下载日志已记录，仅供内部参考

中信建投证券 CHINA SECURITIES

证券研究报告·海外行业深度

2Q26 美股科技财报总结及 2H26 前瞻

计算机

核心观点

前沿实验室 ARR 边际放缓主因供应瓶颈与算法降本，底层算力需求依旧紧俏。三大云厂商 Q2 业绩提速且商业 Backlog 突破万亿美元（锁定 5–10 年长约），有力验证了 AI 基建投资回报的长期能见度；后市核心矛盾转向巨额 CapEx 带来的折旧爬坡，具备自研芯片闭环与现金流造血能力的标的（AMZN、GOOGL）具备更优的单位经济模型，MSFT 依赖集群并网放量，META 需观察折旧对利润率的挤压。软件层方面，专有业务图谱与合规权限确立了核心系统的不可替代性，企业预算正加速回流至工作流枢纽（NOW、SAP）、数据栈（SNOW）与安全防线（CRWD），板块分化聚焦于长约锁定度与现金流自洽度。

核心要点

产业循环与前沿实验室：供给瓶颈与算法降本共振，真实需求韧性依然强劲。 将 AI 产业还原为“前沿实验室 ARR → 云厂商 CapEx → 算力供给 → 智能反哺”的飞轮循环，1H26 OpenAI 与 Anthropic 名义 ARR 的边际减速并非需求停滞，而是物理电力与芯片供应受限、算法工程调优（大幅压降推理 Token 成本与系统级冗余）以及企业经历年初 FOMO 采购后进入理性消化期共同作用的结果。深入拆解收入口径可知，前沿实验室的 ARR 强依赖消耗型计费与头部 B 端客户（Top 1% 客户贡献超 80% 营收），与传统 SaaS 锁定 3–5 年不可撤销合同的模式存在本质差异。剔除单一大客户异常采购、分销渠道抽成与非经常性买断后，真实净增量需求依然紧密锚定底层算力扩张，算力供不应求的格局并未改变。

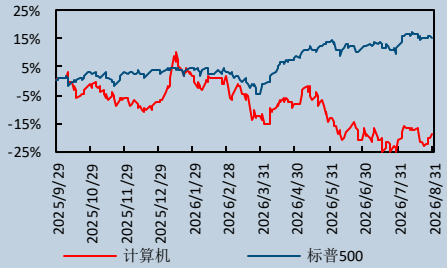
万亿美元长周期订单托底，AI 投资回报（ROI）加速兑现。 Hyperscalers Q2 业绩并不支持市场失速疑虑，Azure（+43%）、AWS（+37%）与 Google Cloud（+82%）均展现出极强的增长弹性。更重要的是，三大公有云的商业待履约合同（RPO/Backlog）分别达到 6,780 亿、4,960 亿与 5,140 亿美元，平均履约周期拉长，且 AWS 长期合同贡献当期收入比例达到 46% 的历史高位，充分验证了 AI 资本开支正在转化为高确定性的未来现金流，破除了“算力过度投资导致空转”的伪命题。尽管单季巨额 CapEx 带来的转固折旧（D&A）爬坡对毛利率与即期自由现金流形成阶段性压制，但云厂商通过自研芯片（Trainium/TPU/Maia）规模化替代与非云核心业务的持续造血，正逐步构筑起单位经济学层面的抗压护城河。

维持 强于大市

于伯韬
yubotao@csc.com.cn
SAC 编号:S1440520110001
SFC 编号:BRR519
许悦
xuyue@csc.com.cn
SAC 编号:S1440523030001

发布日期： 2026 年 09 月 29 日

市场表现



相关研究报告

07.06.11	股权变更获准
07.03.29	增资白敬宇制药持有 30%股份
07.03.05	63%控股鼓楼宿迁人民医院

本报告由中信建投证券股份有限公司在中华人民共和国（仅为本报告目的，不包括香港、澳门、台湾）提供。在遵守适用的法律法规情况下，本报告亦可能由中信建投（国际）证券有限公司在香港提供。同时请务必阅读正文之后的免责条款和声明。

下载日志已记录，仅供内部参考

软件板块价值重塑：打破“席位灭顶”悲观预期，IT 预算加速回流核心生产力底座。企业级软件板块正经历从过度抛售到理性重估的转折点。LLM 作为概率型推理工具，无法脱离专有业务图谱、合规审计链条与跨系统权限而独立运行，系统记录层（System of Record）与行动系统（System of Action）的核心壁垒在智能体时代不仅未被颠覆，反而进一步强化。头部 SaaS 正积极推进“席位底费 + 智能体交互/成果抽成”的混合商业模型以对冲席位收缩风险；同时，在缺乏高质量私域数据清洗与工作流编排的情况下，通用大模型难以直接进入核心生产环节，这驱动企业 IT 预算从前期的通用 API 试水加速回流至现代数据栈（SNOW/MDB）、业务调度中枢（NOW）与网络安全防线（CRWD/PANW）。

进入 2H26，投资者的关注从“算力军备竞赛想象力”全面转向“订单可见度、自研芯片单位成本优势与现金流自洽度”。在大型科技巨头方面，各标的因基建路径与商业模式差异呈现清晰的分化特征：亚马逊（AMZN）依托较长的 Backlog 锁定期、自研 Trainium 闭环以及零售业务充沛的现金流支撑，展现出较强的单位经济学与抗风险优势；谷歌（GOOGL）凭借全栈 TPU 构筑的算力成本壁垒以及搜索 TAC 结构性下行的潜在契机，具备稳固的利润安全垫；微软（MSFT）后续的关键变量在于 Copilot 及其他 AI 应用的推进商业化节奏；Meta（META）在缺乏外部企业级云订单对冲的背景下，高额 CapEx 带来的刚性折旧对利润率的潜在侵蚀仍需持续观察。在软件板块方面，行业正由前期的估值收缩转向业务实质分化：ServiceNow（NOW）依托 Pro Plus 提价通道与企业 AI 编排中枢地位展现出较强韧性，CrowdStrike（CRWD）受益于自动化威胁升级下的安全预算平台化整合，SAP 享有 2027 年 ECC 支持终止带来的长周期刚性迁徙红利，而 Snowflake（SNOW）则逐步体现出企业私域数据原位计算带来的用量回升。整体而言，板块后续的分化主线主要围绕长期订单锁定度、自研硬件单位经济学以及资本支出与现金流的自洽度展开。维持“强于大市”评级。

下载日志已记录，仅供内部参考

目录

AI Labs ARR 复盘 1
行业长期订单改善，AI 收入/Capex 覆盖率快速提升 15
北美科技财报总结及前瞻 17
微软（Microsoft / MSFT）：Azure 43%提速破千亿，Copilot 加速转化带动叙事改善 17
谷歌（Alphabet / GOOGL）：云持续加速，关注核心广告业务利润率 18
Meta（META）：广告推荐算法变现效率持续验证，关注 C 端 AI 变现的边际 ROI 18
亚马逊（Amazon / AMZN）：AWS 37% 创 18 季度新高，自研芯片造就最强单位经济模型 19
软件层：分化与 AI 转型仍在早期阶段 19
投资评价和建议 22
风险分析 23

