综合

2026 年 11 月 8 日

投资评级：看好（维持）

海外科技：算力交付不确定性上升，Anthropic 短期算力成本承压，多模态能力持续升级

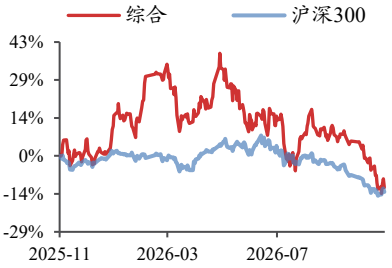
——行业周报

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

叶彬慧（分析师）
yebinhui@kysec.cn
证书编号：S0790526070007

李秋莹（联系人）
liqiuying@kysec.cn
证书编号：S0790126060038

行业走势图



数据来源：聚源

相关研究报告

《海外科技：CSP 资本开支韧性延续，CPO 产业化提速—行业周报》-2026.11.1

《海外科技：Agent 能力向长程专业任务延伸，SpaceX 与 Meta 业绩均超预期—行业周报》-2026.10.25

《海外科技：AMD 推出 Helion 机架级平台，ASML 评估进一步扩产—行业周报》-2026.10.18

● **大模型：Anthropic 算力投入与成本压力显现，多模态理解与生成能力升级**

本周 AI 商业化投入与多模态模型能力升级同步推进。Anthropic 更新 IPO 招股书，2026Q3 收入达 169.56 亿美元，环比增长约 47%，增速较 Q2 放缓；受额外算力交付及短期高价算力采购影响，调整后营业利润率由 19%降至 15%。截至 2026 年 9 月 30 日及期后，公司累计承诺支出约 8,954 亿美元，绝大部分涉及算力和基础设施。我们认为，Anthropic 收入规模持续扩大，但短期高价算力采购对利润率形成压力，叠加大规模长期基础设施承诺，未来资金需求持续增加。

模型能力方面，腾讯发布混元 Hy4，首次引入视觉理解与 Computer Use 能力；OpenAI 推出 GPT-Image-2.6；美团发布 LongCat-2.5。我们认为，近期更新体现了多模态理解、复杂任务执行与内容生成能力的同步升级，有望推动 AI 在专业办公、软件开发等场景进一步应用。

● **云计算：中期选举重塑数据中心政策预期，算力租赁进入交付确定性重定价**

中期选举提高数据中心建设的不确定性，算力交付确定性成为市场定价重点。民主党在德克萨斯、俄亥俄、佐治亚等数据中心重镇胜选，相关政策可能通过暂停或附加条件审批、提高电力与水资源成本承担要求，拉长项目建设周期并抬升开发成本。Neocloud 交付缺口大、现金流承压，受影响更明显：CoreWeave、Nebius、IREN 等公司大量已签约电力尚待转化为实际运营容量，面临更明显的通电与现金流压力；大型 CSP 凭借已建成算力、资金实力及跨区域布局，抗风险能力相对较强。我们认为，中期选举并未改变 AI 算力需求趋势，但市场将更重视项目按期通电和交付的确定性，数据中心成本分担与地方审批进程仍需持续关注。

● **投资建议：**

大模型与云计算：推荐标的：智谱、德适-B，受益标的：谷歌、亚马逊、微软、MiniMax。

● **风险提示：**模型及产品推出不及预期、AI 算力需求不及预期风险、算力供给超预期释放风险、宏观经济增长放缓。

目 录

1、 大模型：Anthropic 算力投入与成本压力显现，多模态理解与生成能力持续升级 3
1.1、 Anthropic：更新 IPO 招股书，收入增速放缓与算力成本压力显现 3
1.2、 腾讯：混元发布 Hy4，视觉理解与长程 Agent 能力升级，专业工作场景进一步拓展 3
1.3、 OpenAI：发布 GPT-Image-2.6，图像生成速度提升 50%，多轮编辑与 GIF 生成能力升级 5
1.4、 美团：发布 LongCat-2.5，原生多模态与长程 Agent 能力升级，国产算力训练及推理效率提升 5
2、 云计算：中期选举重塑数据中心政策预期，算力租赁进入交付确定性重定价 7
2.1、 民主党的数据中心立场：侧重成本内部化与地方同意 7
2.2、 对 Neocloud 与 CSP 的影响：交付缺口集中于 Neocloud，现金流约束放大延期影响 8
3、 港股周度更新 10
4、 模型与算力价格数据更新 12
4.1、 模型信息更新 12
4.2、 算力租赁价格更新 15
5、 投资建议 15
6、 风险提示 16

图表目录

图 1： Hy4 在多项可比评测中较 Preview 版本提升 4
图 2： LongCat-2.5 版本大幅提升，在 AutomationBench、BrowseComp 等评测中得分领先 6
图 3： 民主党赢得的州长选举地区多为数据中心重镇 8
图 4： CoreWeave 与 IREN 已投运容量距离近期目标容量仍存在差距（单位：GW） 9
图 5： Nebius 2026 年底 Connected Power 与签约电力目标对比（单位：GW） 9
图 6： CSP 与 Neocloud 股价在中期选举后出现分化 9
图 7： 本周恒生指数（2026.11.02-2026.11.06）在全球主要市场中涨跌幅靠前（单位：%） 10
图 8： 本周（2026.11.02-2026.11.06）恒生科技涨幅居首（%） 10
图 9： 本周（2026.11.02-2026.11.06）恒生传媒、恒生科技涨幅居前（%） 10
图 10： 2026 年 10 月以来港股通当日买入成交净额有所波动（单位：亿元） 11
图 11： 恒生沪深港通 AH 股溢价指数自高位回落 12
图 12： OpenRouter 的 token 调用量持续提升 14
图 13： 国产 AI 在 ZenMux 的 Token 市场份额约三分之二 14
图 14： B200 租赁价格高位运行（单位：$/小时） 15
图 15： H200 租赁价格高位运行（单位：$/小时） 15
图 16： H100 租赁价格震荡上行（单位：$/小时） 15
图 17： A100 租赁价格基本持平（单位：$/小时） 15

表 1： 本周（2026.11.02-2026.11.06）中芯国际、华虹宏力与中国海洋石油流入资金居前 11
表 2： 9 月以来新上线模型 12

1、 大模型：Anthropic 算力投入与成本压力显现，多模态理解与生成能力持续升级

1.1、 Anthropic：更新 IPO 招股书，收入增速放缓与算力成本压力显现

Anthropic 更新 IPO 招股书，2026Q3 收入环比增长约 47%，增速较 Q2 放缓。11 月 2 日，Anthropic 向美国 SEC 提交 S-1/A 修订版招股书，拟在纳斯达克全球精选市场上市，股票代码为 ANT。根据招股书，2026Q3 公司实现收入 169.56 亿美元，环比增长约 47%，低于 2026Q2 约 144%的环比增速。企业客户规模持续扩大，截至 2026 年 9 月 30 日，过去 12 个月贡献收入不低于 10 万美元、100 万美元及 1000 万美元的客户分别超过 11,900 家、1,900 家及约 250 家，净收入留存率（NRR）约 900%。

