计算机

2026 年 11 月 08 日

北美科技巨头持续加码算力，企业 Agent 加快落地

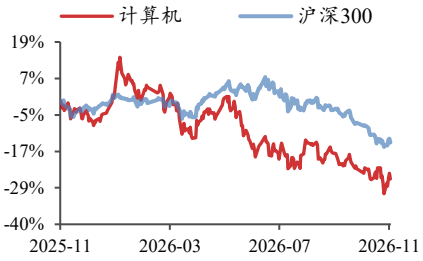
——行业周报

张越（分析师）
zhangyue1@kysec.cn
证书编号：S0790524090003

叶进新（联系人）
yejinxin@kysec.cn
证书编号：S0790126060053

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

《Agent 推理驱动服务器升级，AI 产业迈向全栈协同—行业周报》-2026.11.1

《Anthropic 递交 S-1 招股书启动 IPO，模型与产品双轮驱动兑现盈利拐点—行业点评报告》-2026.10.28

《Kimi K3.1 Pro 智能指数全球第四，闭源差距再缩小—行业点评报告》-2026.10.25

● 北美四巨头资本投入持续扩张，长期订单支撑 AI 基础设施建设

北美四大科技巨头延续 AI 基础设施投入，第三季度现金购置物业与设备支出环比均有增长，服务器、网络设备与数据中心成为重点方向。需求端，谷歌、亚马逊与微软剩余履约义务持续扩大，企业通过多年期合同提前锁定算力，为后续产能建设提供需求支撑；Meta 则通过模型迭代、内容推荐与广告提效推动投入回报。AI 云商业兑现可沿资本投入、订单锁定与收入确认逐步跟踪，客户结构、合同期限及交付进度成为重要观察指标。随着投入规模提升，经营现金流、外部融资与融资租赁共同参与扩产，设备交付、电力及基建配套对实际投放节奏的影响进一步增强。产业关注点正从投资规模延伸至供给兑现与运营效率，具备稳定交付能力和系统配套优势的企业，有望持续受益于算力建设与长期需求释放。

● 企业 Agent 走向托管与跨工具协作，模型能力及国产生态同步完善

本周 AI 产业围绕企业部署、多模态交互与任务执行持续迭代。模型端，腾讯发布混元 Hy4 正式版，新增视觉与电脑操作能力；百度开放 ERNIE 5.5 Flash Base 权重，支持长上下文推理；OpenAI 上线 GPT-Image-2.6，拓展动图与矢量图生成。应用端，DeepSeek Harness 推出创造模式，支持智能体按需生成并安装插件；Kimi Hosted Agents 提供企业智能体托管环境，结合工具接入、多智能体协作与权限管理，降低长程任务部署门槛。国产生态方面，LongCat-2.5 推进国产算力训练与低精度推理优化，小米开源轻量文档解析模型，腾讯 Hyra 2.0 探索科研及处理器优化。模型、工具与运行环境加快融合，有望推动 Agent 从单点功能向持续业务服务演进，拓展 AI 应用与配套算力需求。

● 计算机板块涨幅居前，AI 算力与应用共同呈现活跃行情

11 月 2 日至 11 月 6 日，计算机指数上涨 9.60%，涨幅超过沪深 300、上证指数、创业板指及科创 50，在 31 个一级行业中排名第 3 位。AI 细分方向同步上涨，AI 应用指数累计上涨约 7.71%，AI 算力指数累计上涨约 4.73%，应用方向相对领先约 2.98 个百分点。个股层面，青云科技-U、优刻得-W、开普云分别上涨 28.69%、20.58%和 19.47%，均录得两位数涨幅；AI 应用方向涨幅前十个股周涨幅均超过 18%，易点天下上涨 29.51%。结合北美科技巨头资本投入延续、企业 Agent 托管服务落地与国产模型生态完善，产业催化持续丰富，可关注算力建设向订单及交付转化，以及 AI 能力融入企业工作流带来的细分领域机会。

● 受益标的

国产算力：寒武纪、海光信息、禾盛新材、沐曦股份、浪潮信息、工业富联等；**海外算力：**东山精密、胜宏科技、联特科技、德邦科技等；**算力租赁：**宏景科技、利通电子、协创数据等；**大模型：**智谱、MiniMax、阿里巴巴、腾讯控股；**液冷：**盾安环境、同飞股份、恒铭达等。

● 风险提示：大模型技术迭代不及预期风险、AI 技术应用不及预期风险、相关政策不及预期风险、AI Infra 建设不及预期风险等。

目　录

1、 北美科技四强资本投入加码，供给融资双约束定扩产节奏 3
1.1、 四大巨头资本开支高增延续，三大指标跟踪 AI 云商业兑现 3
1.2、 资本开支的利润与现金流代价显现，外部融资或成增量资金来源 8
2、 计算机产业动态更新 10
2.1、 国内 CSP 厂商与大模型厂商最新动向 10
2.2、 海外 CSP 厂商与大模型厂商最新动向 11
2.3、 国产算力产业最新进展 11
2.4、 海外算力产业最新进展 12
3、 上周市场表现回顾（2026/11/2-2026/11/6） 13
4、 风险提示 14

图表目录

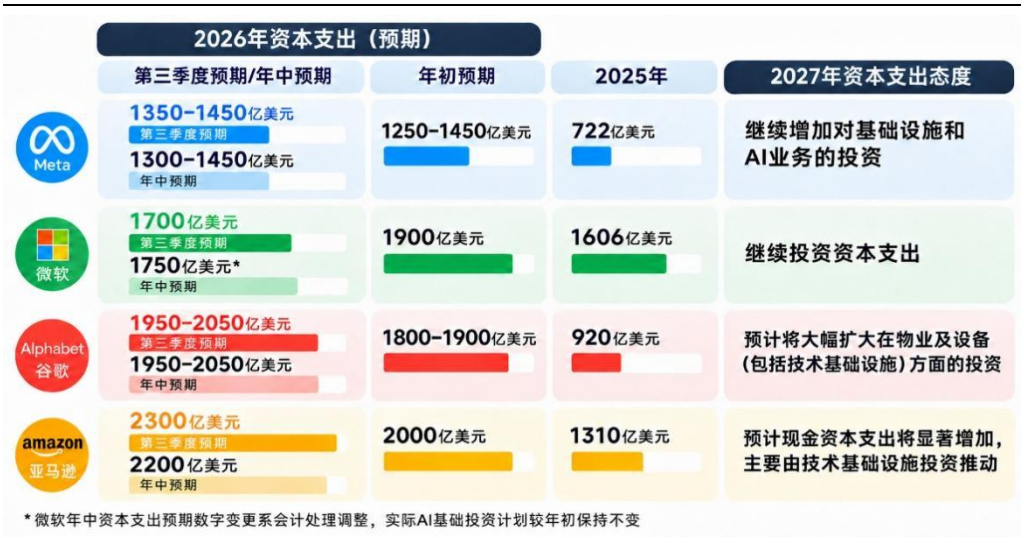
图 1： 北美四大科技巨头资本投入节奏出现分化 3
图 2： AI 云的商业兑现需要经历资本投入、需求锁定、确认收入的循环 4
图 3： 北美四大科技巨头 2026 年 Q3 资本开支环比全部上升 4
图 4： 北美四大科技巨头资本开支单季投入增量创出新高，但投入出现增速分化 5
图 5： 亚马逊、微软、谷歌三大云厂商剩余履约义务规模快速增长 5
图 6： 谷歌剩余履约义务高速扩张，可确认收入占比维持较高水平，业务结构健康 6
图 7： 亚马逊剩余履约义务高速扩张，合同履约周期拉长，订单远期占比抬升 6
图 8： 2025 年末微软短期可确认收入占比显著下行，由大额长期算力订单拉长所至 7
图 9： 北美四大科技公司营收与利润仍维持双位数增长，但盈利结构开始分化 7
图 10： 2025 年 Q4 至 2026 年 Q3 四家合计口径下的资本开支压力呈现明显的分档 8
图 11： AI 算力投入首次烧穿四家巨头的自由现金流防线，单季合计自由现金流首次转负 9
图 12： 本周多数一级行业上涨，计算机板块涨幅居前 13
图 13： 本周 AI 细分方向均有所上涨，AI 应用指数整体表现优于 AI 算力指数 13
图 14： AI 算力指数整体上涨，青云科技-U 等个股表现突出 14
图 15： AI 应用方向指数整体上涨，易点天下涨幅居前 14