图目录

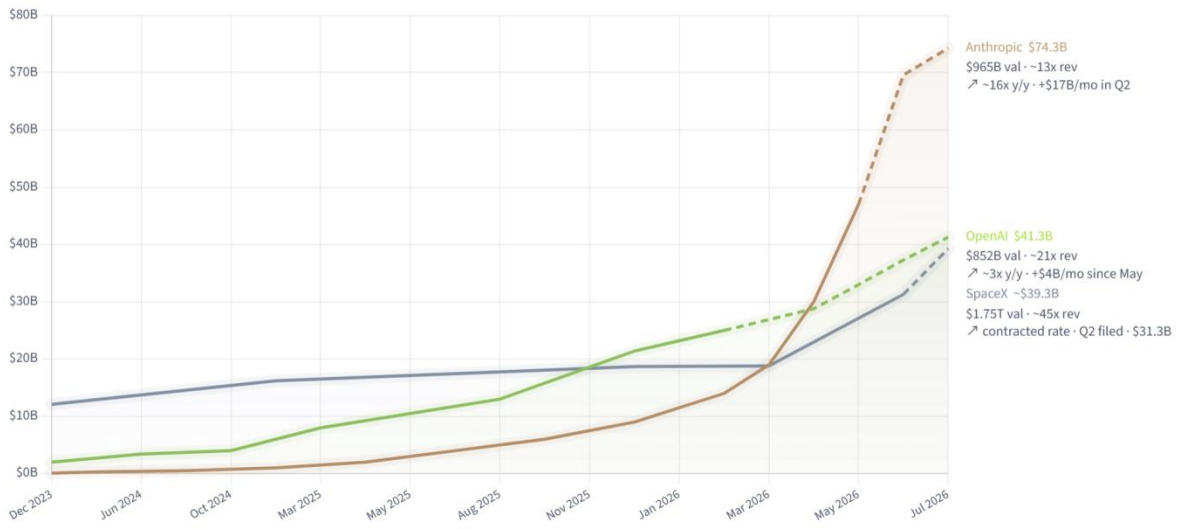
图 1:OpenAI/Anthropic ARR 趋势（1H26 明显加速） 1
图 2:OpenAI/Anthropic 净增 ARR 趋势 1
图 3:SOX 年内涨幅趋势与 NNARR 相关 2
图 4:OpenAI/Anthropic ARR~80%由 top 1%企业贡献 2
图 5:Codex/Cowork 活跃用户增长趋势 4
图 6:Claude/GPT 订阅套餐的实际额度 5
图 7:Codex 1%额度的动态变化 5
图 8:Opus 5.5 近期固定测试集表现略有回落 6
图 9:Opus 5.5 近期固定测试集运行时间大幅提升 6
图 10:GPT-6 Sol 固定测试集表现明显恶化 7
图 11:GPT-6 Sol 固定测试集运行时间大幅提升 7
图 12:OpenAI/Anthropic AIDC 供给趋势（能耗估计） 8
图 13:benchmark 性能接近的对标模型，消耗 token 比例 9
图 14:benchmark 性能接近的对标模型，消耗成本比例 9
图 15:benchmark 性能接近的对标模型，消耗 token 比例 10
图 16:benchmark 性能接近的对标模型，消耗成本比例 10
图 17:Claude 系列模型分词器 token 消耗显著较高 11
图 18:分词器导致 Claude 系列模型处理不同类型的任务 token 消耗均高于其他模型 11
图 19:Claude 系列模型的 token 消耗膨胀较快 12
图 20:Claude 在搜索查询中消耗的 token 数量是 GPT 的 4-5 倍 13
图 21:Anthropic ARR 与 IDC 供给趋势 14
图 22:OpenAI ARR 与 IDC 供给趋势 14
图 23:AI Labs ARR/CSP Capex 覆盖率 15
图 24:AZURE/AWS/GCP RPO/backlog 订单增长趋势 16
图 25:即使剔除 AI Labs 订单，其他云订单的增长也好于长期趋势 16

下载日志已记录，仅供内部参考

AI Labs ARR 复盘

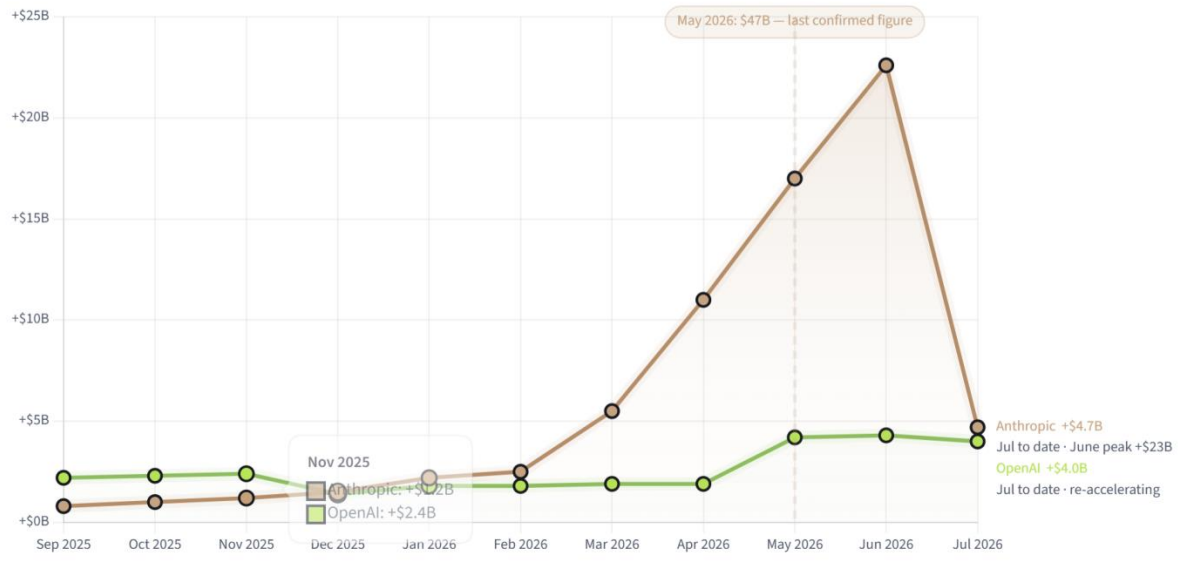
把 AI 行情还原为一个循环：模型智能 → 前沿实验室 ARR → 云/新云 CapEx → 算力供给 → 反哺智能。判断"供给创造需求"是否仍成立，市场直接跟踪的主要是 Anthropic 与 OpenAI 的 ARR 增量（二阶变量）。

