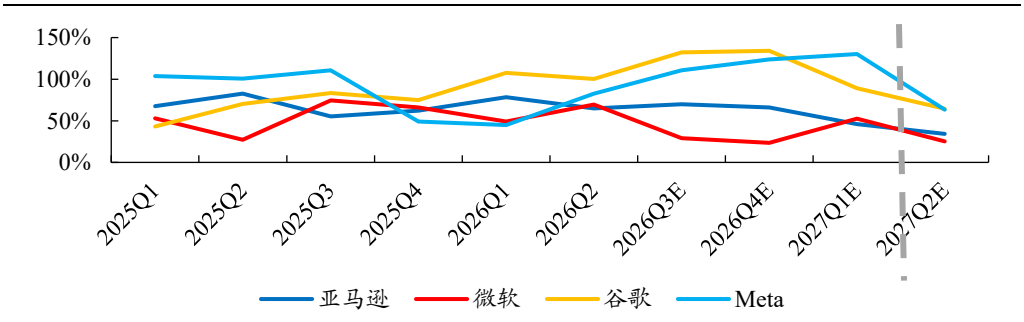
图26：预计 2028 年海外五家龙头的资本开支仍将保持增长



数据来源：Bloomberg、开源证券研究所（微软为财年数据，2026 财年以 2026/6/30 为结尾）

海外云厂收入加速期有望持续至 2028 年下半年。根据 Bloomberg 一致预期，海外 CSP 厂商的高速资本开支扩张将维持至 2027Q1，后续因高基数原因增速逐步平稳。根据我们在《云计算行业深度报告：AI 驱动变革，云计算进入新一轮扩张周期》中得出的结论：”云计算从 Capex 提速到收入兑现存在约 1-1.5 年时间差”，我们判断海外云厂收入加速期有望持续至 2028 年下半年，并从 2027 年下半年开始，伴随资本开支的平稳而迎来利润加速释放周期。

图27：资本开支增速：海外 CSP 厂商的资本开支增速预计在 2027Q2 后放缓



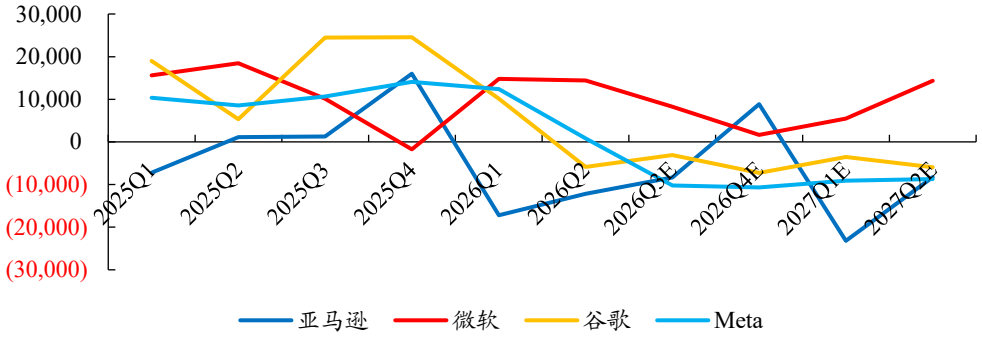
数据来源：Bloomberg、开源证券研究所

下载日志已记录，仅供内部参考

AI 资本开支高增正在加剧云厂商之间的财务与信用分化，自由现金流、债券融资与 CDS 定价三个维度相互印证。

微软内生造血能力最强、重资产扩张者外部融资与信用压力上升。从自由现金流看，在四大厂商同步加大 AI 资本开支的背景下，微软是 2026 年以来唯一持续维持正向自由现金流、且近期重新加速增长的厂商，而亚马逊、谷歌、Meta 的季度自由现金流已先后转负，反映其资本开支前置正在吞噬经营性现金。微软已构建“算力底座 —Azure 云服务 —Copilot 企业软件”的三层变现体系，2026Q2 Azure 年度营收规模突破 1,000 亿美元、Microsoft 365 Copilot 付费席位突破 3,000 万，商业化回款对资本开支形成了更有效的覆盖。