1、 北美科技四强资本投入加码，供给融资双约束定扩产节奏

1.1、 四大巨头资本开支高增延续，三大指标跟踪 AI 云商业兑现

海外四大云厂商对 2027 年资本开支保持一致的扩张口径，AI 基础设施投资的规模化承诺成为财报季最强音。Meta、微软、谷歌及亚马逊四家公司管理层均未给出 2027 年资本开支具体指引，亦无一家表态放缓，但在 2026 年资本支出预期上产生一定分化：Meta 上调 2026 年资本开支区间下限至 1350 亿美元，以支持 AI 发展与核心业务；微软小幅下调 2026 年资本开支预期，但在财报中明确表示将继续投资资本支出，以支撑云产品组合与 AI 训练基础设施；谷歌预计 2026 年大幅增加服务器、网络设备及数据中心投资，后续将扩大物业及设备投入；亚马逊预计 2026 年现金资本支出约 2200 亿美元，2027 年将显著增加，由技术基础设施投资推动。

图1：北美四大科技巨头资本投入节奏出现分化



资料来源：财联社公众号、各公司公告、开源证券研究所

资本开支直接决定云厂商的远期算力供给能力，Meta 的投入则同时锚定模型训练上限与用户承载规模。资本开支指企业购置服务器、算力芯片、网络设备及建设数据中心等长期资产的投入，相关支出在资产使用年限内通过折旧分期计入成本，是支撑营收增长的核心前置指标。业务结构上，谷歌、亚马逊、微软均已布局成熟的对外 AI 云业务，AI 云相较传统云计算投入周期更长、资金规模更大、风险也更集中；Meta 当前算力基建全部服务内部需求，尚未推出商业化公有云服务，仅处于算力外售方案的规划阶段。

AI 云的商业化链路可通过资本开支、剩余履约义务（RPO，企业已签约但尚未确认收入的合同金额）、收入与利润三大指标跟踪，各家的兑现质量各有差异。云厂商先以大额资本开支采购算力芯片、建设或租赁数据中心形成稳定算力供给；再与企业客户签订 3 至 5 年长期合同，将需求锁定为剩余履约义务；最终按履约节奏分期确认收入与利润，完成从资本投入到现金流回流的闭环。四家质量排序清晰：谷歌业务结构最健康，亚马逊经营韧性最强，微软 AI 云对 OpenAI 依赖度最高；Meta 无对外公有 AI 云业务，算力投入主要用于广告提效、自研大模型迭代及消费端 AI 生态建设，回报依托广告业务兑现，算力外售仅为远期备选方向。

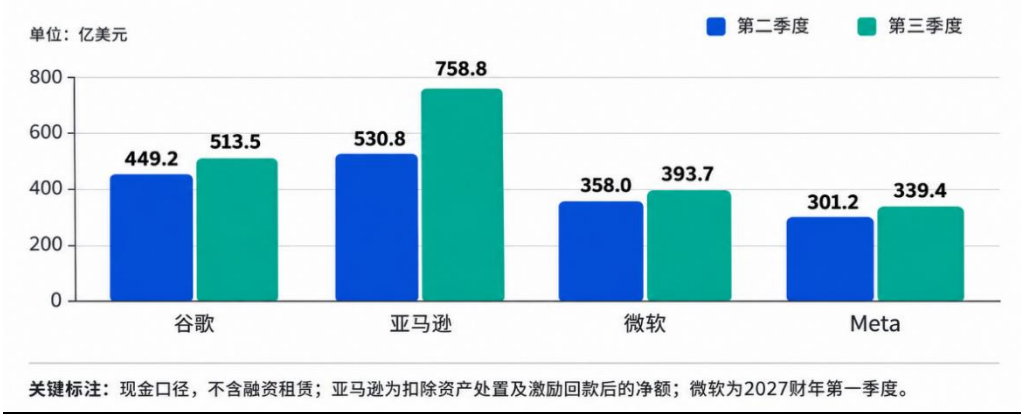
图2：AI 云的商业兑现需要经历资本投入、需求锁定、确认收入的循环



资料来源：财经杂志公众号

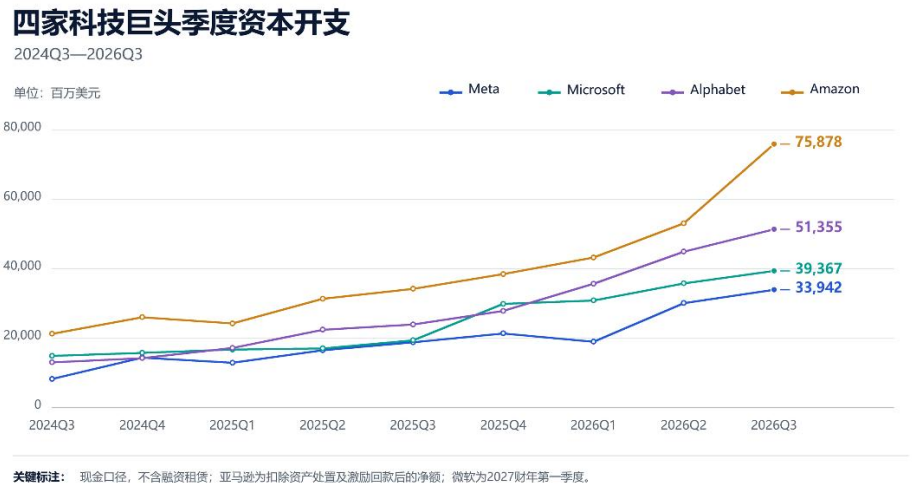
四大云厂商资本开支环比仍在加速，单季投入增量创出新高，指引口径的规模量级同步上移。按现金购置物业与设备计算，第三季度谷歌从第二季度的 449.24 亿美元增至 513.55 亿美元，Meta 从 301.16 亿美元增至 339.42 亿美元，微软从 358.02 亿美元增至 393.67 亿美元，亚马逊从 530.76 亿美元增至 758.78 亿美元；四家单季合计从约 1639 亿美元增至约 2005 亿美元，单季环比增加约 366 亿美元。年度指引层面，四家 2026 年合计约 7200 亿至 7400 亿美元，该合计数仅可观察量级：四家统计口径并不一致，亚马逊为现金资本开支口径，微软与 Meta 包含融资租赁，微软的 1700 亿美元按日历年计算而其财年自 7 月开始。