图 1:OpenAI/Anthropic ARR 趋势（1H26 明显加速）



数据来源：Dealroom，中信建投证券

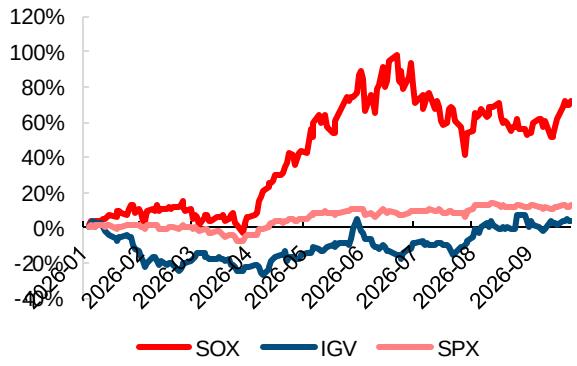
图 2:OpenAI/Anthropic 净增 ARR 趋势



数据来源：Dealroom，中信建投证券

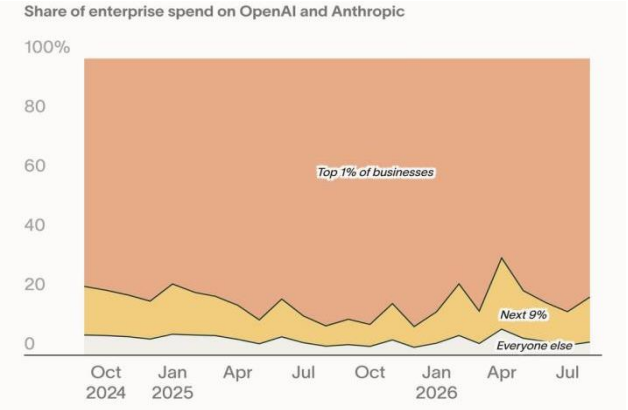
下载日志已记录，仅供内部参考

图 3:SOX 年内涨幅趋势与 NNARR 相关



数据来源：Wind，中信建投证券

图 4:OpenAI/Anthropic ARR~80%由 top 1%企业贡献



数据来源：Ramp，中信建投证券

回到 ARR 数据，市场习惯用传统企业级 SaaS（如 ServiceNow、Workday）的估值框架来审视 AI 前沿实验室，但两者的 ARR 在定义与波动性上存在根本错位：传统 SaaS 企业基于长期固定合同，按年/季度确认收入，履约确定性极高，例如 Workday 明确披露其核心 HCM 及财务管理合同通常为 3 至 5 年，且均为不可撤销协议，客户仅在极罕见的严重违约情况下才有权终止，且报表端按天/月直线法分摊（Ratable Recognition），收入确认的确定性较高。

头部 SaaS 的订阅收入几乎全部由长周期合同（RPO）覆盖，且拥有极高的存量锁定比例。Workday 明确披露其核心 HCM 及财务管理合同通常为 3 至 5 年，且均为不可撤销协议（Non-cancelable Agreements），客户仅在极罕见的严重违约情况下才有权终止。

AI Labs 的 ARR 采用近期月度（MRR × 12）或连续 4 周收入（4-Week Revenue × 13）进行年化，强依赖 Usage/Token 消耗，对业务峰值、大模型版本上线极度敏感。其商业模式绝大部分是“按 Token 实际消耗量计费（Usage-based / API Consumption）”与“月度无绑定订阅（ChatGPT Plus/Pro 每月 $20–$200）”，传统意义上 3 年期锁死的 Committed ARR 占比较小。

依据 ASC 606 准则，GAAP 收入必须按照履约义务的履行进度严格确认（例如：按实际消耗 Token 的周期结转，或预付 Credit 按消耗比例分摊）。2026 年 Q1，OpenAI 披露的实际入账 GAAP 营收为 $5.7B，Q2 为 $6.7B（上半年合计入账仅约 $12.4B）；但其对外宣称的 Run-rate ARR 在 3 月达到 $25B，8 月达到 >$40B，如果按期间确认收入/披露 ARR 比例看，OpenAI/Anthropic 低于一般的 SaaS 企业，业务的波动性较高。

早在 2026 年 4 月，OpenAI 首席营收官（CRO）Denise Dresser 联合 CFO 在四页员工备忘录中针对 Anthropic 发起公开质疑，指出 ASC 606 会计准则下的毛额法（Gross）虚增与非经常性合同问题，OpenAI 认为①Anthropic 的 ARR 中包含非经常性一次性合同$3.5B - $4.0B (部分大企业客户为锁定折扣，一次性买断数亿美元预付额度，被直接乘以 12 计入 ARR)；② 云厂商间接渠道抽成 (AWS/GCP Take Rate)约有$4.0B - $4.5B[1]，最终可审计真实净额 ARR 约 $22B 左右。

我们认为，这为 Anthropic IPO 文件调整 ARR 口径埋下伏笔，7 月 Bloomberg 等媒体报道 Anthropic ARR 达

[1] 据 Semianalysis，Anthropic 1H26 可能有~40%收入来自 AWS Bedrock / Google Cloud 等间接渠道，分成率一般在 15-20%（一些合同分成基数为总收入，一些合同分成基数为“扣除硬件基础设施成本后的毛利”）。

下载日志已记录，仅供内部参考

650 亿美元，Atreides Management 投资人 Gavin Baker 提到，Anthropic IPO 文件中的 ARR 数据变化主要是一些风险管理因素导致的，例如从 Gross 口径转向 Net 口径，同时剔除了 Meta 和中国相关的收入来源，确保对于单一大客户/竞争因素导致的收入波动[2]。另一方面，市场投资者以第三方机构的高频 ARR 数据作为锚点本身存在不可衡量的误差，① 三方机构通过网站点击量来评估 OpenAI 的商业化走势“从方法论上就是错误的”。因为随着企业客户将工作流深度集成至 API 以及专用算力集群，高价值、高付费密度的商业负载正在全面移出 Web 界面[3]。② 信用卡抽样数据仅能反映部分特定地域的小规模行为，无法穿透 OpenAI 面向《财富》500 强企业的直销体系。

三方数据跟踪的本质是特定渠道（例如 C 端订阅/网页渠道），特定客群（C 端用户+SMB 企业）的用量趋势，因此在营收结构变化的情况下，例如 OpenAI 企业端收入占比从 26 年初的 40%提升到 26 年 7 月的 50%+，6-7 月 OpenAI 企业端 ARR 环比增长 32%，超越整体 20%的环比增长。这意味着三方数据跟踪的误差可能越来越大（C 端/小 B 的消耗趋势与整体背离）[4]，整体 ARR 的结构中~88%来自 Top 1%的大型企业，而三方数据跟踪的优势模块在于中小企业/个人用户，且这一结构并不稳定。

回顾 OpenAI Codex/ChatGPT Work 的周度活跃用户，7 月 Tibo（OpenAI 核心产品负责人）开启 Codex 额度重置后至今不足 3 个月周度活跃用户从 600 万增长至 2500 万（合并了 ChatGPT Work），但同期的 ARR 并未翻 4x+，原因在于订阅套餐的商业模式锁死了收入上限，重度用户越多毛利率反而越低（固定收入，但是用量越多成本越高）。

表 1:Codex/Cowork 用户量增长里程碑

时间节点	官方披露/确认指标	关键驱动事件与产品动态
2026 年 5 月 31 日	活跃用户突破 500 万	首次因“里程碑”进行全员速率重置，拉开与竞品（Cursor/Claude）额度补贴战序幕。
2026 年 7 月 09 日	活跃用户突破 600 万	修复社区反馈的代码评审并发 Bug，宣布全量 Reset。
2026 年 7 月 13 日	活跃用户突破 700 万	上线针对 GPT-5.6 Sol 的底层 Token 节省算法；全量推出“库存重置（Banked Reset）”。
2026 年 7 月 15 日	活跃用户突破 800 万	Codex 与 ChatGPT Work 合计周活暴增；解除 5 小时硬性时间锁。
2026 年 7 月 16 日	活跃用户突破 900 万	仅耗时 1 天再增 100 万，连续多日全量充值额度。
2026 年 8 月 12–13 日	正式跨过 1,000 万关口	Tibo 发推：“承诺过 1,000 万前每增 100 万就 Reset 一次，现已远超”；同时灰度测试**“8 美元单次额度直接重置（Paid Reset）”。
2026 年 8 月 21 日	活跃用户突破 2,000 万	金十/币安等引述官方数据，伴随企业端广泛引入自动代码安全审查，Codex 活跃量翻倍。