算力基础设施费用增加，GAAP 营业利润转负，调整后营业利润率下降。2026Q3 Anthropic 算力及基础设施费用为 98.63 亿美元，较 Q2 的 55.58 亿美元明显增加，主要受额外算力交付（含短期承诺算力）影响。公司披露，Q3 算力费用中有 50 亿美元来自短期承诺算力，其目前成本显著高于多年期承诺下采购的可比算力。同期 GAAP 营业利润由 Q2 的 9.74 亿美元转为亏损 5.67 亿美元，调整后营业利润由 22.35 亿美元增至 24.63 亿美元，但调整后营业利润率由 19%降至 15%。此外，Q3 营业利润受到 20 亿美元捐赠匹配计划相关非现金费用影响。

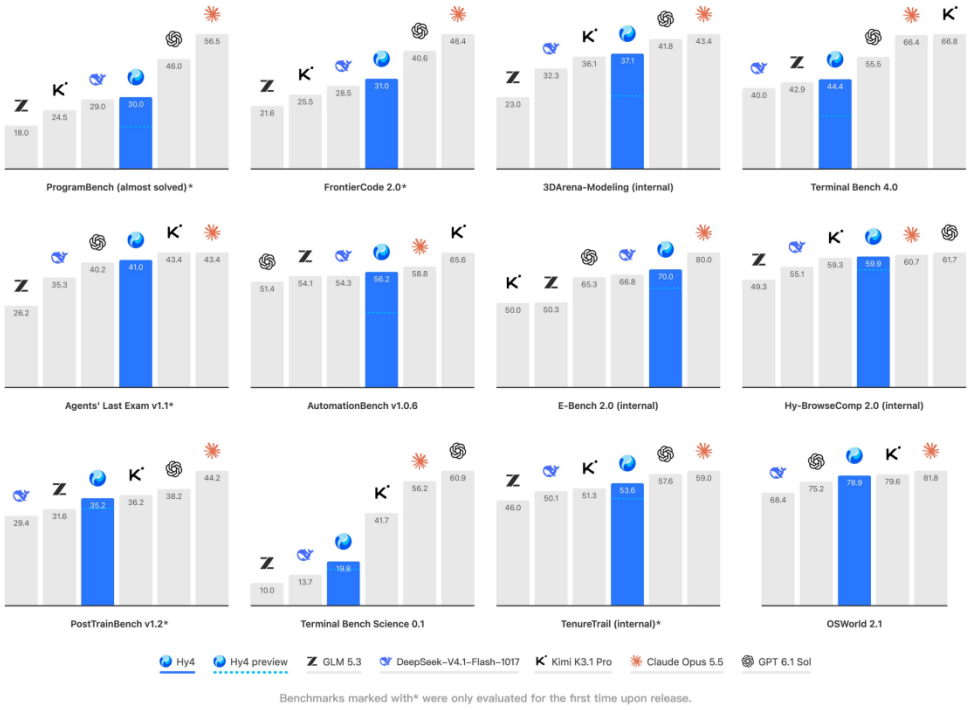
算力采购与基础设施承诺规模扩大，长期资本需求持续增加。根据 Anthropic 招股书，截至 2026 年 9 月 30 日及期后，Anthropic 签订的采购承诺、租赁协议及尚未起租租赁等累计承诺支出约 8,954 亿美元，绝大部分涉及算力和基础设施。其中，2030 年底前预计发生约 4,475 亿美元支出，其余支出绝大部分预计在 2031—2045 年发生；2026 年 9 月 30 日后 12 个月内到期约 698 亿美元。公司正由仅通过固定期限云协议采购算力的模式，逐步转向结合专用数据中心及直接租赁芯片的混合模式，并通过 AWS Trainium、Google TPU、NVIDIA GPU 及 AMD GPU 等多种芯片架构优化算力配置。

Anthropic 收入规模快速扩张，长期算力投入同步增加。公司企业客户持续增长，但短期高价算力采购推高基础设施费用，使调整后营业利润率环比下降。与此同时，大规模长期算力承诺及专用数据中心与直租芯片等基础设施投入增加了未来资金需求；收入增长、算力利用效率及长期合同兑现将成为后续盈利能力与现金流表现的重要影响因素。

1.2、 腾讯：混元发布 Hy4，视觉理解与长程 Agent 能力升级，专业工作场景进一步拓展

腾讯混元发布 Hy4，扩展预训练及后训练规模，强化视觉理解与长程 Agent 能力。11 月 4 日，腾讯混元发布新一代大语言模型 Hy4。相比 Hy4 Preview，Hy4 显著扩大预训练及后训练的数据与训练规模，首次引入视觉理解能力，并大幅提升指令遵循、长文理解和长程智能体任务表现。新模型结合视觉理解、推理与工具调用能力，可执行细粒度图像处理、复杂图表分析及数学视觉推理等任务。同时，Hy4 首次支持 Computer Use，能够结合命令行与图形界面操作。混元针对 Blender、CAD 等专业软件及跨应用长程任务开展专项训练，大幅提升专业任务的规划与执行能力。

图1：Hy4 在多项可比评测中较 Preview 版本提升



资料来源：腾讯官网

专业办公与金融分析能力提升，模型与产品协同优化工作流执行效果。Hy4 全面提升了 WorkBuddy 的用户体验。在 WorkBuddy 的评测中，Hy4 平均效果得分为 83.18 分，通过率达 96.15%，12 个场景中 9 个排名国内第一。金融方面，混元基于真实金融场景构建数千个工作空间和上万个高质量任务，让模型学习基金研究、证券分析及量化策略等方向的专业流程与规范。在 213 位金融专家参与的众测中，基于 579 个真实金融任务，Hy4 众测得分为 3.04，高于 DeepSeek-V4.1-Flash-1017（2.97）、Kimi K3.1 Pro（2.68）和 GLM-5.3（2.68）。此外，混元针对时效检索、多源信息整合及事实一致性校验进行优化，自建测试中的模型幻觉率由 12.95%降至 7.2%，猜测导致的错误率由 23.1%降至 5.6%。

软件开发与科学研究能力进一步拓展，强化复杂任务的端到端执行。混元针对前端复杂交互、业务逻辑、状态管理及前后端联动进行优化，并实现对 UE、Unity 和 Godot 等主流引擎的深度适配。相比 Hy4 Preview，新模型在 GameDev 和 GameLogic 评测中分别提升 19.1%和 9.8%。科学研究方面，Hy4 结合复杂推理、科学编程及长程任务执行能力，进一步提升端到端推进科研工作的能力。在覆盖七大学科、123 项可执行研究任务的 TenureTrail 内部评测中，Hy4 在数学构造、物理建模、地球环境、系统工程和社会经济决策等领域展现出扎实的科研能力。同时，Hy4 与 Hyra2.0 脚手架优化了内部同架构小模型的预训练基础设施；混元团队还开源了由 Hy4 协同 Hyra2.0 研发的单细胞表征基础模型 HyCell 和物理场基础模型 HyPDE。