图3：北美四大科技巨头 2026 年 Q3 资本开支环比全部上升



数据来源：各公司公告、开源证券研究所

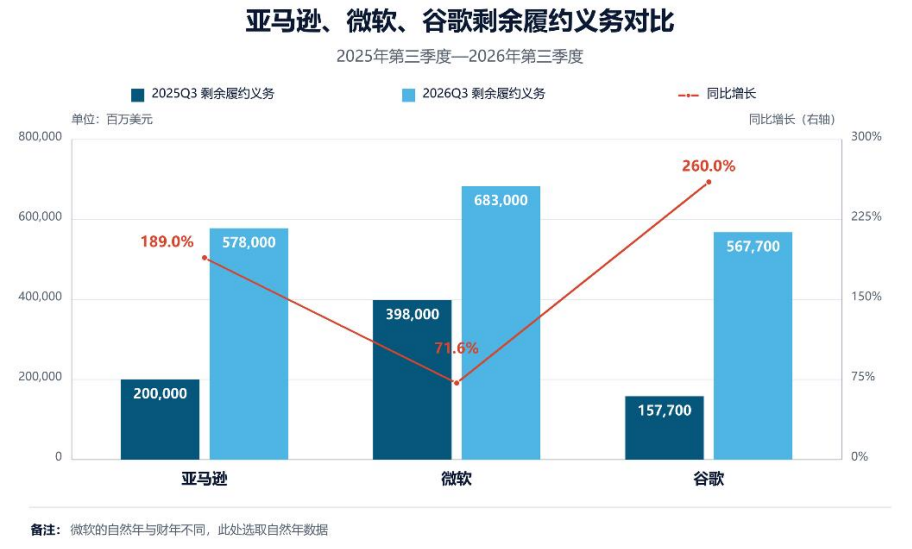
图4：北美四大科技巨头资本开支单季投入增量创出新高，但投入出现增速分化



数据来源：各公司公告、开源证券研究所

AI 算力需求推动企业以多年期合同提前锁定算力资源，三大云厂商剩余履约义务规模快速增长，长期订单的收入质量成为跟踪重点。截至 2026 年第三季度，亚马逊剩余履约义务由上年同期的约 2000 亿美元增至 5780 亿美元、同比增长 189%；谷歌由 1577 亿美元增至 5677 亿美元、同比增长 260%；微软由 3980 亿美元增至 6830 亿美元、同比增长 71.6%（按自然年口径）。不同厂商披露的剩余履约义务收入质量不同，衡量长期订单可靠程度有两个关键指标：一是合同平均履约年限，二是未来 12 个或 24 个月可确认收入的占比，两项指标均按季度披露。

图5：亚马逊、微软、谷歌三大云厂商剩余履约义务规模快速增长

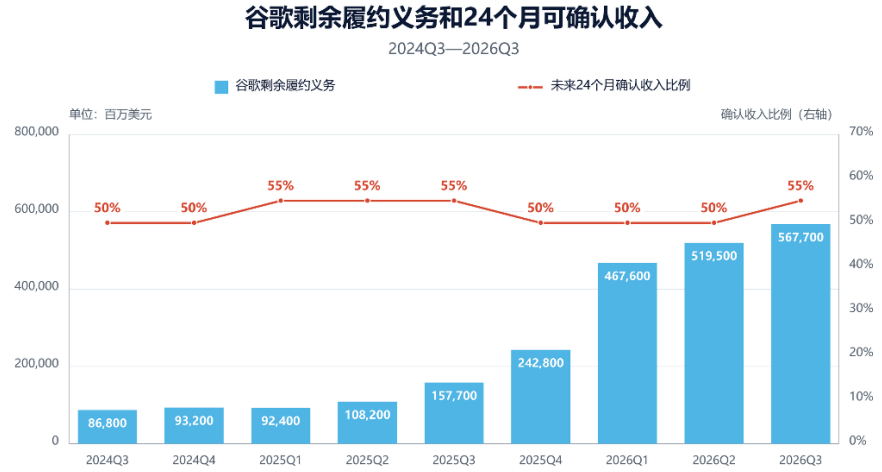


数据来源：各公司公告、开源证券研究所

谷歌剩余履约义务以最高增速扩张，但客户结构分散使其成为三家中订单质量最健康的一家。2026 年第三季度，谷歌剩余履约义务达 5677 亿美元、同比增长超过 260%，主要由 GCP 的基础设施资源、Gemini 大模型与企业级 AI 产品拉动。订单质量层面，谷歌未来 24 个月可确认收入的占比长期维持在 50%至 55%，九个季度内始终稳定，表明新增订单的履约周期未明显拉长。背后逻辑在于客户结构：谷歌 GCP 的 AI 客户相比亚马逊 AWS 与微软 Azure 更为分散，既包括 Anthropic、Midjourney

等明星 AI 创业公司，也包含大量中小创业公司，其使用习惯更轻更快，倾向以 API 调用模型获取推理服务，订单规模适中但兑现期限更短。

图6：谷歌剩余履约义务高速扩张，可确认收入占比维持较高水平，业务结构健康



数据来源：谷歌公告、开源证券研究所

亚马逊的长期订单兑现风险最小，稳定且多元的大客户结构是其剩余履约义务的核心支撑。亚马逊剩余履约义务绝大部分由 AWS 贡献，其全球最大客户涵盖 Salesforce、苹果、字节跳动、Netflix 等大型科技公司，以及 Anthropic、OpenAI 等 AI 创业公司，头部客户每年采购数十亿乃至上百亿美元规模云资源，订单结构相对稳定、兑现概率更高。规模层面，2026 年第三季度亚马逊剩余履约义务达 5780 亿美元、同比增长 189%。合同平均履约年限自 2026 年起显著拉长，由 2025 年第三季度的 3.8 年升至 2026 年第三季度的 6.1 年，长周期合同在锁定未来收入的同时，亦对履约管理提出更高要求。