[2] https://minutesof.com/gavin-baker/

[3] https://www.eweek.com/news/chatgpt-hits-400m-weekly-active-users/

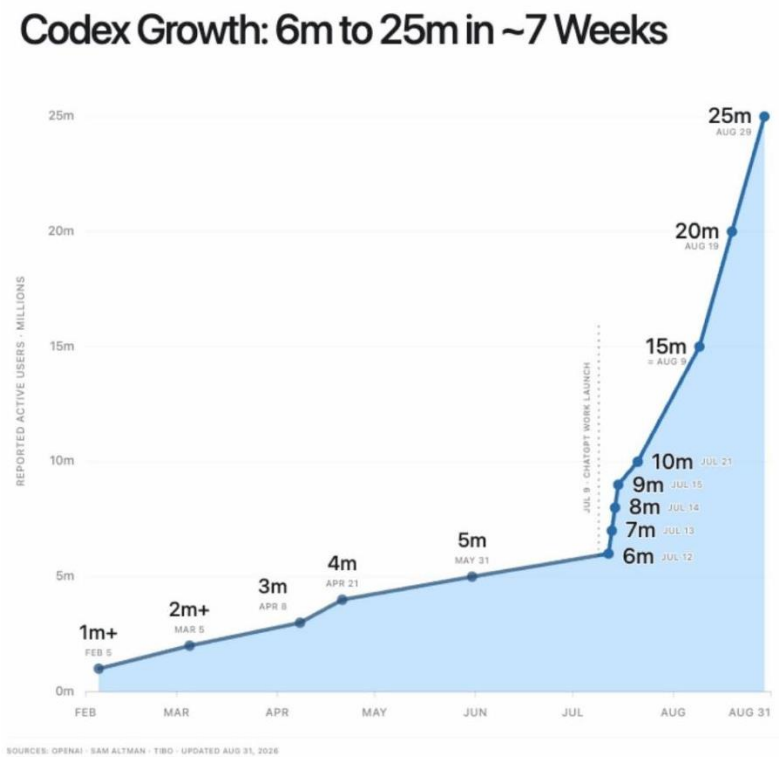
[4] https://www.pymnts.com/news/artificial-intelligence/2026/openai-and-anthropic-get-80percent-of-revenue-from-1percent-of-customers/

下载日志已记录，仅供内部参考

2026 年 8 月 30 日	突破 2,500 万活跃用户	Tibo 发推庆祝达到 25M Active Users，再次为所有 ChatGPT Work 及 Codex 付费用户发放全员 Reset。

资料来源：OpenAI，Tibo，中信建投证券

图 5:Codex/Cowork 活跃用户增长趋势



数据来源：OpenAI，Tibo，中信建投证券

Semianalysis 实际测算 2026 年 6 月 Anthropic/OpenAI 的套餐价格和实际用量，这里的实际用量是按照目录价格（API 展示的 token 价格）估算的，并非实际成本。理论上用量最深的一批用户可能完全打穿订阅套餐的收入，也就是负毛利提供服务，这种情况下模型团队实质上是对外提供补贴。2026 年 6 月 30 日,The Information 报道称，OpenAI 工程师通过系统级优化，在不增加新芯片的情况下，将模型推理成本降低了超过 50%，如今仅需数百块英伟达 GPU 即可服务 ChatGPT 的全部匿名用户流量。核心技术手段包括：量化压缩、KV-Cache 优化、连续批处理、投机解码以及优先级调度。

下载日志已记录，仅供内部参考

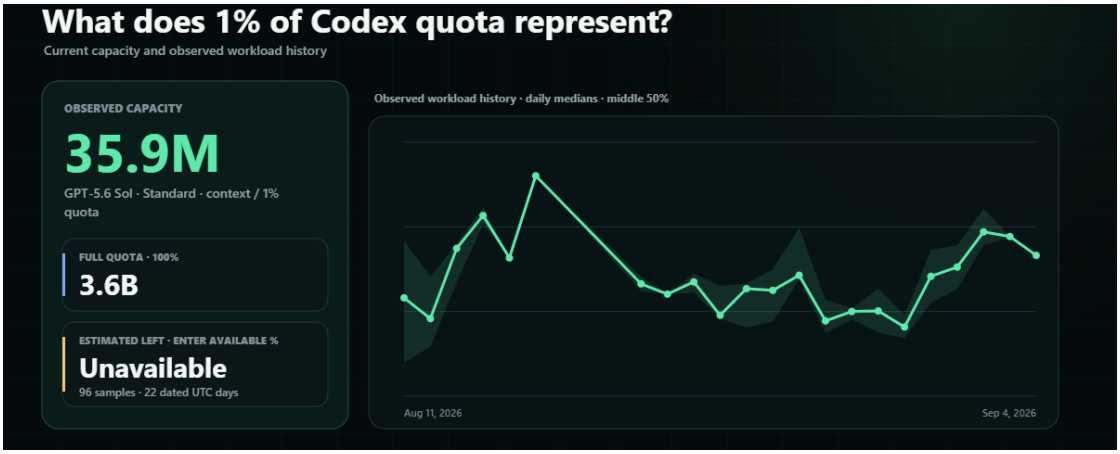
图 6:Claude/GPT 订阅套餐的实际额度

PLAN	PLAN PRICE	MAX POSSIBLE SPEND (APPROX.)
ANTHROPIC — CLAUDE		
claude-pro	$20/mo	$400/mo
claude-max-5x	$100/mo	$2,000/mo
claude-max-20x	$200/mo	$8,000/mo
OPENAI — CHATGPT / CODEX		
chatgpt-plus	$20/mo	$700/mo
chatgpt-pro-5x	$100/mo	$3,500/mo
chatgpt-pro-20x	$200/mo	$14,000/mo

数据来源：Semianalysis，中信建投

OpenAI Developer Community（2026 年 8 月 6 日帖）[5]：Pro 用户报告过去两周 usage limits 大幅下降。实测：消耗 195.5M tokens（GPT-5.6 Sol High-xHigh）= 周度限额的 28% → 推算周度限额≈ 700M tokens。对比同是 $100 的 Claude 套餐 ≈1.5B tokens（同类型工作）。多人跟帖确认限额大幅下降，以前重置卡都用不完，现在需严格控制。2026 年 6 月也有类似情况[6]，Pro 20x 用户报告 周度限额大致从 3B tokens 骤降至 0.5–0.7B tokens，单对话 30M tokens 吃掉 6%。Codex 实际用户用量跟踪数据也显示花完 1% Codex 额度（图中示例为 Pro 20x 套餐）消耗的 token 并不是固定的，而是区间内波动。

图 7:Codex 1%额度的动态变化



数据来源：Codex quota[7]，中信建投证券

[5] https://community.openai.com/t/chatgpt-pro-usage-limits-have-dropped-significantly/1389324