图28：自由现金流：微软仍维持正向的自由现金流，其余 CSP 厂商转负



数据来源：各公司公告、Bloomberg、开源证券研究所

为弥补资本开支造成的现金缺口，其余厂商转向外部举债，融资强度显著抬升。2025 年 11 月至今，亚马逊、Alphabet、Meta 分别新发债券 44、64、6 只，合计发行规模达 2,381 亿美元（分别为 1,072.8 亿、1,058.1 亿、250.0 亿美元），而微软年内未新发债券，“内生现金流覆盖开支”与“依赖债券融资扩张”的路径差异清晰可见。

表6：2025 年 11 月至今亚马逊、谷歌、Meta 合计发行债券金额 2681 亿美元，微软未新发债券

公司	2026 年表内新发债券数	2026 年初至今发行金额合计（十亿美元）
Amazon	44	$107.28
Alphabet	64	$105.81
Meta	6	$25.00
Microsoft	—	0
合计	**114**	**$238.09**

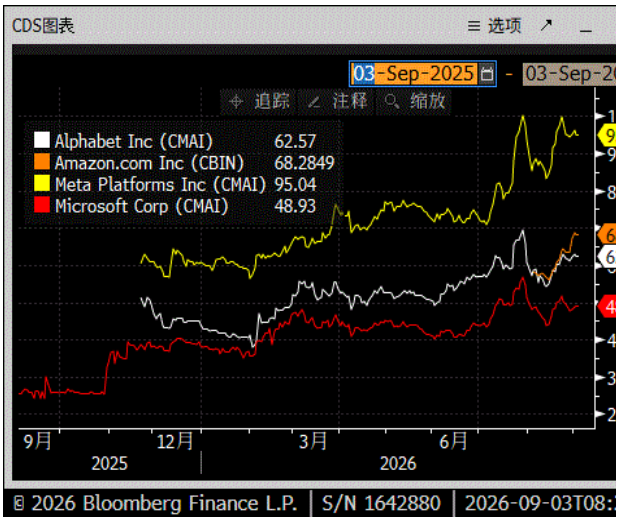
数据来源：Bloomberg、开源证券研究所

信用市场已对这种分化作出定价。五年期 CDS 方面，截至 2026 年 9 月 3 日，科技巨头内部呈明显梯度：微软最低，仅 48.93bp，Alphabet、亚马逊、Meta 依次

下载日志已记录，仅供内部参考

为 62.57bp、68.28bp、95.04bp，Meta 较微软高出约 46bp；而高杠杆的算力链主体信用利差显著更宽 ——CoreWeave 高达约 839bp，约为微软的 17 倍，甲骨文为 212.47bp，约为微软的 4.3 倍。CDS 本质是市场对违约风险的定价，新兴 GPU 云厂商（如 CoreWeave）单一客户依赖度高、负债重、回款周期长，其高企的信用利差反映出市场对其激进扩张模式可持续性的担忧；相比之下，超大规模云厂商凭借多元业务与强融资能力，CDS 仍处 100bp 以内的低风险区间。

图29：5 年期 CDS：截至 2026 年 9 月 3 日 Meta 为 95.04，微软为 48.93



资料来源：Bloomberg

图30：5 年期 CDS：截至 2026 年 9 月 3 日 Coreweave 为 839.05，甲骨文为 212.47



资料来源：Bloomberg

整体看，本轮 AI 资本开支竞赛已从“谁投得多”进入“谁能更稳健地为投入融资”的阶段。微软以正向自由现金流、零新增发债与最低 CDS，成为资金与信用市场共同认可的财务最稳健标的；亚马逊、谷歌、Meta 虽凭借巨头资质仍能以低成本发债、风险整体可控，但杠杆水平与外部融资依赖度在抬升；真正的信用脆弱点集中于 CoreWeave、甲骨文等高杠杆算力供应商。后续建议将“自由现金流能否回正、发债节奏与利率成本、CDS 利差变化”作为衡量 AI 基建扩张可持续性的核心跟踪指标。