模型已接入腾讯产品及第三方开发平台，价格维持不变并计划推出 Flash 版本。Hy4 目前已上线腾讯云 TokenHub、WorkBuddy、CodeBuddy、元宝、企业微信智能助理大圆及 ima 等产品，API 同步接入 OpenRouter、Hermes 和 OpenCode 等平台。相比 Hy4 Preview，新模型保持原有定价，每百万 Token 输入、输出及缓存命中价格分别为 6 元、18 元和 0.3 元。此外，即日起至 2026 年 12 月 31 日，Hy4 在腾讯云 TokenHub 按刊例价五折计费。

腾讯通过模型能力与产品协同优化，推动长程 Agent 向专业工作流拓展。Hy4 在视觉理解、Computer Use、金融分析及软件开发等场景中实现能力提升；混元通过专项训练与产品适配强化其复杂任务执行能力，并在模型能力升级的同时维持原有 API 价格，进一步改善专业任务的能力与成本表现。

1.3、 OpenAI：发布 GPT-Image-2.6，图像生成速度提升 50%，多轮编辑与 GIF 生成能力升级

OpenAI 发布 GPT-Image-2.6，图像生成速度与编辑一致性进一步提升。11 月 4 日，OpenAI 推出图像生成模型 GPT-Image-2.6，重点优化生成速度、视觉质量及多轮编辑中的细节一致性。新模型图像生成速度较此前提升 50%，并改善角色、产品及品牌元素在连续编辑中的一致性。API 提供 GPT-Image-2.6 Flare 和 Sunburst 两个版本，其中 Flare 在图像质量设为低时最快 6 秒生成图像，并改进海报、产品图片及其他专业视觉作品的布局效果；Sunburst 适用于细节和精度要求较高的任务，在图像质量设为中或高时可生成更高质量的图像，并在多轮编辑中更好地保持一致性。

新增 GIF 动画生成及精细编辑功能，扩展多模态内容创作能力。GPT-Image-2.6 支持制作 GIF 动图，可用于品牌吉祥物动画、社交媒体表情动图和梗图，以及视频制作前的动态分镜等场景。图像编辑方面，OpenAI 将 ChatGPT 已有的图像标注功能引入 API，支持用户指定修改位置和内容，并支持导出可用于网站、应用及印刷品的 SVG 文件。同时，API 新增复杂度参数，允许开发者调整模型处理请求时的规划程度：较低复杂度适用于快速编辑，较高复杂度适用于元素较多或限制条件较复杂的场景；图像质量参数则用于控制视觉细节及保真度。

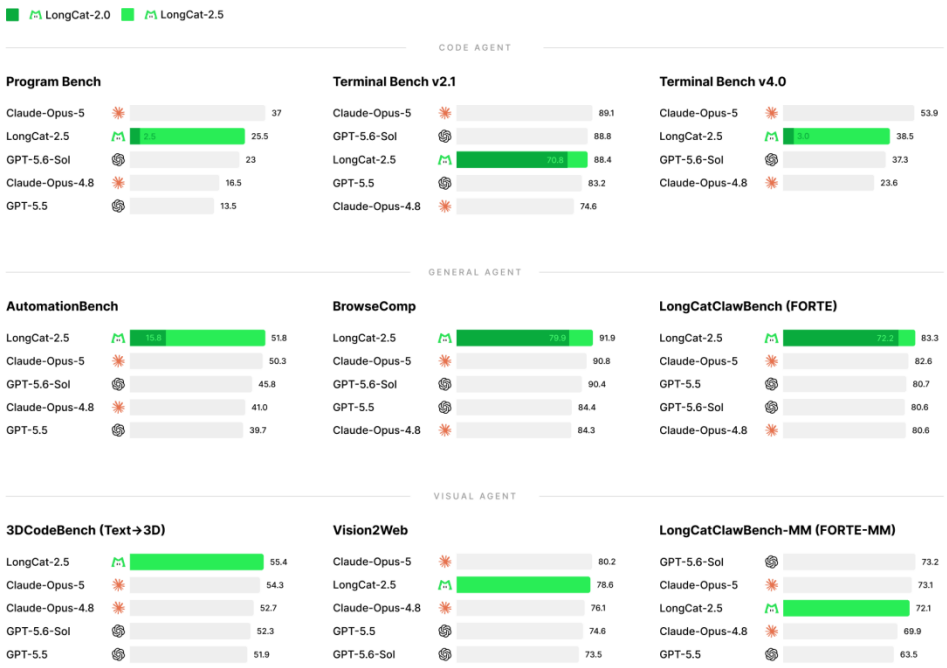
模型已覆盖 ChatGPT、Codex 及 API，图像生成使用规模持续扩大。Images 2.6 已向全球各订阅档位的 ChatGPT 及 Codex 用户开放，覆盖桌面端、移动端及网页端，Flare 和 Sunburst 均可通过 API 调用。OpenAI 披露，目前全球用户通过 ChatGPT、Codex 及 API 每周生成 40 亿张图像。

OpenAI 围绕图像生成效率、内容可控性及创作工作流持续优化。GPT-Image-2.6 通过提升生成速度、强化多轮编辑一致性及增加 GIF、SVG 等输出形式，进一步拓展图像模型在专业设计、营销内容制作及动态视觉创作中的应用范围。

1.4、 美团：发布 LongCat-2.5，原生多模态与长程 Agent 能力升级，国产算力训练及推理效率提升

美团发布 LongCat-2.5，强化原生视觉理解与长程 Agent 编程能力。11 月 2 日，美团正式发布 LongCat-2.5，定位为面向长程智能体任务的原生多模态模型，总参数量为 1.6 万亿，每 Token 激活约 480 亿参数。相比 LongCat-2.0，新模型进一步强化原生视觉理解与智能体编程能力。技术方面，LongCat-2.5 将视觉信号编码为多层级离散 Token，使主干模型能够以统一的 Token 形式处理视觉与文本输入，并通过视觉 Token 离线预处理及视觉嵌入联合学习，既支持原生多模态融合，也简化训练中的视觉计算与负载管理，支持 Visual Coding、创意设计及 3D 等任务。

图2：LongCat-2.5 较 2.0 版本大幅提升，在 AutomationBench、BrowseComp 等评测中得分领先



资料来源：美团技术团队

扩大长程强化学习训练规模，提升复杂 Agent 任务的端到端执行能力。美团围绕数据合成、环境构建、异步强化学习及跨 Harness 混合训练完善 Agentic RL 训练链路，增加长程任务轨迹，并将真实 Agent 失败案例转化为可复现、可验证的训练任务。在训练过程中，采用 Partial Rollout 异步训练策略缓解长尾任务带来的效率损失，同时对重复思考和重复工具调用施加额外惩罚，并通过重新分配正样本中的 Reward 抑制过度思考和长度增长过快。此外，美团沿用 LongCat-2.0 的专家组机制，通过多专家在线教学，将 Agent、推理及交互体验（含指令遵循与安全防御）等领域的专家能力融入主线模型；跨多任务、跨 Scaffold 的混合训练则可显著提高模型泛化能力。