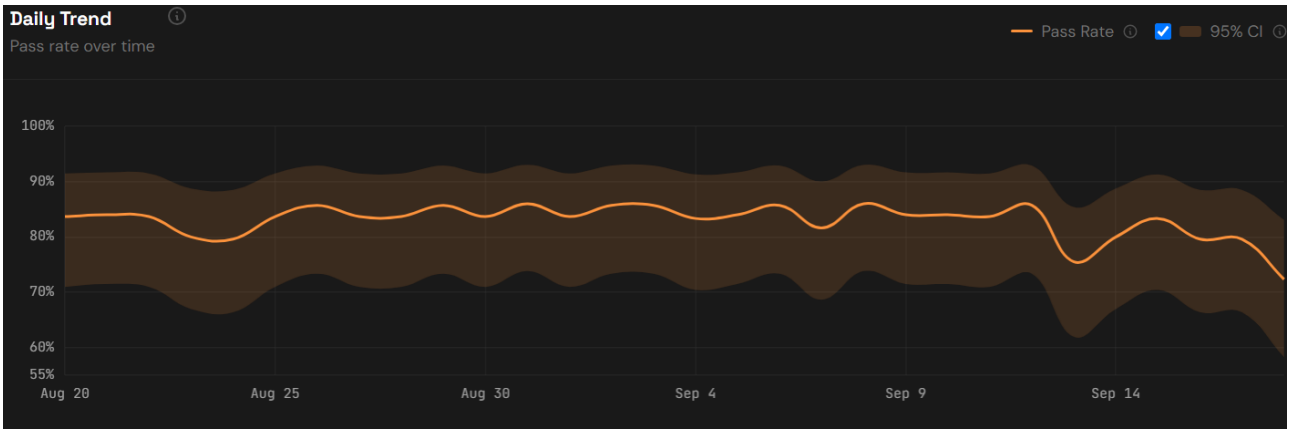
[6] https://community.openai.com/t/abnormal-pro-20x-weekly-limit-consumption/1383944

[7] https://codex-quota.manetli.com/?model=gpt-5.6-sol&plan=pro_x20&unit=tokens#economics

下载日志已记录，仅供内部参考

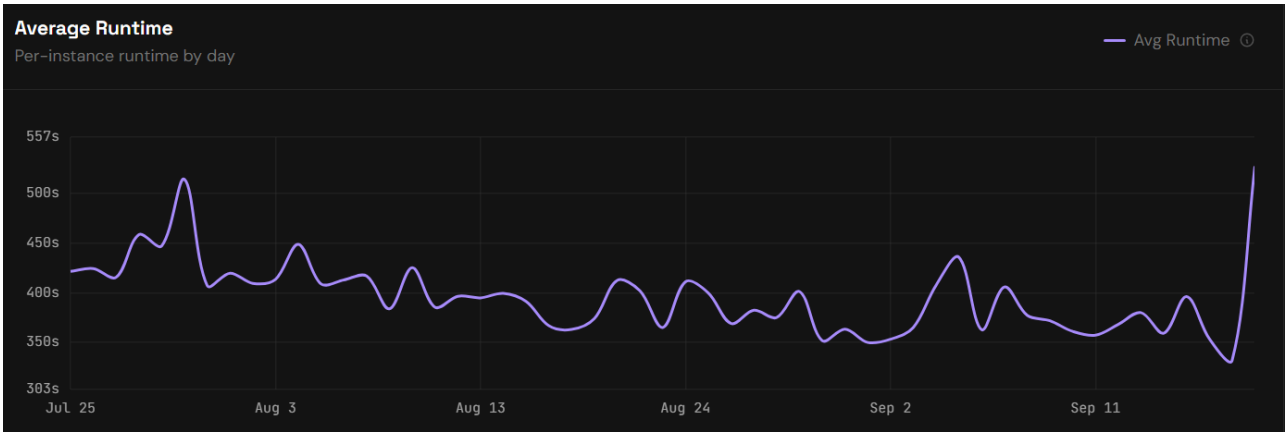
另一个角度是用固定测试精选集跑benchmark的pass rate，Claude Code上Opus 5.5的表现基本稳定，pass rate过去30天在84%左右，但是消耗token趋势性下降（反映了内部降本的成效），input/output token过去30天也下降了约20%。

图 8:Opus 5.5 近期固定测试集表现略有回落



数据来源：Marginal Lab[8]，中信建投证券

图 9:Opus 5.5 近期固定测试集运行时间大幅提升



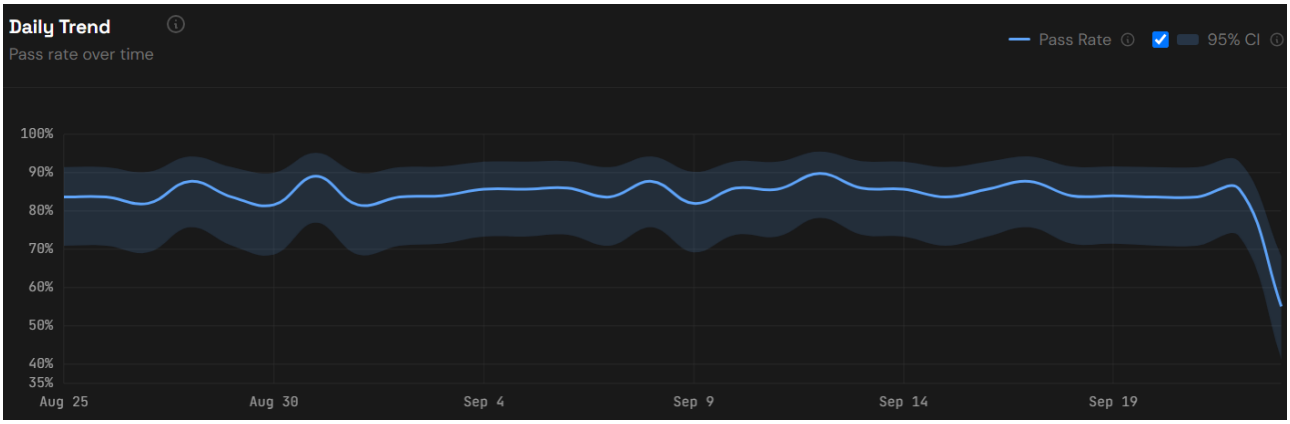
数据来源：Marginal Lab，中信建投证券

比起收入侧的变化，成本/毛利端的变化更能反映商业化进展（调整定价似乎更难以被广泛接纳）。过去几年，供应限制下OpenAI/Anthropic通过工程优化、算法迭代等不断降低单位token的推理成本，使得订阅套餐的毛利率提升。根据The information报道，OpenAI 24年1月付费用户毛利率35%，24年底提升至52%，25年10月付费客户板块的计算毛利率一度被优化至~70%，整体~40%的毛利率主要受免费chat用户拖累。Anthropic 25年毛利率约为40%，26年上半年推理侧毛利率（仅考虑计算相关的COGS后，未考虑向渠道支付的佣金分成）~70%。