4.4、 新云厂：收入与积压订单快速增长，新签订单回本周期缩短至 2 年内

新云厂收入延续三位数高增长，Nebius 的盈利修复斜率显著快于 CoreWeave。

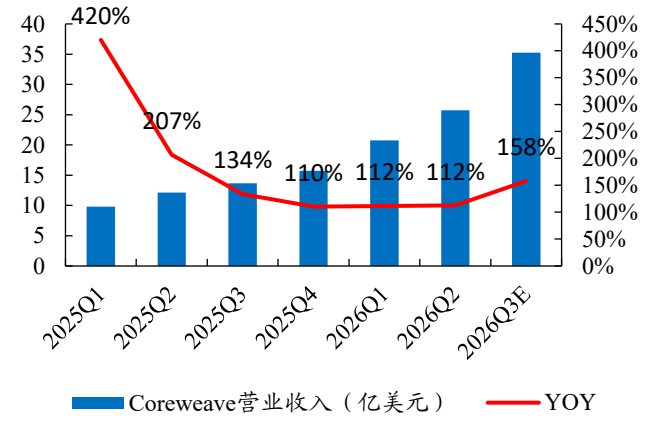
收入端：CoreWeave 2026Q2 营收 25.75 亿美元、同比增长 112%，2026Q3 指引同比增速上升至 158%，环比重新加速、规模再上台阶。Nebius 2026Q2 营收约 5.82 亿美元、同比增长 454%，体量约为 CoreWeave 的四分之一，但季度同比增速中枢更高（多在 400%-680% 区间），仍处扩张早期。截至 2026 年 6 月 Nebius 的年运行收入达到 30 亿美金，预计到 2026 年底将提升至 70-90 亿美金。

盈利端：两家调整后 EBITDA 盈利能力强、但调整后净利润仍亏损，呈现投入期重资产模式特征。2025Q2 以来 CoreWeave 调整后 EBITDA 利润率连续多季度稳定于 56%-62%（2026Q2 为 59%），但调整后亏损率仍达 -22%，反映资本开支扩

下载日志已记录，仅供内部参考

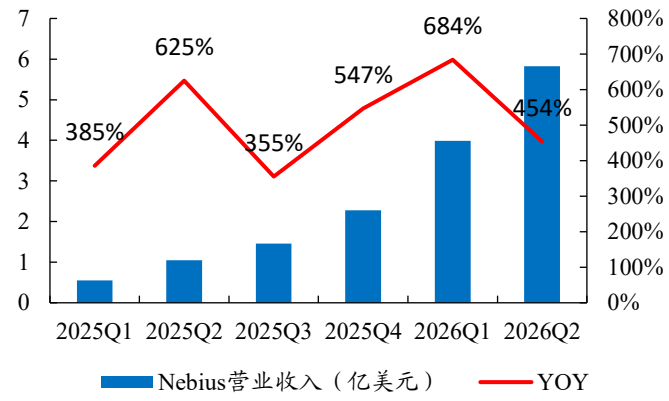
张期高杠杆下财务费用对盈利质量的压力。Nebius 则沿经营杠杆曲线快速修复，调整后 EBITDA 利润率抬升至 2026Q2 的 41%，AI 业务 EBITDA margin 达到 50%，调整后亏损率同步由 2025 年 Q2 的 -87% 大幅收窄至 2026 年 Q2 的-6%，已逼近盈亏平衡。