优化国产算力上的长上下文训练系统，提升大规模模型训练效率。针对国产 AI 加速器单卡显存不足 80GB 的限制，美团通过超节点并行策略、双重分片 FSDP、动态专家复制及通信计算重叠等技术降低显存占用、改善负载均衡。在长序列监督微调任务中，优化后的 FSDP 配置使单卡吞吐量较 DDP 基线提升 86%。同时，美团通过训推数值一致性及长上下文 Rollout 优化，支撑大规模长程强化学习的高效运行。

W4A4 量化推理优化计算与通信开销，Prefill 和 Decode 吞吐量分别提升 43.3%和 42.77%。美团针对 1.6 万亿参数 MoE 模型及国产 AI 加速器特点，将专家权重与激活值均量化至 4-bit，并结合原生 INT4 矩阵计算、算子融合及推理流水优化提高推理效率。根据官方线上压测，相比 W8A8 INT8 基线，W4A4 方案的 Prefill 和 Decode 吞吐量分别提升 43.3%和 42.77%，首 Token 延迟（TTFT）及每 Token 输出延迟（TPOT）分别下降 15.91%和 19.72%，GMM 算子平均实现 1.73 倍加速。

美团围绕原生多模态、长程强化学习及国产算力适配，同步推进模型能力与计算效率优化。LongCat-2.5 通过统一视觉与文本表征、扩展长程 Agent 训练链路，提

升复杂编程及多模态任务执行能力，同时通过国产算力训练系统与 W4A4 推理优化改善模型训练和部署效率。

2、 云计算：中期选举重塑数据中心政策预期，算力租赁进入交付确定性重定价

美国中期选举成为海外算力板块重要的政策变量之一。在美国当地时间 11 月 3 日举行的中期选举中，民主党赢得参众两院多数席位，并拿下德克萨斯、俄亥俄、佐治亚等数据中心重镇州长席位；11 月 5 日纳斯达克指数下跌 5.52%、费城半导体指数下跌 12.25%，CoreWeave、Nebius 与 IREN 分别下跌 9.83%、14.94%、16.95%。我们认为，选举不改变 AI 需求的斜率，但改变了算力的交付确定性：此前市场主要按已签约容量给 Neocloud 定价，此后或更多按项目能否按期通电定价；CSP 因选址分散、现金流较为充沛以及已建成算力资源丰富，受影响或相对较小。

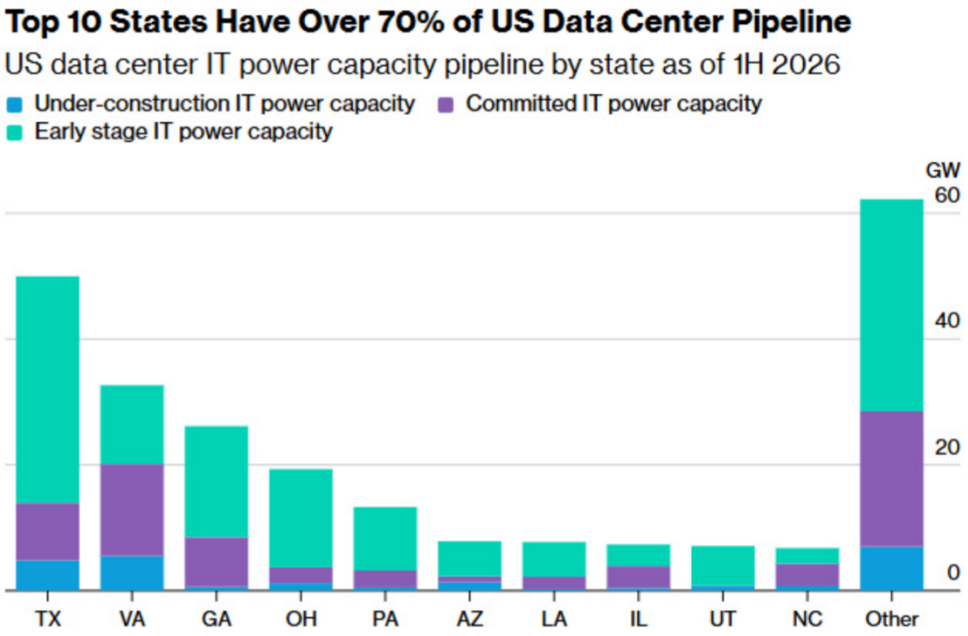
数据中心建设逐渐成为选举议题，核心矛盾在于经济收益与基础设施成本的分配，居民主要担忧电价上涨及水、电等公共资源消耗。NBC News 2026 年 8-9 月民调显示，69%的受访者反对在当地社区新建 AI 数据中心，其中 45%强烈反对。随着议题进一步进入选举视野，NBC News 2026 年 11 月出口民调显示，56%的选民认为 AI 数据中心是其国会选举投票决策中的关键或重要因素。我们认为，居民抵触情绪并非单纯针对 AI 技术，而是反映出数据中心商业收益与电网扩容、能源消耗及水资源使用等成本在企业与社区之间分配不均。未来政策或更侧重明确成本分担机制，推动数据中心承担相应的基础设施建设及资源使用成本，而非全面限制 AI 产业发展。

2.1、 民主党的数据中心立场：侧重成本内部化与地方同意

从已当选州长的竞选纲领看，政策工具箱集中在三类：一是直接暂停审批，德克萨斯州（TX）当选州长 Gina Hinojosa 主张立即暂停新建数据中心，直至州议会通过保护居民的立法；二是有条件暂停，俄亥俄州（OH）当选州长 Amy Acton 要求数据中心承担 100%公用事业成本、使用工会劳工、提高运营透明度；三是评估期暂停，佐治亚州（GA）当选州长 Keisha Lance Bottoms 主张全数据中心建设，以评估对公用事业成本与水资源的影响。

值得关注的是，民主党在德克萨斯、俄亥俄和佐治亚三个重要数据中心市场均赢得州长选举，意味着政策收紧风险集中出现在美国算力基础设施建设的核心区域。根据 Pew Research Center 2026 年 4 月发布的统计，三州分别拥有 466 个、223 个和 235 个数据中心，其中规划建设项目分别为 170 个、57 个和 141 个。德州、佐治亚州规划数据中心数量位居全美前列。三州当选州长均提出不同程度的数据中心建设限制，政策影响可能不止于新增项目审批，还可能通过提高电力接入、环境评估、劳工及社区协商要求，拉长项目建设周期并抬升开发成本。