图7：亚马逊剩余履约义务高速扩张，合同履约周期拉长，订单远期占比抬升

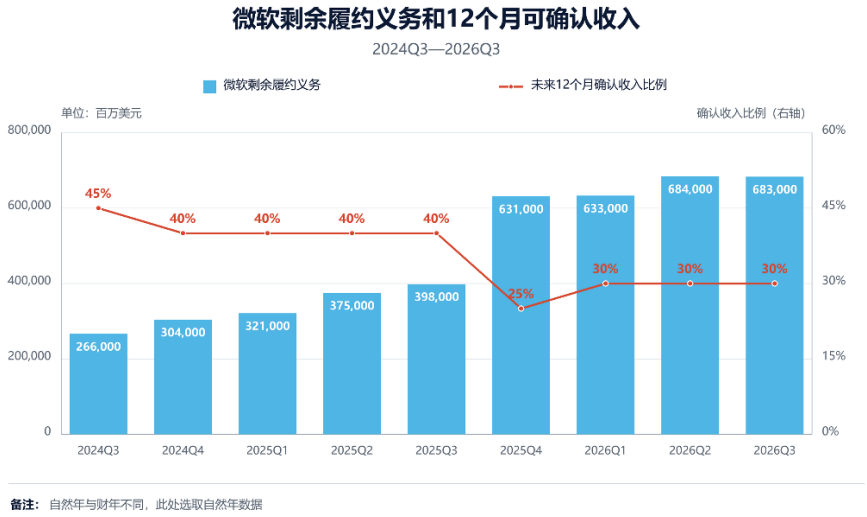


数据来源：亚马逊公告、开源证券研究所

微软剩余履约义务的膨胀由单一大单驱动，对 OpenAI 的深度绑定使其兑现风险在四家中最为突出。2025 年末微软剩余履约义务达 6310 亿美元，不到一年增长超 3000 亿美元；未来 12 个月可确认收入占比由 45%骤降至 25%，源于 2025 年 10 月 OpenAI 与微软签署的 2500 亿美元长期算力大单，兑现周期因此拉长，管理层披露平均履约年限由 2 年增至 2.5 年，企业客户 45%的剩余履约义务与 OpenAI 相关。口

径辨析上，微软剩余履约义务虽为四家最高，但相当部分来自 Office、Windows 等 SaaS 订阅授权，履约周期仅两三年、现金回收更易；Azure 云合同周期更长，2025 年 9 月后新增部分主要与 Azure 相关。

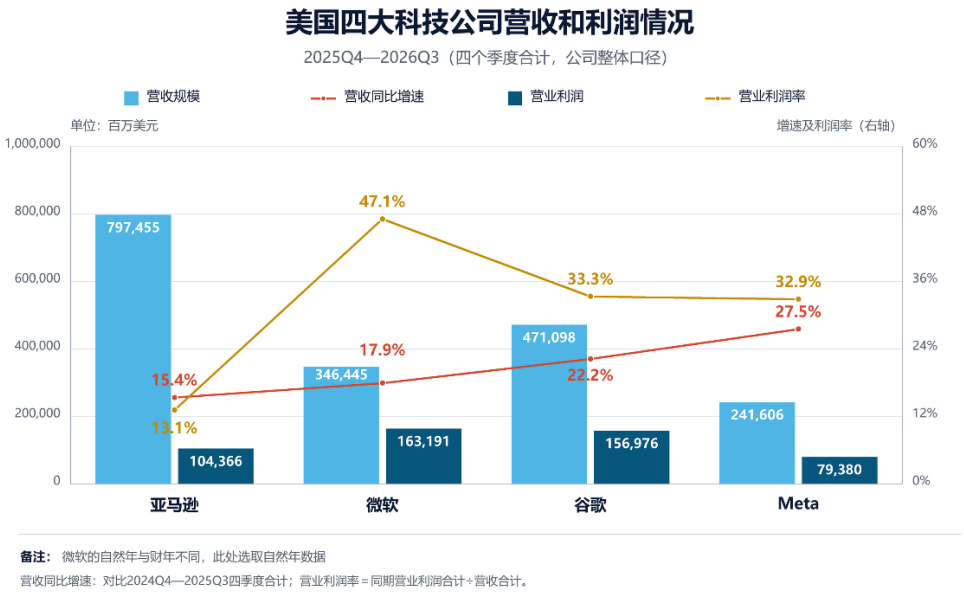
图8：2025 年末微软短期可确认收入占比显著下行，由大额长期算力订单拉长所至



数据来源：微软公告、开源证券研究所

高强度资本开支尚未侵蚀四家的盈利根基，营收与利润仍维持双位数增长，但盈利结构开始分化。按 2025 年第四季度至 2026 年第三季度四季度合计（公司整体口径、微软取自然年），亚马逊营收 7975 亿美元、同比增长 15.4%，营业利润 1044 亿美元、利润率 13.1%；微软营收 3464 亿美元、同比增长 17.9%，营业利润 1632 亿美元、利润率 47.1%居四家之首；谷歌营收 4711 亿美元、同比增长 22.2%，营业利润 1570 亿美元、利润率 33.3%；Meta 营收 2416 亿美元、同比增长 27.5%为四家最高，营业利润 794 亿美元、利润率 32.9%。当前利润表仍具消化能力，但设备占比上行推动折旧加速，承压差异已开始在现金流层面显现。

图9：北美四大科技公司营收与利润仍维持双位数增长，但盈利结构开始分化

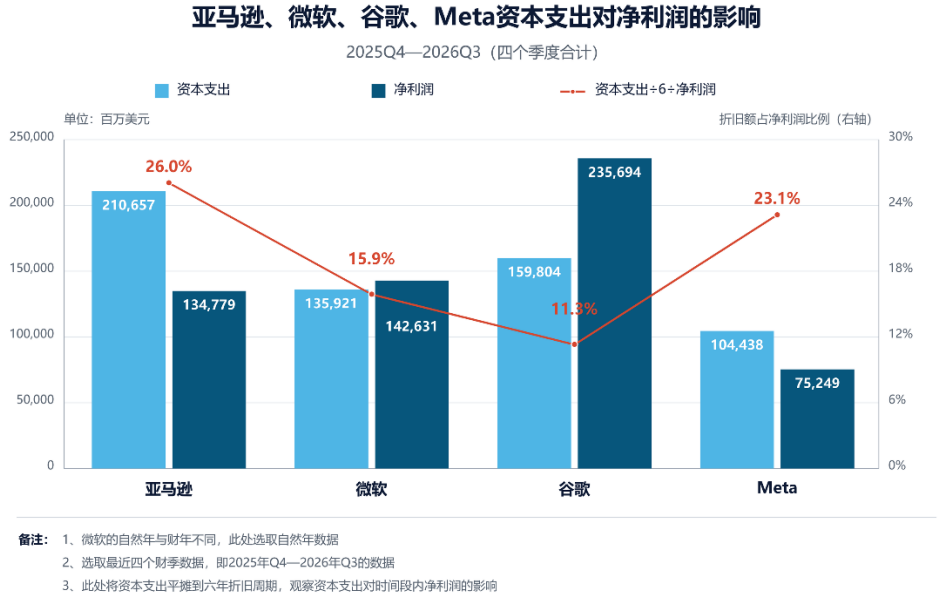


数据来源：各公司公告、开源证券研究所

1.2、 资本开支的利润与现金流代价显现，外部融资或成增量资金来源

资本开支的构成决定折旧对利润表的挤压节奏，设备占比越高，同等投入年均摊入成本的折旧越多。服务器、GPU 与网络设备的折旧年限通常为 5 至 6 年，数据中心建筑则长达十几年至数十年，两类资产的期限结构差异直接映射到未来几年的利润弹性。以六年折旧周期粗略估算（资本支出除以六年再除以净利润），2025 年第四季度至 2026 年第三季度四家合计口径下的压力梯度清晰：亚马逊资本支出 2106.57 亿美元，年均折旧约 351 亿美元，相当于当年净利润的 26.0%；Meta 为 23.1%；微软 15.9%，谷歌 11.3%，四家呈现明显的分档。