图31：2026Q2 Coreweave 收入为 25.75 亿美元，同比增长 112%



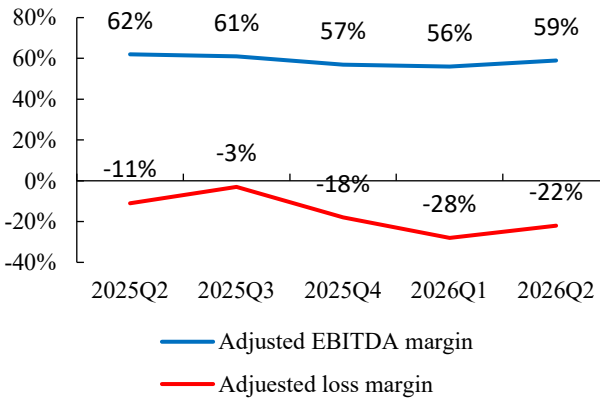
数据来源：Coreweave 公司公告、Wind、开源证券研究所

图32：2026Q2 Nebius 收入为 5.82 亿美元，同比增长 454%



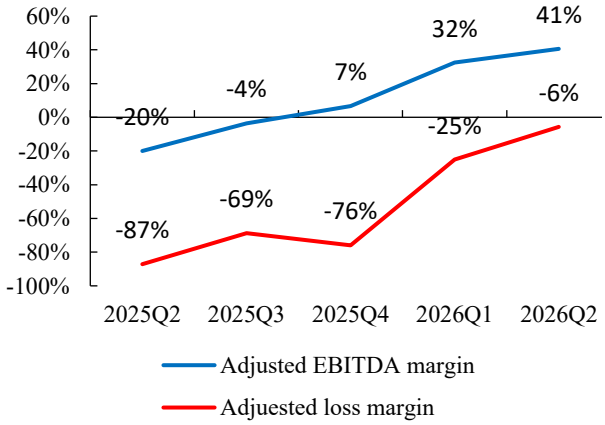
数据来源：Wind、开源证券研究所

图33：2026Q2 Coreweave 的 Adjusted EBITDA margin 为 59%、Adjusted loss margin 为-22%



数据来源：Coreweave 公司公告、开源证券研究所

图34：2026Q2 Nebius 的 Adjusted EBITDA margin 为 41%、Adjusted loss margin 为-6%



数据来源：Nebius 公司公告、开源证券研究所

新云厂在手订单规模同比快速增长、订单盈利质量同步提升，为高增长与现金流修复提供了高确定性支撑。

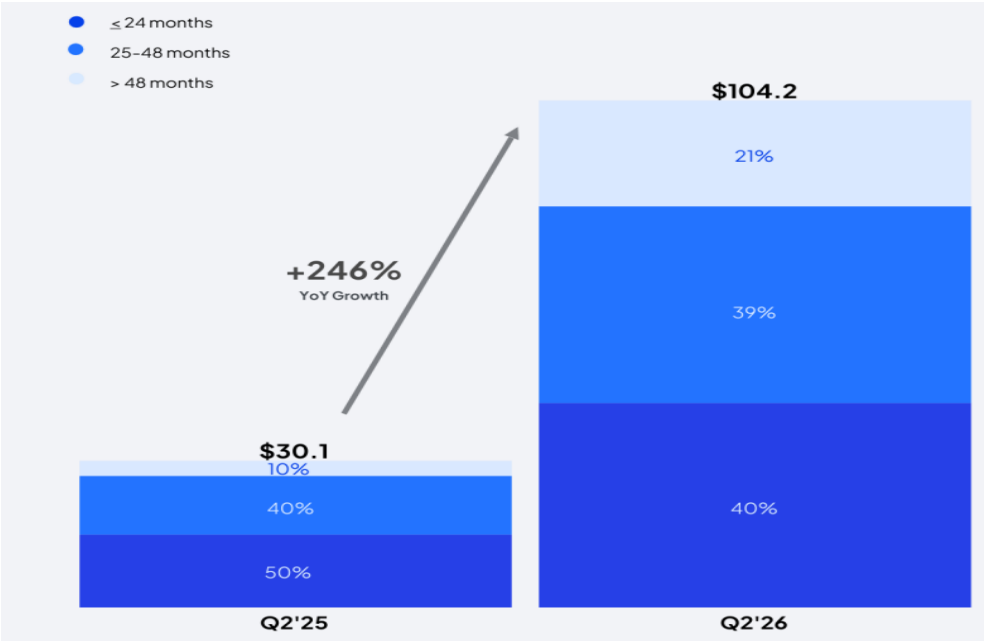
积压订单：截至 2026Q2，CoreWeave 积压订单达 1,042 亿美元、同比增长 246%（2025Q2 仅 301 亿美元）；Nebius 积压订单亦超过 400 亿美元，收入能见度在高基数上进一步抬升。更关键的是订单期限结构持续优化：CoreWeave 期限 25 个月以上的合同占比由 2025Q2 的 50% 提升至 2026Q2 的 60%，其中 48 个月以上长约占比由 10% 翻倍至 21%，长周期合约不仅锁定多年收入，也与数据中心、服务