[8] https://marginlab.ai/trackers

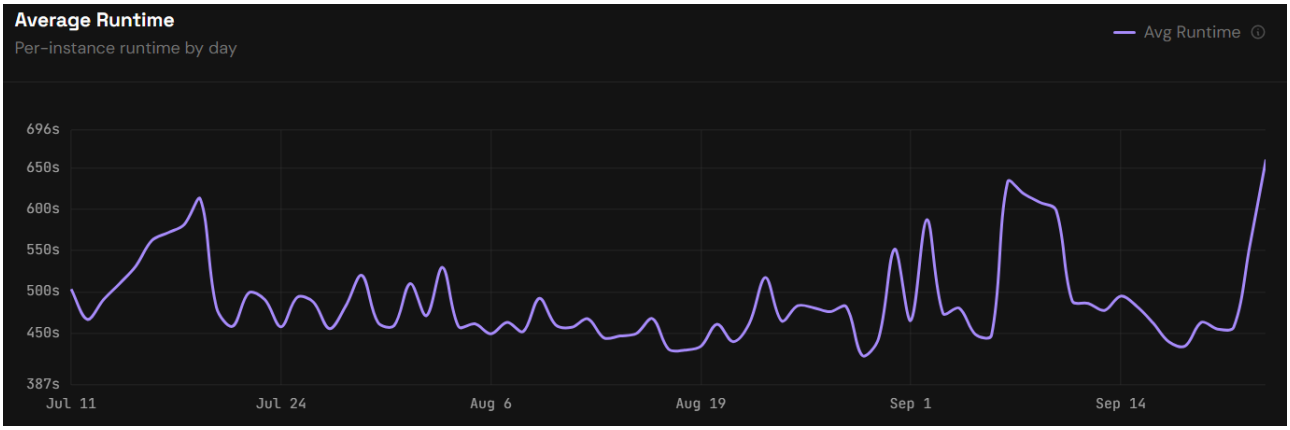
下载日志已记录，仅供内部参考

图 10:GPT-6 Sol 固定测试集表现明显恶化



数据来源：Marginal Lab，中信建投证券

图 11:GPT-6 Sol 固定测试集运行时间大幅提升



数据来源：Marginal Lab，中信建投证券

考虑一点，OpenAI 的 Vera Rubin 首批交货在 8 月 20 日，而且主要用于模型训练，即便考虑替换的算力集群用于填补推理需求，其实质的供应增量也在 8 月。然而 7-9 月 4x+的活跃用户增长期间算力供应增长幅度远远无法匹配需求。2026 年 7–9 月，微软 FY27Q1 提到，2Q26 Fairwater Wisconsin（Mount Pleasant）交付数百上千 GB200 等 GPU 连成单一集群（估算 IT power ~300–369 MW，约 15 万+ GPU 量级/建筑），7–9 月主要是；Oracle 等 FY27Q1（截止 2026/8/31）交付 618 MW（13.1 万块 GPU，主要是 B200/B300），英伟达在 2026 年 8 月 26 日电话会上确认，Vera Rubin 平台于 8 月上旬才开始小批量出货。因此 7-9 月的用户/用量激增主要是软件算法调优带来的：Tibo 在 7 月 13 日详细解释了 OpenAI 优化成本方面的举措，例如推理优化节省~10%算力直接返还给订阅用户;② GPT-5.6 Sol 部署后用于自我改进 →production GPU kernel 改进降低 推理成本约 20%；投机推理提升 token 效率>15%[9]；③ 9 月针对重度用户 cache hit rate、compaction、prefix cache、KV cache 管理进行优化，这部分用户的效率可能提升 3-4x。GitHub Issue #44884[10] 中的实测数据披露 96.6%的 input tokens 来自 cache，说明用户配额消耗高度依赖 cache 命中效率，而非新增 GPU 算力，预计缓存命中等措施为供应提

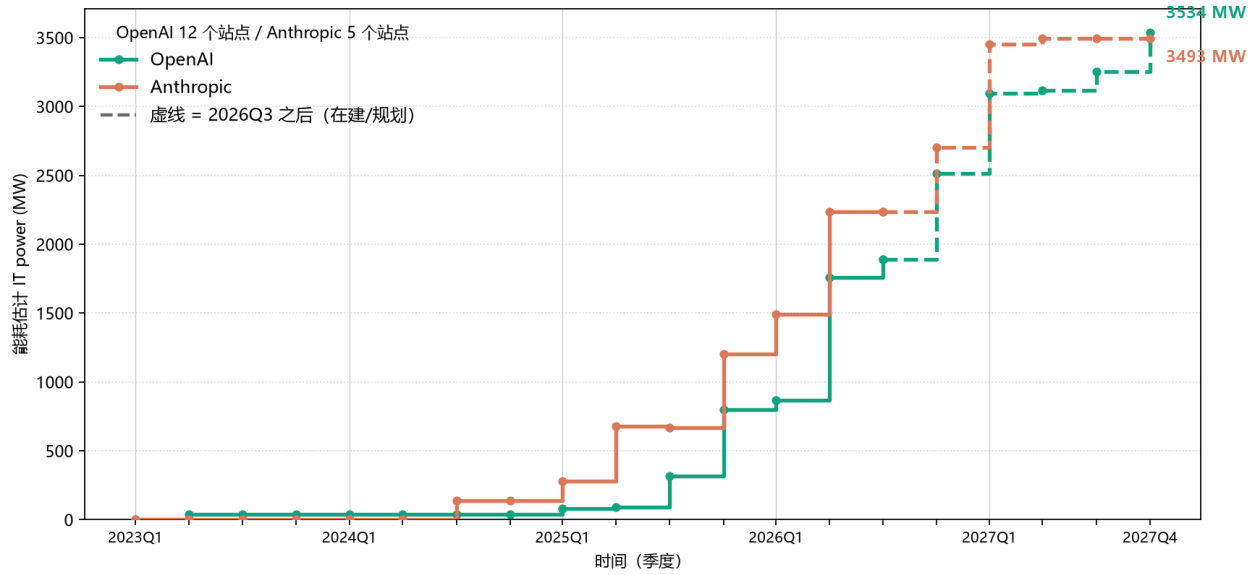
[9] https://openai.com/index/gpt-5-6-frontier-intelligence-efficiency/

[10] https://github.com/openai/codex/issues/44884

下载日志已记录，仅供内部参考

供弹性。

图 12:OpenAI/Anthropic AIDC 供给趋势（能耗估计）



数据来源：Epoch.AI，中信建投

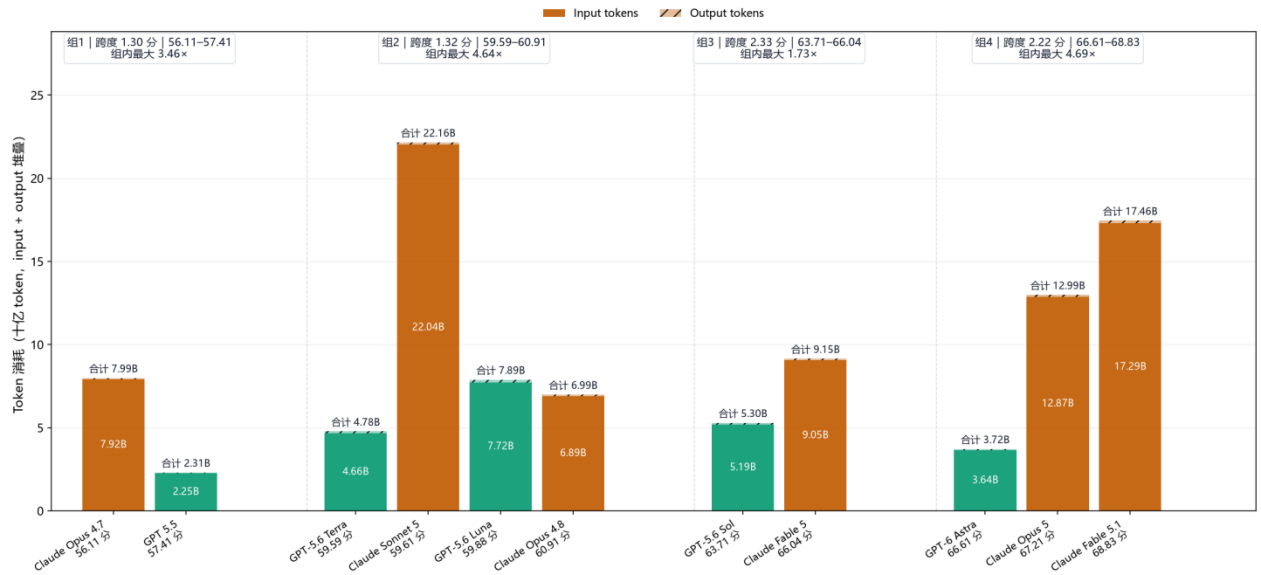
OpenAI 的案例反馈 ARR 增长放缓并非需求停滞，而是物理供应+商业模式制约导致的，OpenAI 的订阅收入相关占比仍然略少于 50%，26 年初也有 60%，这意味着收入增长更平滑，主要依靠订阅用户增长驱动（Plus/Pro 等套餐目录价格并未调整），而 Anthropic 的收入中 80%以上为企业/API 等，订阅占比少于 20%，因此 token/API 用量驱动整体收入波动率更高。