图3：民主党赢得的州长选举地区多为数据中心重镇



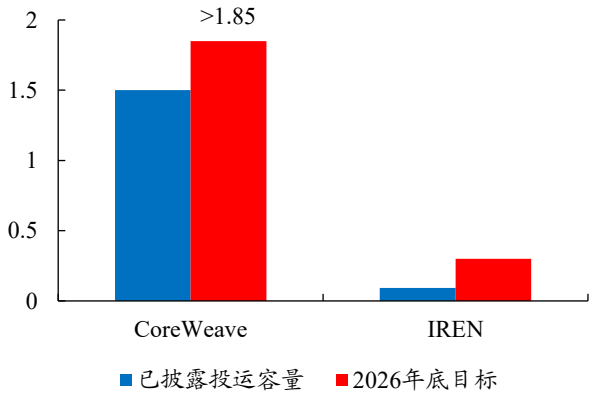
资料来源：BloombergNEF、DC Byte

2.2、对 Neocloud 与 CSP 的影响：交付缺口集中于 Neocloud，现金流约束放大延期影响

CSP 方面，数据中心建设周期拉长可能影响新增算力投放节奏，但头部厂商凭借更大规模的已交付算力储备、更强的资金实力、基础设施建设能力及多区域布局，整体抗风险能力相对更强。以全球基础设施布局为例，AWS、Microsoft Azure 及 Google Cloud 拥有全球分散化布局的数据中心，能够依托跨区域基础设施网络进行资源调度，降低单一区域建设审批及电力供应约束对整体算力部署的影响。同时，在 AI 算力需求持续增长、数据中心新增建设受到土地审批、电网接入及电力供应约束的背景下，头部 CSP 已建成并接入电网的数据中心及算力资源稀缺性有望进一步提升，凸显其先发优势。

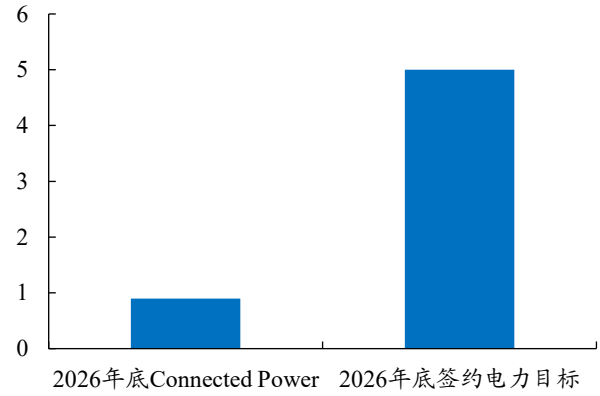
Neocloud 的交付风险主要在于已锁定电力能否按期转化为运营容量。CoreWeave 2026Q2 末签约电力约 3.7GW、已投运约 1.5GW，约 2.2GW 尚未投运，截至 8 月 11 日签约电力增至约 4.2GW；Nebius 预计 2026 年底 Connected Power 为 0.8–1GW，年底签约电力目标上调至 5GW；IREN 的 0.3GW（IT）部署计划对应逾 40 亿美元 ARR 目标，该目标的实现时间受客户要求的设计变更及供应商交付进度影响，由 2026 年底推迟至 2027 年 1 月底，可能延至 2 月底，同时公司计划 2027 年底达到约 0.8GW 数据中心容量。三家公司均有较大规模容量待建设或部署，项目延期可能推迟收入及回款，加大融资压力。

图4：CoreWeave 与 IREN 已投运容量距离近期目标容量仍存在差距（单位：GW）



数据来源：CoreWeave 2026Q2 业绩公告及业绩会纪要、IREN FY2027Q1 业绩材料、开源证券研究所

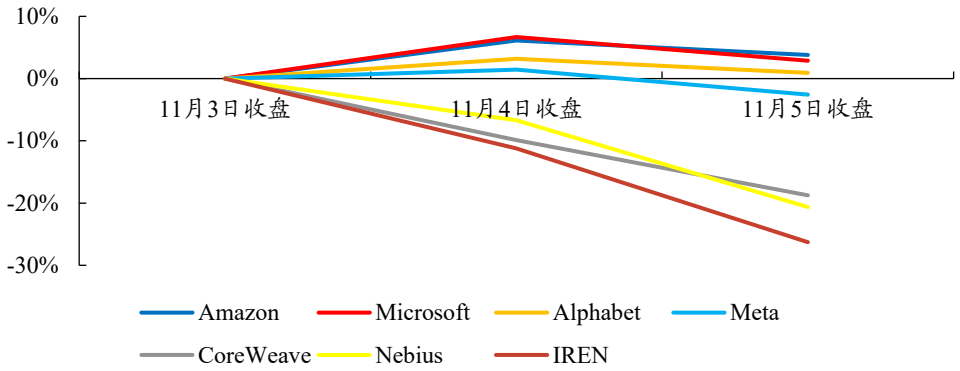
图5：Nebius 2026 年底 Connected Power 与签约电力目标对比（单位：GW）



数据来源：Nebius 2026Q2 股东信及业绩会纪要、开源证券研究所（注：2026 年底 Connected Power 指引为 0.8-1GW，图中取中值 0.9GW）

权益市场出现明显结构性分化，大型 CSP 整体表现相对稳健，Neocloud 及半导体板块明显承压。11 月 4 日选举结果公布后，市场对数据中心建设审批趋严的预期出现重新定价，11 月 4 日至 5 日，Amazon、Microsoft、Alphabet 与 Meta 股价两日累计变动分别为+3.8%/+2.9%/+0.9%/-2.6%，而 CoreWeave、Nebius 与 IREN 两日累计变动分别为-18.8%/-20.6%/-26.3%，大型 CSP 与 Neocloud 之间股价表现明显分化。

图6：CSP 与 Neocloud 股价在中期选举后出现分化

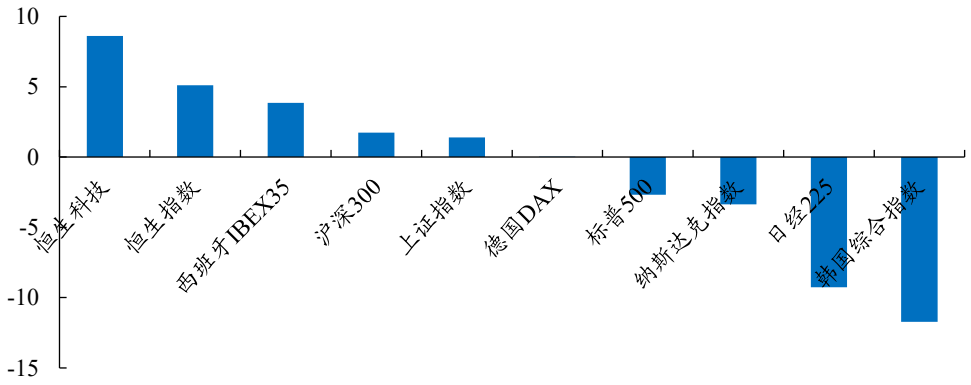


数据来源：Wind、开源证券研究所

3、 港股周度更新

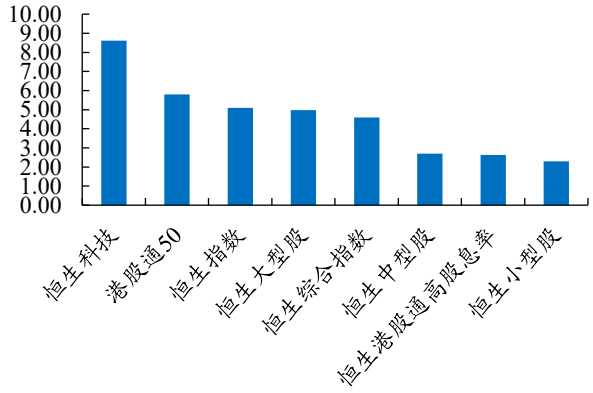
本周恒生指数上涨 5.10%，在全球主要市场中涨跌幅靠前，从成分上看，恒生传媒、恒生科技涨幅居前，恒生科技本周上涨 8.61%。