下载日志已记录，仅供内部参考

器的重资产折旧周期更好匹配，降低产能闲置与续约风险。

订单盈利能力： Nebius 的签约结构最具代表性，单位产能年化合同价值（ACV/GW）由 2026 年初的约 120 亿美元，升至 Q2 新签的超 200 亿美元、Q3 短期容量协议进一步超 400 亿美元，量价齐升凸显算力卖方议价权；Q2 签约总合同额环比增长 4 倍，斩获 Reflection、Cohere 等 4 笔标志性大单、平均单笔规模超 10 亿美元，且新签订单回本周期仅 1 年 10 个月、50%-60% 的资本开支由客户预付款自筹。新云厂并非单纯依赖外部融资“烧钱换规模”，其订单端已具备较强盈利质量与一定自我造血能力，为后续收入兑现和自由现金流改善奠定坚实基础。

图35：截至 2026Q2 Coreweave 的积压订单为 1042 亿美元，同比增长 246%



资料来源：Coreweavw 公司公告

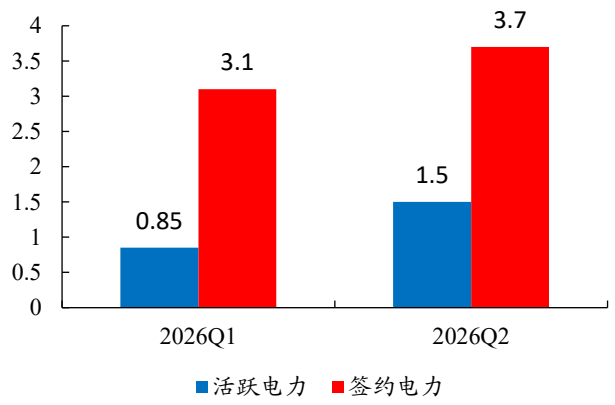
两家新云厂签约电力在一年内连续上台阶，共同印证 AI 算力需求高景气与新云厂算力变现能力的同步上修。

CoreWeave: 截至 2026/8/11 签约电力突破 4.2GW，在线活跃容量 1.5GW，千亿美元级积压订单为后续算力建设与配套供电需求提供强支撑；

Nebius: 2026 年底签约电力目标上调至 5GW，预估 2027 年年新增超 1GW，公司 Q2 达成单兆瓦超 2000 万至 2500 万美元的中长期合同，短期高价值训练合同定价达 4000 万至 5000 万美元/兆瓦，AI 算力变现能力持续上修。