如果只考虑 API/企业订阅方面，OpenAI 年初至 8 月底的 ARR 增量约在 160 亿美元，而 Anthropic 则在 460 亿美元。OpenAI 订阅方面的 ARR 增量仅 40-50 亿美元，Anthropic 订阅 ARR 增量则为 100 亿美元。如果结合 The Information、Bloomberg 等披露的客户结构，Meta 可能贡献 Anthropic ARR 约 50 亿美元，中国区客户~2%的 API 占比（对应 20+亿美元 ARR，假设 token 消耗与其他客户平均水平一致），结合 Anthropic Marketplace 20%分成（对应 40-50 亿美元），剔除这些因素后，Anthropic/OpenAI 26 年增量 ARR 份额大致对应 2：1。

我们注意到 OpenAI/Anthropic 的模型虽然旗舰模型定价基本接近，但是实际完成任务/benchmark 消耗方面存在较大差异，例如 vals.ai 评测的将金融、编程、法律等 agent 任务，GPT 每一代旗舰模型 input/output token 都明显少于 Claude，我们选取发布时间接近/测评分数接近的“对标模型”，Claude 系列模型基本每一代的 token 消耗都高出 GPT，GPT-5.6 sol 算是差距最小的，Fable 5 仅多出~70%的 token 消耗。成本层面（考虑缓存命中等优化措施），Claude 模型基本高于对标 GPT 系列模型的 1x 以上，唯一例外是 Astra vs Opus 5 实际成本基本接近（但是在 Artificial Intelligence Index 中两个模型的性能存在较大差距），更全面地引入 AA index 也支持这一结论（尽管 ARR 增量份额是 2：1，但这并非说明用户 task 执行份额是 2：1，而是相似任务 Anthropic 系列模型的 token 消耗远大于 OpenAI 的 2x，导致实际成本一般也是 OpenAI 对标模型的 2x 左右），反过来讲，OpenAI 的用户增量/实际用户的需求可能略多于 Anthropic，但由于 token 效率较高，实际产生的收入较低。