图10：2025 年 Q4 至 2026 年 Q3 四家合计口径下的资本开支压力呈现明显的分档



数据来源：各公司公告、开源证券研究所

折旧压力分档由各家的业务属性与现金流结构决定，亚马逊与 Meta 构成承压组，微软与谷歌构成缓冲组。亚马逊压力居四家之首，电商业务长期需要物流与仓储投入、本身即重资产模式，云业务无法获得主业反哺；Meta 压力次之，广告业务能提供支撑，AI 可经内容推荐与广告转化率改善反哺主业，但其收入高度依赖广告，一旦广告增长放缓，新增折旧与运营成本将更直接挤压利润。微软与谷歌缓冲垫更厚：Office、Windows 及 SaaS 授权带来充沛现金流，谷歌广告与搜索业务利润率高，可为云业务持续输血。**折旧期内的利润承压能力，构成四家资本开支可持续性的第一道约束。**

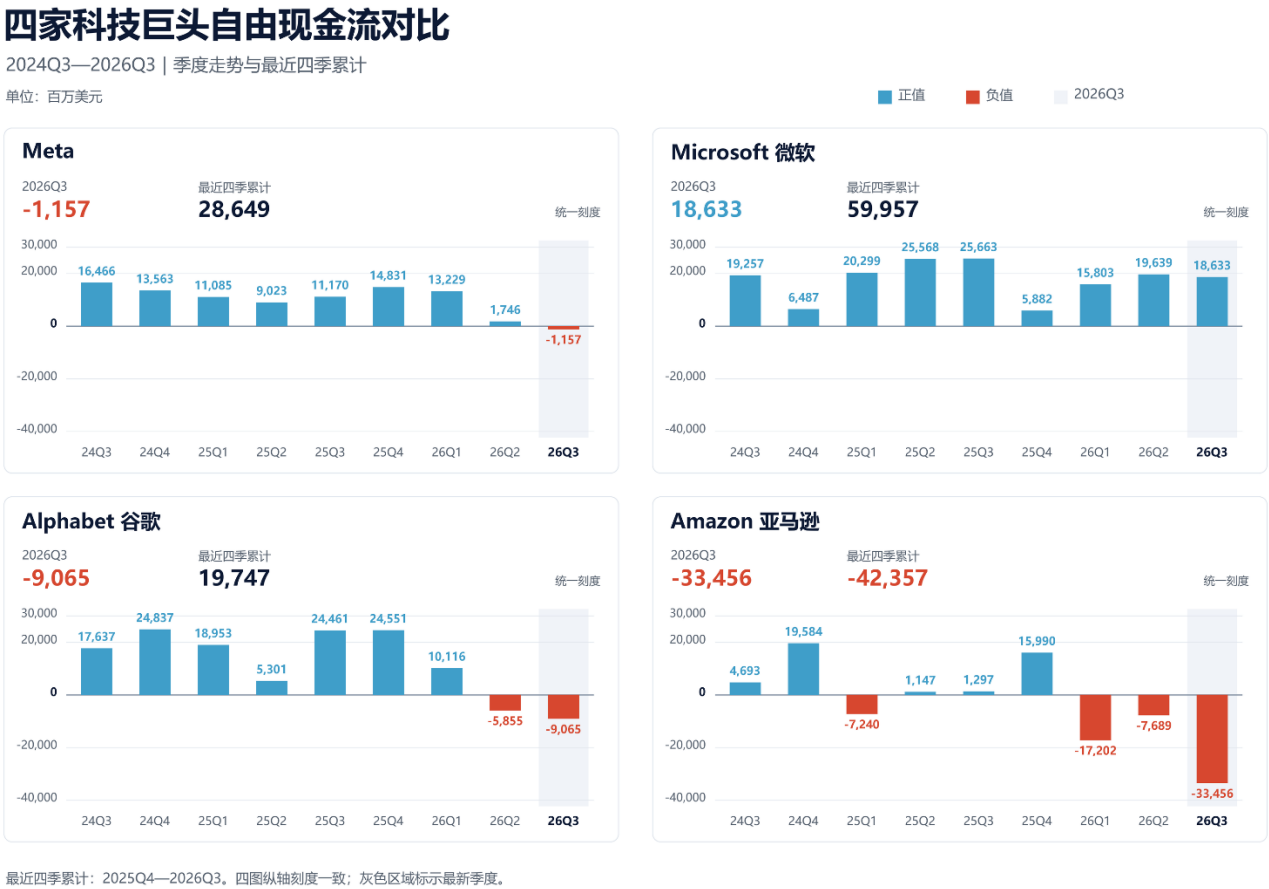
四家资本开支高度集中于短寿命设备资产，折旧加速已在利润表兑现。谷歌第三季度技术基础设施投资约 65%用于服务器与网络设备（年限约 6 年）、35%用于数据中心（7 至 40 年）；微软约四分之三投向以 CPU 和 GPU 为主的短寿命资产，且已包含更高的内存与存储成本；Meta 集中于服务器、数据中心与网络基础设施；亚马逊 759 亿美元现金资本开支主要投向 AWS 和 AI。折旧随之加速：谷歌第三季度折旧 83.69 亿美元、同比增长约 49%，Meta 折旧摊销 71.54 亿美元、上年同期 49.63 亿美元。谷歌首席财务官提示，技术基础设施投资大幅增加，将继续以折旧和能源等运营成本压在利润表上。

2026 年第三季度各公司普遍出现共同现象，部分原定资本开支项目未能按期投放，当前瓶颈已不在需求而在产能交付。微软本季资本开支 452 亿美元，管理层解

释部分原计划于第一财季交付的第三方产能推迟、相关支出移至下半财年；谷歌因供给紧张，第三季度末开始使用第三方算力，第四季度将加大使用并预计近期云业务利润率因此承压；AWS 新增 AI 算力投放后快速被客户消化，算力利用率维持高位。2027 年资本开支落地或将受到产能条件约束，实际投入规模取决于电力、基建与芯片供给等综合因素，并非仅由预算决定。

AI 算力投入首次烧穿四家巨头的自由现金流防线，单季合计转负成为本轮资本开支周期最具信号意义的边际变化。第三季度四家经营现金流合计约 1755 亿美元，自由现金流口径的资本开支合计约 2005 亿美元，自由现金流合计约为负 250 亿美元；二季报时该合计数仍为正，当时仅亚马逊与谷歌转负。各家位置分化明显：微软是唯一保持明显正自由现金流的公司，单季 186 亿美元、但同比下降约 27%；谷歌连续第二个季度为负，由负 58.55 亿美元扩大至负 90.65 亿美元；亚马逊过去 12 个月自由现金流为负 423.57 亿美元、缺口四家最大；Meta 本季转为负 11.6 亿美元，上年同期为正 111.7 亿美元。

图11：AI 算力投入首次烧穿四家巨头的自由现金流防线，单季合计自由现金流首次转负



数据来源：各公司公告、开源证券研究所

四家现金流缺口的性质并不相同，谷歌的消耗部分来自一个可回流的现金循环。谷歌的一部分现金占用是 TPU 系统备货：库存由年初的 24.39 亿美元增至 133.08 亿美元，签下 TPU 系统销售合同后订单先进入积压，公司随即备货、现金先流出，交付开票后现金再流回，按此口径该部分并非传统意义的资本开支，而是设备销售的营运资金循环。亚马逊与 Meta 的缺口则主要来自数据中心和服务器投入本身，属于真金白银的产能开支，现金流出后须经折旧与收入兑现的长周期才能回流。缺口的性质差异，决定四家对外部融资的依赖程度即将分道。