下载日志已记录，仅供内部参考

图36：截至 2026Q2，Coreweave 的签约电力达到 3.7GW（单位：GW）



数据来源：Coreweave 公司公告、开源证券研究所

图37：Nebius 2026 年底签约电力目标上调至 5GW



资料来源：Nebius 公司公告

4.5、 主权基金：中东及亚洲主权资本加速布局，AI 算力相关规划已超 20GW

主权资本正由财务投资者向 AI 基础设施"资本+能源+算力"组织者转变。据不完全统计，PIF（HUMAIN）、G42、MGX/Mubadala 及 GIC 关联项目的披露口径已达 20GW 以上。

沙特HUMAIN 为核心形成全国级算力平台。HUMAIN 规划口径为 2030 年 1.9GW、2034 年 6GW；项目端同步推进 AMD/Cisco 与 STC 各 1GW 合作、AWS 最高 50MW AI Zone、Together AI 250MW、DataVolt Oxagon 1.5GW 项目，以及利雅得东部 6GW 园区招标。

阿联酋主权资本布局兼具全球并购与本土枢纽建设。G42 牵头的 Stargate UAE 园区规划约 5GW，首阶段 200MW 预计于 2026 年上线；MGX 等参与收购的 Aligned Data Centers 拥有 51 个园区、运营及规划容量约 6.4GW；Campus AI 法国项目旗舰园区为 1.4GW、全法网络远期最高 3GW；日本秋田约 500MW 项目仍处拟议阶段。

新加坡 GIC 通过成熟数据中心平台获取全球算力基础设施敞口。其与 Equinix、CPP Investments 设立的美国 xScale 合资平台完全建成后预计新增超过 1.5GW 容量；Vantage APAC 相关平台容量约 1GW，覆盖马来西亚柔佛等亚太节点。相较中东基金主导国家级园区建设，GIC 更偏向通过专业运营商和区域平台参与算力扩张。

表7：据不完全统计，主权资本 AI 算力相关采购合同及规划规模已超 20GW

公司/平台	合作方/项目	国家/地区	金额/进展	期限	GW 数
PIF（HUMAIN）	AMD/Cisco	沙特	up to $10B	2030	1.0
	STC	沙特	未披露	—	1.0
	Together AI	沙特	>$5B 年化收入	—	0.25
	AWS AI Zone	沙特	>$5B；最高 50MW	2028	0.05
	DataVolt（Oxagon）	沙特	$5B	2028	1.5

下载日志已记录，仅供内部参考

	利雅得东部 6GW 园区	沙特	未披露	招标中	6.0
	HUMAIN 全国规划口径	沙特	1.9GW by 2030 / 6GW by 2034	—	1.9 / 6.0
G42	Stargate UAE	阿联酋	$30B+；首阶段 200MW	2026 / 2030	5.0
MGX / Mubadala	Aligned Data Centers	美国	$40B + $5B	2026-07	6.4
	Campus AI 法国	法国	€7.5B	2028	1.4 / 3.0
	日本秋田数据中心（拟议）	日本	$6.3B	early 2030s	0.5
GIC	Equinix xScale JV	美国	>$15B	滚动	>1.5
	Vantage APAC	亚太（马来西亚等）	$1.6B	2025Q4	1.0

资料来源：NextFin、EnterpriseAM、MEED、優分析、复旦发展研究院、华尔街见闻、AI2Work、CIO Middle East、Yahoo Finance 、GIC 官网、开源证券研究所

5、 投资建议

GLM-5.3 系列与 GPT-Astra 的接连突破，不仅是参数和榜单数字的跃升，标志着模型进入帕累托最优与前沿智能提升的双向突围，有力地证明了开源模型的天花板同样很高，国产模型在国际竞争力进一步提升，在性能和成本的平衡下有望打开全球 AI 需求天花板。

模型端：需求天花板提升，推荐标的智谱、德适-B，受益标的 MiniMax-W。

算力端：多模态落地与 Agent 规模化带动训练、推理算力需求长期刚性扩张，推荐标的阿里巴巴-W、百度集团-SW、联想集团、中芯国际，受益标的腾讯控股、天数智芯。