图7：本周恒生指数（2026.11.02-2026.11.06）在全球主要市场中涨跌幅靠前（单位：%）



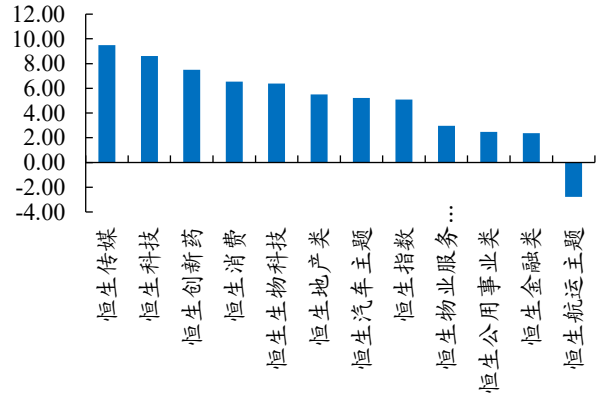
数据来源：Wind、开源证券研究所

图8：本周（2026.11.02-2026.11.06）恒生科技涨幅居首（%）



数据来源：Wind、开源证券研究所

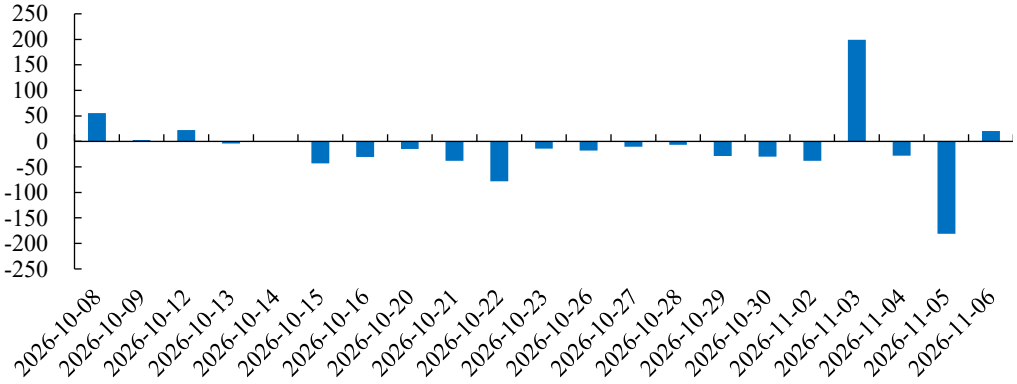
图9：本周（2026.11.02-2026.11.06）恒生传媒、恒生科技涨幅居前（%）



数据来源：Wind、开源证券研究所

本周港股通资金净流出人民币 27.01 亿元，规模较上周收窄；但从头部活跃个股来看，中芯国际（+11.6 亿港元）、华虹宏力（+9.2 亿港元）、中国海洋石油（+7.6 亿港元）资金净流入居前。

图10：2026 年 10 月以来港股通当日买入成交净额有所波动（单位：亿元）



数据来源：Wind、开源证券研究所

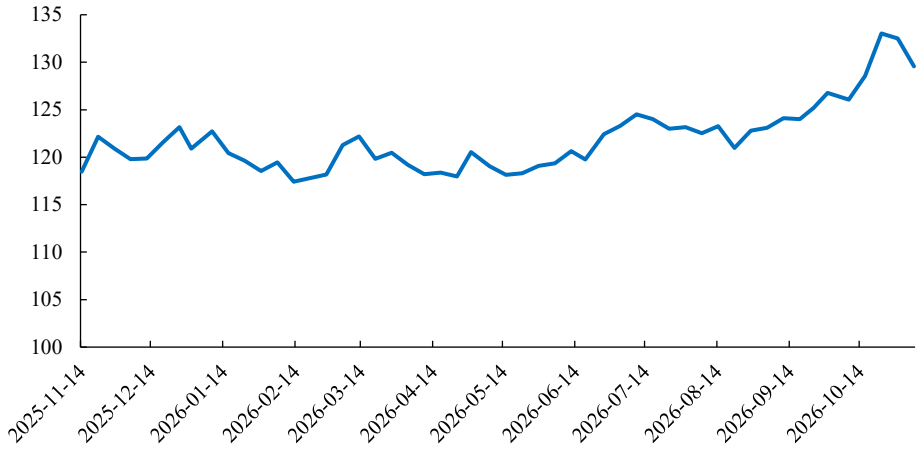
表1：本周（2026.11.02-2026.11.06）中芯国际、华虹宏力与中国海洋石油流入资金居前

股票代码	股票名称	成交净买入（亿港元）	港股通持股数占港股总股数比例
0981.HK	中芯国际	11.6	26.4%
1347.HK	华虹宏力	9.2	14.1%
0883.HK	中国海洋石油	7.6	22.4%
3308.HK	中际旭创	5.7	10.6%
2513.HK	智谱	5.4	16.6%
6166.HK	剑桥科技	-0.1	61.5%
0100.HK	MINIMAX-W	-0.2	12.2%
0939.HK	建设银行	-0.4	15.0%
2331.HK	李宁	-0.5	19.1%
2269.HK	药明生物	-0.7	23.2%
3690.HK	美团-W	-0.9	18.7%
1398.HK	工商银行	-1.4	40.8%
6869.HK	长飞光纤光缆	-1.5	40.5%
1888.HK	建滔积层板	-3.3	9.9%
1548.HK	金斯瑞生物科技	-3.4	19.4%
0005.HK	汇丰控股	-3.9	8.9%
0700.HK	腾讯控股	-16.8	12.0%
1810.HK	小米集团-W	-68.3	19.0%

数据来源：Wind、开源证券研究所

本周恒生沪深港通 AH 股溢价指数为 129.58，环比有所下降，随着后续更多 A 股科技类标的赴港上市，或带动港股板块的“成长性溢价”持续向 A 股靠拢。

图11：恒生沪深港通 AH 股溢价指数自高位回落



数据来源：Wind、开源证券研究所

4、 模型与算力价格数据更新

4.1、 模型信息更新

9 月以来全球大模型行业进入新品密集发布期，海内外头部厂商先后推出约三十款新模型，产品定位与定价梯队分化明显。本周（2026.11.02-2026.11.06）新上线模型主要为美团 LongCat-2.5 和腾讯 Hy4。