四家已全部动用或筹备外部融资工具，增量资金的规模与成本或将决定 2027 年资本开支的落地强度。亚马逊长期债务由 2025 年末 656.48 亿美元增至 9 月末 1731.25 亿美元，九个月新增加超 1000 亿美元、其中三季度单季新发 481 亿美元；谷歌本季按市价发行普通股净融资约 124 亿美元、发债净所得约 287 亿美元，回购为零，期末长期债务由年初 465 亿升至 1251.02 亿美元;Meta 长期债务由 587.44 亿增至 809.22 亿美元,并批准设立无投票权 C 类股,明年 1 月 19 日生效后以一比一股票股息派发、效果类似一拆二，为后续股权融资预留工具；微软本季融资租赁 74 亿美元、按起租本金计入资本开支，表态将在合适时候动用额外资金来源，信用评级赋予其发债成本优势。

AI 投资的约束条件正由投资意愿转向融资成本，外部资金环境对各家的约束强度沿现金流状况分层。信用评级较高、经营现金流充裕的公司，在利率上行或信用利差走阔的环境下仍可维持既定投资节奏；自由现金流已转负且扩产需求持续的公司，对融资条件的边际变化更为敏感。与此同时，融资租赁与数据中心合资的占比上升，资本开支中部分科目由资产负债表内的资产购置，转为摊销期内的租金与利息支出，微软已提示下一季度利息支出将包含数据中心融资租赁利息。**本轮算力基础设施投资的资金来源结构中，外部融资的边际贡献或将超过经营利润的留存再投资。**

2、 计算机产业动态更新

2.1、 国内 CSP 厂商与大模型厂商最新动向

（一）腾讯发布混元 Hy4 正式版，新增视觉与电脑操作能力

11 月 5 日，混元 Hy4 正式版在 preview 基础上扩大预训练与后训练规模，新增视觉和电脑操作能力。公司披露，GameDev、GameLogic 评测分别提升 19.1%、9.8%；在 579 个真实金融任务众测中得分 3.04，自建测试中的幻觉率由 15.76%降至 5.45%。

Hy4 已上线腾讯云与 OpenRouter，API 价格保持不变，模型权重计划开放。11 月 4 日接入 WorkBuddy 后，腾讯将推理集群扩容十倍并提高免费体验额度，但高峰时段仍可能限额。

（二）百度开源 ERNIE 5.5 Flash Base，支持百万 token 推理上下文

11 月 6 日，百度此次开放的 ERNIE-5.5-Flash-Base 为预训练终点权重，采用稀疏 MoE 架构，总参数 280B，每 token 激活 14B。模型以 33.5 万亿 token 从零预训练，训练上下文为 128K，推理支持 1M 上下文。

43 层网络中，37 层采用压缩注意力，每 128 个 token 仅缓存一条；模型配置 512 个路由专家，每 token 选择 10 个。FP8 和 BF16 两版权重已上线 Hugging Face，此次开放的是基座权重，经过后训练的版本将随后推出。

（三）DeepSeek Harness 上线创造模式，支持智能体生成并安装插件

11 月 4 日，DeepSeek Harness 新增“创造模式”，允许智能体根据用户的自然语言需求编写插件，并安装到自身界面。官方公开七套示例提示词，覆盖 3D 建筑展示、视频剪辑、海报制作、论文阅读及 token 计费等场景。用户可在新会话中选择创

造模式并输入需求。官方同时说明，部分演示经过多轮修改，单条提示词未必能够完整复现。

（四）Kimi 推出 Hosted Agents，提供企业智能体托管服务

11 月 6 日，面向企业智能体部署，Kimi 通过 Hosted Agents 将运行所需的 Harness 作为云端服务提供。企业无需自行搭建运行框架，即可使用全天候托管环境，支持长程任务和多智能体协作。服务提供 140 余项 Skills 和 30 余个插件，可接入企业自有工具与数据源。每个智能体配置独立沙箱和网络权限控制，企业可中止会话并导出审计记录；结合 Kimi Inference Hook，可在每次推理前按企业规则拦截敏感数据。目前服务仅向企业认证用户开放，首批提供八个金融智能体模板。

2.2、 海外 CSP 厂商与大模型厂商最新动向

（一）OpenAI 全量上线 GPT-Image-2.6，扩展动图与矢量图输出

11 月 6 日，GPT-Image-2.6 已在 ChatGPT、Codex 及 API 全量上线。OpenAI 披露，模型整体生成时间缩短 50%，最快约 6 秒出图。API 提供 Flare 和 Sunburst 两个型号，分别侧重生成速度与排版、多轮修改一致性，并新增 complexity 参数，将场景规划算力与画质设置分开控制。

动图通过先生成 3×2 的六帧序列图、再切片合成 GIF 实现，减少调用视频模型带来的等待时间。模型还支持原生 SVG 输出，局部圈选重绘同步向 API 开放。

（二）OpenAI 拟收购 Pierre，补充 Codex 的 Git 协作工作流

11 月 6 日，OpenAI 宣布收购 AI 代码托管公司 Pierre，拟改善面向人与智能体协作的 Git 工作流，使 Codex 覆盖从创建分支到部署上线的开发过程。Pierre 由 Bootstrap 联合作者 Jacob Thornton 创办，其开源库 Trees 和 Diffs 已用于 Codex 界面。Pierre 的 Code Storage 面向 AI 代码托管，单仓库容量最高 32TB，测试中每秒可处理超过 500 次 Git 请求。交易仍需获得监管批准。

（三）Anthropic 开源 FlashProve，以零知识证明核验模型来源

11 月 5 日，Anthropic 开源的 FlashProve 采用零知识证明，可在不公开模型权重的情况下核验回答是否由承诺的模型生成。在 4000 亿参数 MoE 模型 Llama 4 Maverick 上，单张 H100 的证明吞吐约为每秒 110 个 token，单份证明约 18.78MB，CPU 验证约需 17 秒。

该方案仍有较高计算成本：生成证明消耗的 GPU 时间约为 vLLM 推理的 70 至 95 倍，目前仅支持 1024 token 上下文、整数模型和贪婪解码。该系统仅核验回答是否出自指定模型，不验证回答本身的正确性。

2.3、 国产算力产业最新进展

（一）LongCat-2.5 披露国产算力训练与低精度推理优化

11 月 3 日，LongCat-2.5 采用原生多模态架构，总参数 1.6 万亿、每 token 激活约 480 亿参数，将视觉信号表示为多层级离散 token，与文本统一处理。训练加入长程轨迹数据和大规模长程强化学习，并将真实智能体的失败轨迹转为可验证任务。

团队披露，在国产算力平台上，长序列监督微调的单卡吞吐较 DDP 基线提升 86%；W4A4 推理相较 W8A8 基线，预填充和解码吞吐分别提升 43.3%、42.77%。