表8：核心标的盈利预测及估值表

代码	公司	市值（亿港币）	财报货币	营收（亿元）			归母净利润（亿元）			估值方法	估值			评级
				2026E	2027E	2028E	2026E	2027E	2028E		2026E	2027E	2028E	
2513.HK	智谱	3,618	CNY	62.7	195.6	360.1	-41.0	-20.5	-2.0	PS	49.2	15.8	8.6	买入
2526.HK	德适-B	365	CNY	4.9	9.3	17.2	-0.8	-0.7	0.1	PS	63.5	33.6	18.1	买入
0100.HK	MiniMax-W	1,000	USD	4.2	11.6	23.4	-7.6	-9.2	-8.8	PS	30.7	73.7	36.5	未评级
9988.HK	阿里巴巴-W	22,828	CNY	10865	11996	14573	1087	1439	1894	PE	17.9	13.5	10.3	买入
9888.HK	百度集团-SW	2,436	CNY	1313	1391	1498	137	159	170	PE	15.1	13.1	12.2	买入
0992.HK	联想集团	4,561	USD	1000	1231	1387	40	56	64	PE	14.5	10.4	9.1	买入
0700.HK	腾讯控股	41,074	CNY	8314	9154	9978	2329	2378	2603	PE	15.1	14.7	13.5	买入
9903.HK	天数智芯	1,126	CNY	29	69	113	-4	8	26	PE	-221.2	114.5	37.1	未评级
0981.HK	中芯国际	7,495	USD	118	142	168	12	17	23	PE	80.6	56.8	42.2	买入

数据来源：Wind、开源证券研究所，注未评级标的盈利预测为 Wind 一致性预期，已评级标的为开源证券研究所预测，汇率：CNYHKD=1.1714，USDHKD=7.8452。

6、 风险提示

软件及产品推出不及预期：科技行业软件更新较快，市场竞争激烈，若产品推出不及时或不及预期，或影响企业未来竞争力和创收情况。

产能及供应链风险：若产能爬坡速度较慢，供给端受限，创收弹性无法充分释

下载日志已记录，仅供内部参考

放，则相关行业产业链业绩或受到影响。

监管政策变动：科技政策及垂类行业相关政策的变动，或将影响公司增长景气度及战略决策，进一步影响公司的增长节奏。

宏观经济增长放缓：若宏观经济增长放缓影响终端消费需求、企业 IT 支出等，将影响产业链相关企业的创收能力。

地缘政治风险：地缘政治变动或伴随国际政策调整，或对科技行业供应链等相关企业带来限制。

下载日志已记录，仅供内部参考

特别声明

《证券期货投资者适当性管理办法》、《证券经营机构投资者适当性管理实施指引（试行）》已正式实施。根据上述规定，开源证券评定此研报的风险等级为R4（中高风险），因此通过公共平台推送的研报其适用的投资者类别仅限定为境内专业投资者及风险承受能力为C4、C5的普通投资者。若您并非境内专业投资者及风险承受能力为C4、C5的普通投资者，请取消阅读，请勿收藏、接收或使用本研报中的任何信息。

因此受限于访问权限的设置，若给您造成不便，烦请见谅！感谢您给予的理解与配合。

分析师声明

本研究报告的署名人员具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，以勤勉的职业态度，独立、客观地出具本报告，并对内容和观点负责。本报告清晰准确地反映了署名人员的研究观点，所包含的分析基于各种假设，不同假设可能导致分析结果出现重大不同。本报告采用的各种估值方法及模型均有其局限性，估值结果不保证所涉及证券能够在该价格交易。本报告署名人员保证他们报酬的任何一部分不曾与，不与，也将不会与本报告中具体的推荐意见或观点有直接或间接的联系。