表2：9 月以来新上线模型

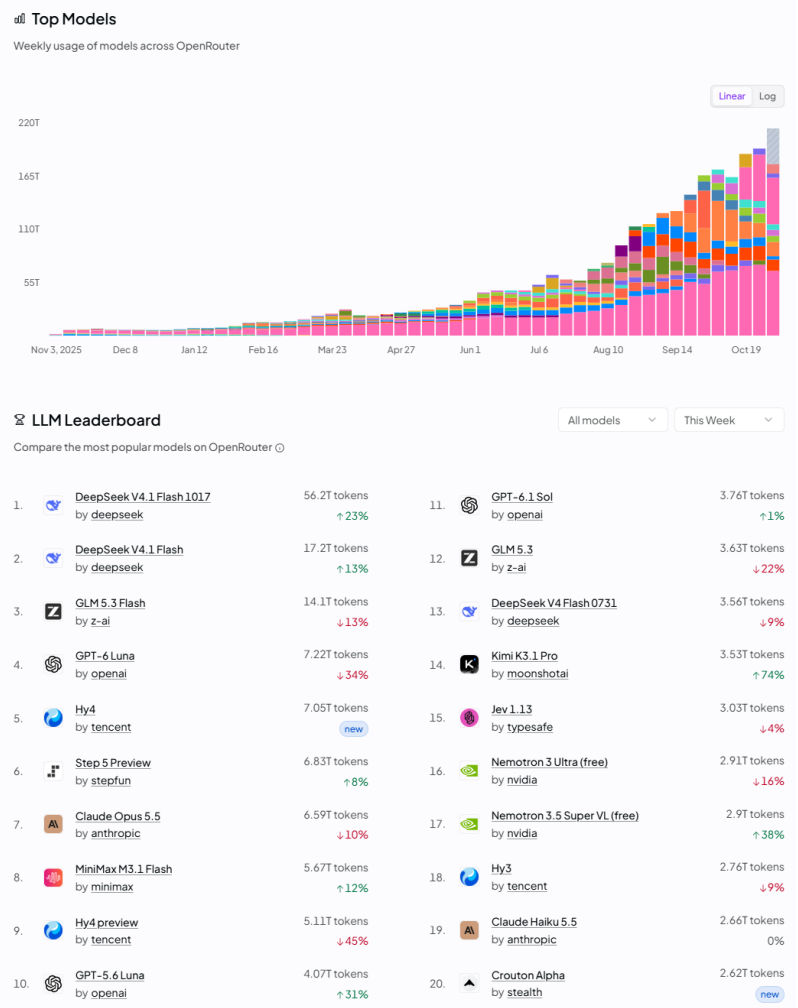
已发布模型	发布时间	厂商	输入价格（/M）	缓存命中输入价格（/M）	输出价格（/M）	上下文窗口
Claude Fable 5.1	2026-09-01	Anthropic	$10.00	$0.25	$50.00	1M
Claude Mythos 5.1	2026-09-01	Anthropic	$10.00	$0.25	$50.00	1M
Gemini 3.8 Flash	2026-09-02	Google	$0.75	$0.075	$3.75	1M
Muse Spark 1.3	2026-09-02	Meta	$1.25	$0.15	$4.25	1M
GPT-6 Astra	2026-09-03	OpenAI	$10.00	$1.00	$50.00	1.05M
Kimi K2.8 Preview	2026-09-11	月之暗面	目前只在 Kimi Code 与 Kimi Work 上线			1M
GLM-5.3-FlashX	2026-09-18	智谱	¥2.00	¥0.57	¥7.00	1M
Step 5 Preview	2026-09-20	阶跃星辰	¥7.00	¥0.35	¥20.00	1M
MiMo-V2.6-Pro	2026-09-22	小米	¥3.00	¥0.025	¥6.00	1M
MiMo-V2.6-Flash	2026-09-22	小米	¥1.00	¥0.02	¥2.00	1M
Grok 4.7	2026-09-22	SpaceXAI	$2.00	$0.50	$6.00	500K
GPT-6 Sol	2026-09-23	OpenAI	$2.00	$0.20	$10.00	1.05M
GPT-6 Luna	2026-09-23	OpenAI	$0.10	$0.01	$0.50	1.05M
Claude Opus 5.5	2026-09-23	Anthropic	$4.00	$0.20	$20.00	1M
Claude Sonnet 5.5	2026-09-28	Anthropic	$2.00	$0.10	$10.00	1M
GPT-6.1 Sol	2026-09-29	OpenAI	$2.00	$0.10	$10.00	1.05M
Gemini 4 Argon	2026-10-01	Google	$2.00	$0.10	$10.00	1M
Beam	2026-10-05	Reflection AI		暂无		1M
Claude Haiku 5.5	2026-10-08	Anthropic	$0.10	$0.01	$0.50	1M

已发布模型	发布时间	厂商	输入价格（/M）	缓存命中输入价格（/M）	输出价格（/M）	上下文窗口
Muse Wonder Full	2026-10-12	Meta	目前只在 Muse Code 订阅中开放			1M
Muse Wonder	2026-10-12	Meta	$4.00	$0.20	$20.00	1M
Muse Spark Open	2026-10-15	Meta	$1.25	$0.15	$4.25	1M
DeepSeek-V4.1-Flash-1017(高峰)	2026-10-17	DeepSeek	$0.30	$0.006	$1.20	1M
DeepSeek-V4.1-Flash-1017(空闲)	2026-10-17	DeepSeek	$0.15	$0.003	$0.60	1M
MiniMax M3.1 Flash	2026-10-19	MiniMax	$0.15	$0.015	$0.60	1M
Kimi K3.1 Pro Highspeed	2026-10-24	月之暗面	$4.00	$0.20	$20.00	1M
Kimi K3.1 Pro	2026-10-24	月之暗面	$2.00	$0.20	$10.00	1M
GPT-6.1 Terra	2026-10-27	OpenAI	$1.00	$0.05	$5.00	1.05M
Gemini 4 Carbon	2026-10-28	Google	$2.00	$0.10	$10.00	1M
LongCat-2.5	2026-11-02	美团	¥2.00	¥0.04	¥8.00	1M
Hy4	2026-11-04	腾讯	¥6.00	¥0.30	¥18.00	1M

资料来源：各模型厂商官网、IT 之家、开源证券研究所

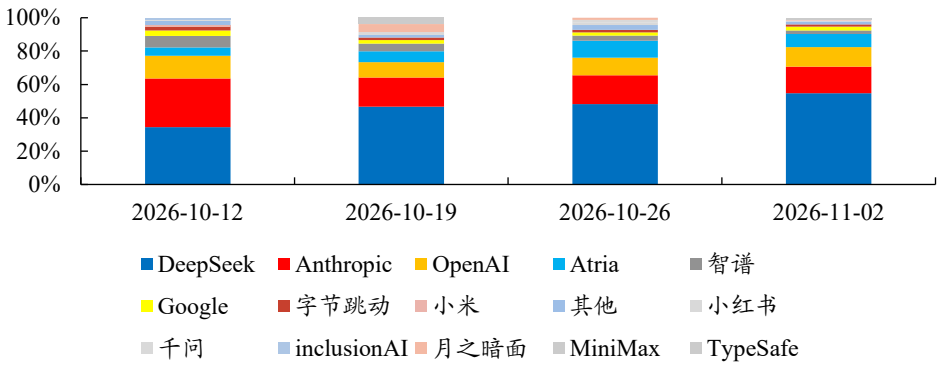
本周（2026.11.02-2026.11.06）OpenRouter 的 token 调用量持续提升，其中国产模型 DeepSeek V4.1 Flash 1017、GLM 5.3 Flash 与 Hy4 位居前列。截至 11 月 2 日，在 ZenMux 统计的 Token 市场份额中，国产 AI 占比约 67%，DeepSeek 市场份额保持第一，占比为 54.76%。依托开源生态及较低的 API 定价，国产模型市场份额维持较高水平。