（二）小米开源 Xiaomi-OCR-0，以轻量模型统一文档解析任务

11 月 2 日，小米将文档解析、关键信息抽取与视觉问答整合至 8 亿参数的 Xiaomi-OCR-0，并开放权重、代码和 MCP 服务。模型基于 Qwen3.5-0.8B-Base，以约 1.7 亿条样本训练，在 Real5-OmniDocBench 取得 95.24 分，在 OmniDocBench v1.6 取得 96.83 分，五项视觉问答任务平均得分 83.2。相关技术已融入超级小爱，用于取餐码展示、自动记账等应用。

（三）腾讯混元发布 Hyra 2.0，支持科学模型研发与处理器优化

11 月 3 日，腾讯混元的 Hyra 2.0 采用 Overall、Mission、Worker 三层协作架构，连接企业微信、计算平台和代码仓库，并将研究经验沉淀为可复用 Skills。系统可组织模型训练、工具开发及计算实验。团队披露，从头训练的 196 万参数 HyInv 较 ProteinMPNN 快 57 倍，6.5 亿参数 HyPLM-Flash 在参与对比的 12 个模型中推理速度最快。处理器优化实验中，香山处理器每 GHz 得分提升 5.79%，模糊测试发现 10 个此前未报告的缺陷。

2.4、 海外算力产业最新进展

（一）Marvell 拟收购 Lumentum，扩展 AI 数据中心光互联布局

11 月 2 日，Marvell 宣布拟以全股票方式收购 Lumentum，隐含企业价值约 1150 亿美元。Lumentum 在激光器及光通信器件领域的产品，与 Marvell 的数据中心互联业务形成衔接，交易涉及近封装光学与共封装光学等 AI 基础设施配套方向。根据公布的方案，每股 Lumentum 股票将换取 4.5 股 Marvell 股票，交易完成后双方原股东持股比例约为 67%和 33%。交易尚待完成。

（二）谷歌披露 TPU 在轨运行试验，太空算力仍处验证阶段

11 月 2 日，谷歌披露 Project Suncatcher 试验进展，谷歌原型卫星于 10 月 1 日入轨，搭载四颗 TPU 并运行 Gemma 4，完成 AI 响应生成与回传。卫星最高供电约 1kW，每次仅启动一颗 TPU，运行约 20 分钟，后续将测试热应力和辐射影响，计划于 2027 年再发射两颗配备光通信的卫星。

项目测算认为，若发射成本降至每公斤 200 美元，太空算力才可能接近地面成本。项目负责人表示，五年内难以比地面更便宜。项目仍处于在轨验证阶段。

（三）得州拟延长数据中心审批冻结，在建项目不受新许可暂停影响

11 月 6 日，财联社报道，得州候任州长提出，在立法完成前暂停数据中心新许可和并网审批，在建项目除外，同时要求大型用电户承担新增发电、输电及并网成本，并逐步取消数据中心税收减免。

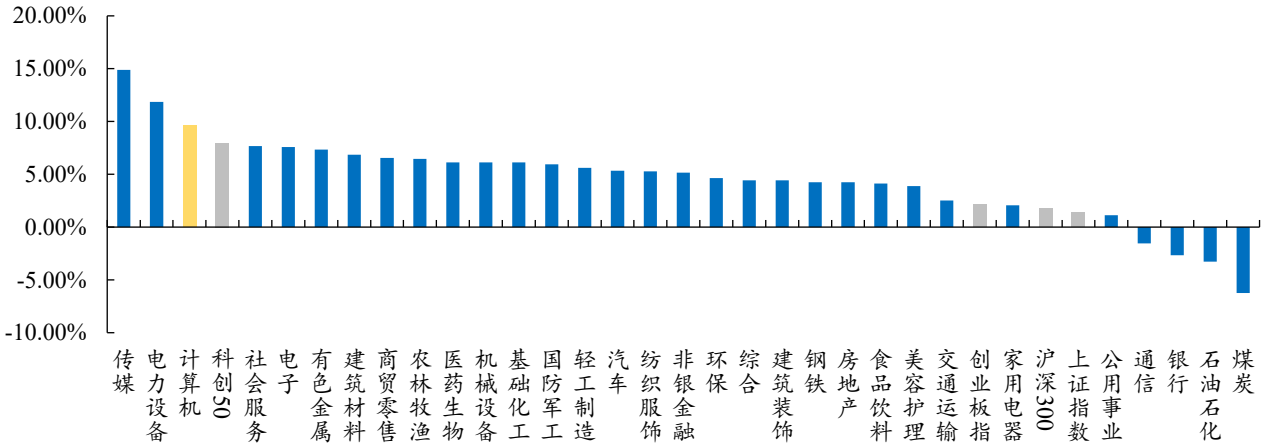
得州现任州长已于 8 月要求审查排队并网的数据中心，9 月暂停全部数据中心许可。州议会将于 2027 年 1 月 12 日开会，常规会期至 5 月 31 日结束；电网运营商 ERCOT 计划于 12 月 10 日提交审查报告，但报告本身不会重启审批，冻结何时解除仍取决于后续立法进展。

3、 上周市场表现回顾（2026/11/2-2026/11/6）

计算机指数上涨 9.60%，相对主要市场指数表现较强。同期沪深 300 指数上涨 1.73%，上证指数上涨 1.39%，创业板指上涨 2.14%，科创 50 指数上涨 7.96%，计算机指数分别跑赢 7.88、8.21、7.47 和 1.65 个百分点。

横向比较一级行业，本周传媒、电力设备、计算机表现居前，分别上涨 14.90%、11.85%和 9.60%；计算机板块上涨 9.60%，在 31 个一级行业中排名第 3 位。与此同时，通信、银行、石油石化和煤炭分别下跌 1.55%、2.66%、3.26%和 6.26%。整体来看，在本周 TMT 板块内部走势分化的背景下，计算机板块相对表现较强。

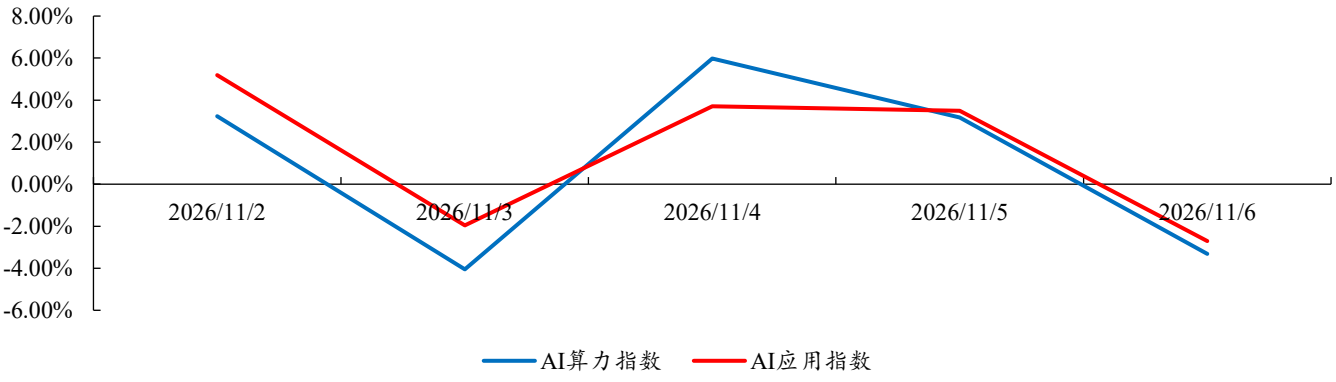
图12：本周多数一级行业上涨，计算机板块涨幅居前



数据来源：Wind、开源证券研究所

本周 AI 细分方向呈现一定分化，**AI 应用指数整体表现明显优于 AI 算力指数。**按照日度涨跌幅复合计算，11 月 2 日至 11 月 6 日 AI 应用指数累计上涨约 **7.71%**，AI 算力指数累计上涨约 **4.73%**，AI 应用相对 AI 算力取得约 **2.98 个百分点**的超额收益。