股票投资评级说明

	评级	说明
证券评级	买入（Buy）	预计相对强于市场表现 20%以上；
	增持（Outperform）	预计相对强于市场表现 5%～20%；
	中性（Neutral）	预计相对市场表现在－5%～＋5%之间波动；
	减持（Underperform）	预计相对弱于市场表现 5%以下。
行业评级	看好（Overweight）	预计行业超越整体市场表现；
	中性（Neutral）	预计行业与整体市场表现基本持平；
	看淡（Underperform）	预计行业弱于整体市场表现。
备注：评级标准为以报告日后的 6~12 个月内，证券相对于市场基准指数的涨跌幅表现，其中 A 股基准指数为沪深 300 指数（北交所基准指数为北证 50 指数）、港股基准指数为恒生指数、新三板基准指数为三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的）、美股基准指数为标普 500 或纳斯达克综合指数。我们在此提醒您，不同证券研究机构采用不同的评级术语及评级标准。我们采用的是相对评级体系，表示投资的相对比重建议；投资者买入或者卖出证券的决定取决于个人的实际情况，比如当前的持仓结构以及其他需要考虑的因素。投资者应阅读整篇报告，以获取比较完整的观点与信息，不应仅仅依靠投资评级来推断结论。		

下载日志已记录，仅供内部参考

法律声明

开源证券股份有限公司是经中国证监会批准设立的证券经营机构，已具备证券投资咨询业务资格。

本报告仅供开源证券股份有限公司（以下简称“本公司”）的客户使用。本公司不会因接收人收到本报告而视其为客户。本报告是发送给开源证券客户的，属于商业秘密材料，只有开源证券客户才能参考或使用。

本报告是基于本公司认为可靠的已公开信息，但本公司不保证该等信息的准确性或完整性。本报告所载的资料、工具、意见及推测只提供给客户作参考之用，并非作为或被视为出售或购买证券或其他金融工具的邀请或向人做出邀请。本报告所载的资料、意见及推测仅反映本公司于发布本报告当日的判断，本报告所指的证券或投资标的的价格、价值及投资收入可能会波动，过往的业绩表现不应作为其日后表现的预示。在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。客户应当考虑到本公司可能存在可能影响本报告客观性的利益冲突，不应视本报告为做出投资决策的唯一因素。本报告中所指的投资及服务可能不适合个别客户，不构成客户私人咨询建议。本公司未确保本报告充分考虑到个别客户特殊的投资目标、财务状况或需要。本公司建议客户应考虑本报告的任何意见或建议是否符合其特定状况，以及（若有必要）咨询独立投资顾问。在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议。在任何情况下，本公司不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任。若本报告的接收人非本公司的客户，应在基于本报告做出任何投资决定或就本报告要求任何解释前咨询独立投资顾问。投资者应自主作出投资决策并自行承担投资风险，任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。

开源证券在法律允许的情况下可参与、投资或持有本报告涉及的证券或进行证券交易，或向本报告涉及的公司提供或争取提供包括投资银行业务在内的服务或业务支持。开源证券可能与本报告涉及的公司之间存在业务关系，并无需事先或在获得业务关系后通知客户。

本报告的版权归本公司所有。本公司对本报告保留一切权利。未经本公司事先书面授权，本报告的任何部分均不得以任何方式制作任何形式的拷贝、复印件或复制品，或再次分发给任何其他人，或以任何侵犯本公司版权的其他方式使用。所有本报告中使用的商标、服务标记及标记均为本公司的商标、服务标记及标记。

开源证券研究所

上海
地址：上海市浦东新区世纪大道1788号陆家嘴金控广场1号楼3层
邮编：200120
邮箱：research@kysec.cn

深圳
地址：深圳市福田区金田路2030号卓越世纪中心1号楼45层
邮编：518000
邮箱：research@kysec.cn

北京
地址：北京市西城区西直门外大街18号金贸大厦C2座9层
邮编：100044
邮箱：research@kysec.cn

西安
地址：西安市高新区锦业路1号都市之门B座5层
邮编：710065
邮箱：research@kysec.cn