图12：OpenRouter 的 token 调用量持续提升



资料来源：OpenRouter 官网

图13：国产 AI 在 ZenMux 的 Token 市场份额约三分之二



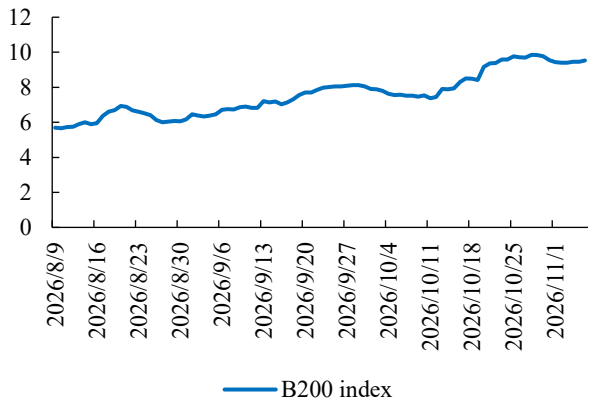
数据来源：ZenMux、开源证券研究所

4.2、 算力租赁价格更新

GPU 租赁价格整体高位运行，不同性能 GPU 价格走势有所分化。根据 Ornn 统计的基于 neocloud 市场的租赁交易价格，高性能型号 B200 与 H200 租赁价格处于高位但近期有所回落，H100 租赁价格震荡上行，A100 租赁价格基本持平。

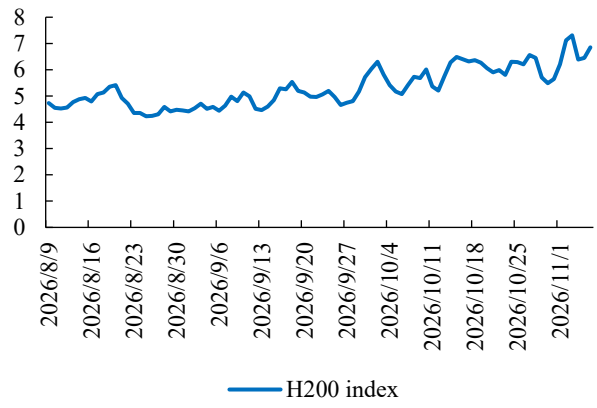
截至 11 月 6 日，B200 租赁价格为 9.53$/小时；H200 租赁价格为 6.85$/小时；H100 租赁价格为 3.28$/小时；A100 租赁价格为 1.05$/小时。

图14：B200 租赁价格高位运行（单位：$/小时）



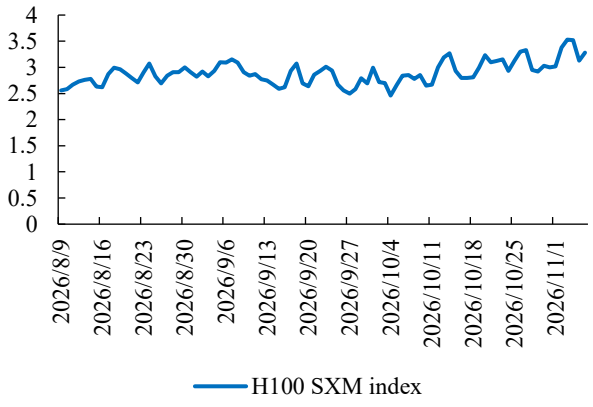
数据来源：Ornn、开源证券研究所

图15：H200 租赁价格高位运行（单位：$/小时）



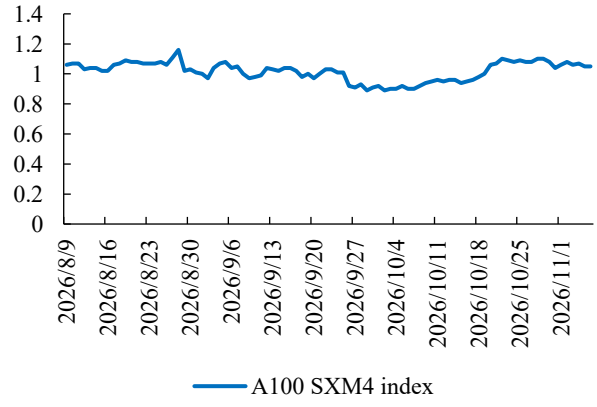
数据来源：Ornn、开源证券研究所

图16：H100 租赁价格震荡上行（单位：$/小时）



数据来源：Ornn、开源证券研究所

图17：A100 租赁价格基本持平（单位：$/小时）



数据来源：Ornn、开源证券研究所

5、 投资建议

大模型：Anthropic 收入保持较快增长，但短期高价算力采购推高成本，调整后营业利润率环比下降，叠加大规模长期基础设施承诺，头部模型厂商对稳定算力供给的需求持续提升；腾讯 Hy4、美团 LongCat-2.5 及 OpenAI GPT-Image-2.6 持续强化视觉理解、长程 Agent 及内容生成能力，推动模型在专业办公、软件开发及内容创

作等场景落地，有望带动 Token 调用和推理算力需求增长。建议关注具备模型迭代、产品落地和客户拓展能力的厂商，推荐标的：智谱、德适-B，受益标的：MiniMax。

云计算：中期选举后数据中心审批趋严及成本分担要求提高的预期升温，算力交付确定性的重要性进一步提升。已建成并接入电网的算力资源稀缺性有望提升，具备存量数据中心、资金实力及跨区域布局的头部厂商相对受益；同时，高额资本开支的回报仍需跟踪，收入兑现与利润率压力可能带来短期估值波动。受益标的：谷歌、亚马逊、微软。

6、 风险提示

模型及产品推出不及预期：大模型行业更新较快，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。

AI 算力需求不及预期风险：若下游模型迭代放缓、客户融资或算力利用率低于预期，可能导致新增订单放缓、RPO 转化周期拉长或数据中心产能交付延后。

算力供给超预期释放风险：本轮按需算力价格上涨建立在供需缺口之上，若 GPU 与内存供给快速释放，或新云厂商产能集中投放，按需与长约价格可能回落，算力租赁的盈利弹性随之收窄。

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。

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）》已于2017年7月1日起正式实施。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。

分析师承诺

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、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。		

分析、估值方法的局限性说明

本报告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。

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