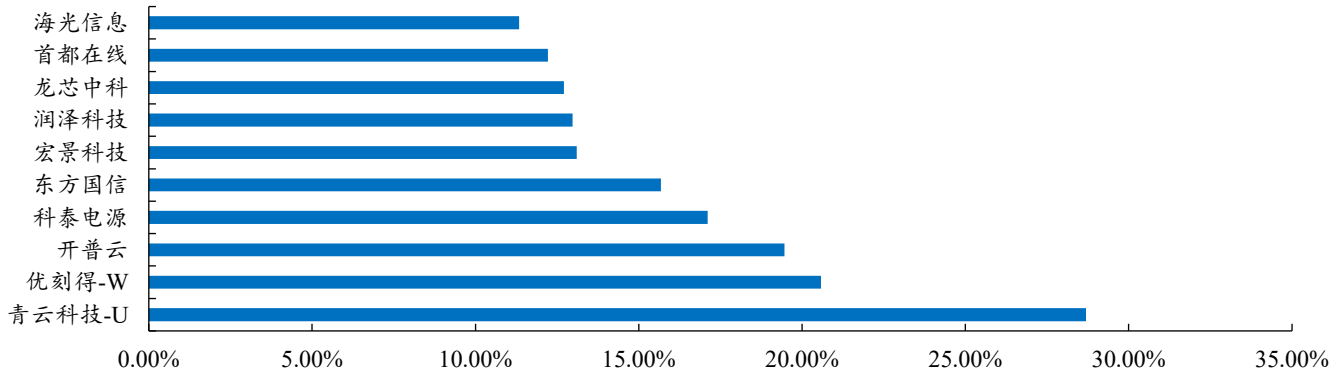
图13：本周 AI 细分方向均有所上涨，AI 应用指数整体表现优于 AI 算力指数



数据来源：Wind、开源证券研究所

AI 算力指数整体上涨，但**板块内部仍存在较为明显的结构性行情**。青云科技-U 本周上涨 **28.69%**，优刻得-W 上涨 **20.58%**，均录得两位数涨幅。

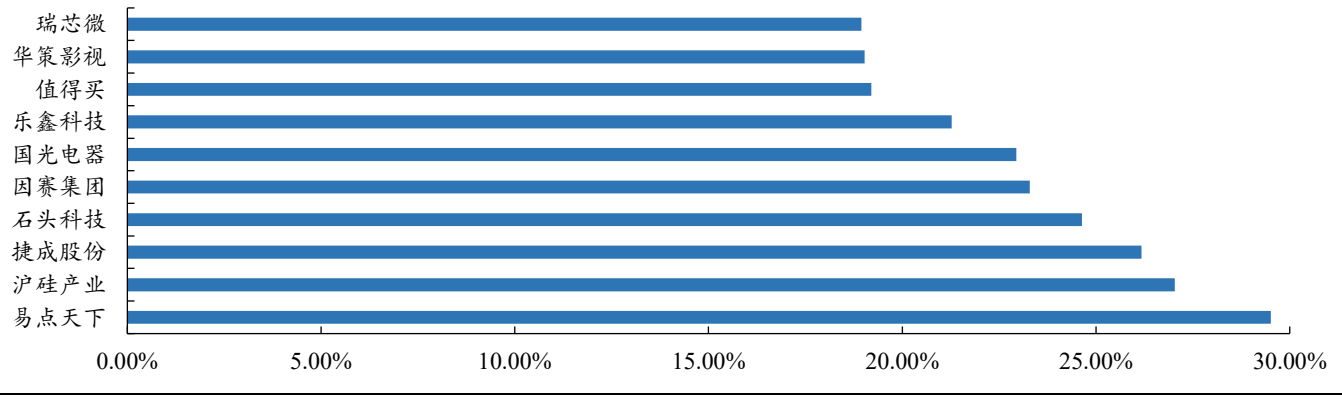
图14：AI 算力指数整体上涨，青云科技-U 等个股表现突出



数据来源：Wind、开源证券研究所

AI 应用方向指数涨，龙头个股赚钱效应更为突出。涨幅前十个股周涨幅全部超过 18%，易点天下本周上涨 **29.51%**，位居 AI 应用板块涨幅首位。

图15：AI 应用方向指数整体上涨，易点天下涨幅居前



数据来源：Wind、开源证券研究所

4、 风险提示

大模型技术迭代不及预期风险。大模型技术迭代是推动产业能力边界拓展与下游场景深化的核心驱动力，当前行业正处于多模态感知、复杂推理、智能体执行等前沿技术的攻坚突破期，技术突破节奏、能力落地效果均存在一定不确定性。若后续核心技术突破进度、模型能力提升幅度低于市场预期，无法有效支撑下游高价值场景的深度适配与功能升级，可能导致产业落地节奏整体延后，行业成长进程存在阶段性放缓的可能性。

AI 技术应用不及预期风险。下游场景规模化渗透是大模型产业价值兑现的核心支撑，当前大模型正从技术概念验证期向垂直行业规模化应用期过渡，落地进度受行业数字化基础、客户付费意愿、投入产出比验证等多重因素共同制约。若后续垂直行业适配进度、企业付费转化节奏不及市场预期，产业商业价值的兑现周期可能有所拉长，行业长期成长动能的释放节奏存在一定波动可能。

相关政策不及预期风险。大模型产业发展高度依赖政策环境的引导规范与配套支持，当前人工智能领域处于快速发展阶段，数据安全合规、技术监管标准、产业

扶持政策、市场准入规则等相关制度仍处于持续完善过程中。若后续行业监管要求趋严、扶持政策落地进度慢于预期，或合规要求抬升导致企业研发与运营成本增加，可能对大模型技术迭代、商业化落地及行业竞争格局形成阶段性扰动，行业发展节奏存在相应调整的可能性。

AI Infra 建设不及预期风险。AI 基础设施是大模型技术迭代与产业规模化落地的底层算力与工程支撑，算力集群建设、高速互联部署、存储体系配套等基建环节具备投入规模大、建设周期长、产业链协同要求高的特征，落地进度受硬件供给、资金投入、工程能力等多重因素约束。若后续高端算力供给受限、基建项目建设进度慢于市场预期，可能形成算力供给瓶颈，制约大模型训练迭代与应用规模化落地，行业整体发展节奏存在阶段性承压的可能。

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R3（中风险），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专业投资者及风险承受能力为C3、C4、C5的普通投资者。若您并非专业投资者及风险承受能力为C3、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。		

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