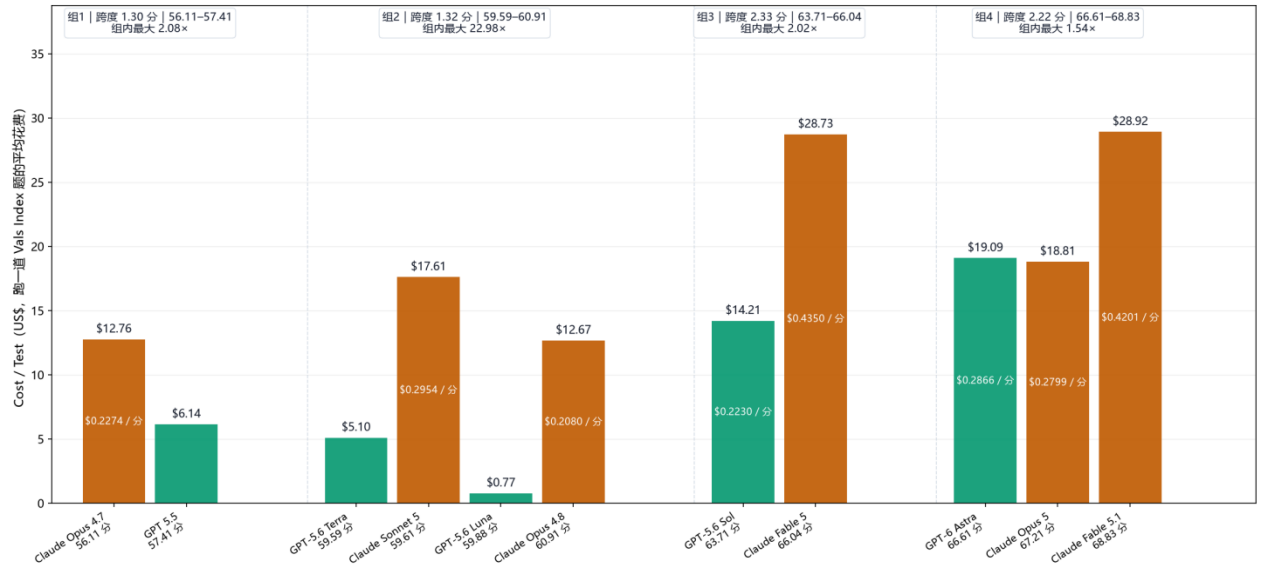
下载日志已记录，仅供内部参考

图 13:benchmark 性能接近的对标模型，消耗 token 比例



数据来源：Artificial Analysis，中信建投

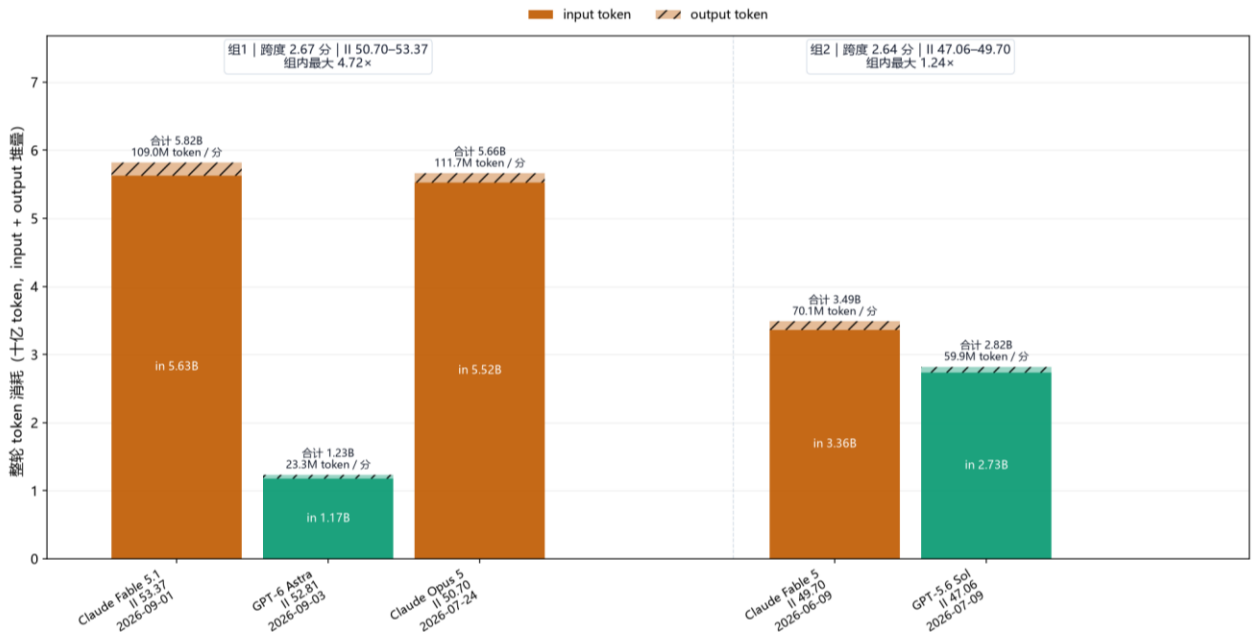
图 14:benchmark 性能接近的对标模型，消耗成本比例



数据来源：Artificial Analysis，中信建投证券

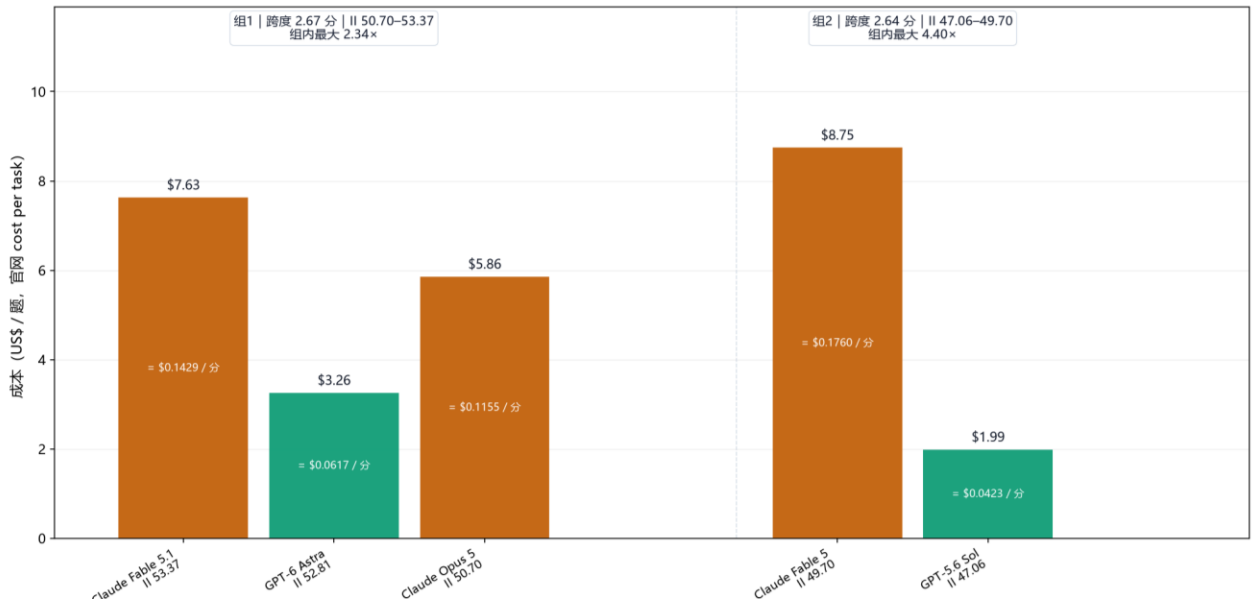
下载日志已记录，仅供内部参考

图 15:benchmark 性能接近的对标模型，消耗 token 比例



数据来源：Artificial Analysis，中信建投证券

图 16:benchmark 性能接近的对标模型，消耗成本比例



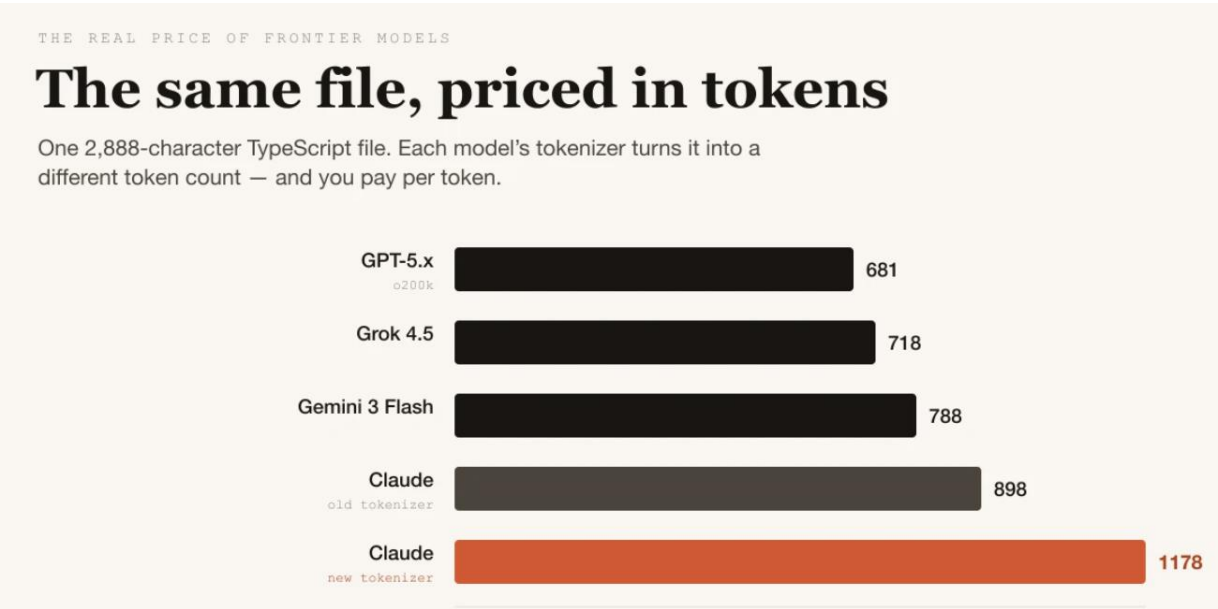
数据来源：Artificial Analysis，中信建投证券

为什么 Anthropic 系列模型的 token 消耗/成本总是高于同级别的 GPT 模型？首先是，Claude 4.7 系列模型调整了 tokenizer，导致相同任务下的 token 膨胀。不同分词器下 Fable 5.1 处理相同文件消耗的 token 可能达到

下载日志已记录，仅供内部参考

GPT-5.x o200k 分词器的~2x[11]。

图 17:Claude 系列模型分词器 token 消耗显著较高



数据来源：playcode[12]，中信建投证券

图 18:分词器导致 Claude 系列模型处理不同类型的任务 token 消耗均高于其他模型

Content	Claude (new)	Claude (old)	Gemini 3.x Flash	Grok 4.5
TypeScript	**1.73x**	1.32x	1.16x	1.05x
Rust	**1.58x**	1.22x	1.19x	1.05x
JavaScript	**1.52x**	1.26x	1.23x	1.11x
Python	**1.50x**	1.22x	1.20x	1.09x
HTML page	**1.36x**	1.18x	1.08x	1.04x
English prose	**1.40x**	1.05x	1.01x	1.00x
Chinese prose	**1.44x**	1.45x	0.85x	0.86x
Chinese chat	**1.53x**	1.55x	0.91x	0.92x

数据来源：Playcode，中信建投证券

[11] https://playcode.io/blog/real-price-of-frontier-models
[12] https://playcode.io/blog/real-price-of-frontier-models?via=dailydev

下载日志已记录，仅供内部参考

以 o200k 分词器处理以下内容的 token 消耗为基准，Claude 新分词器相比于旧分词器进一步提升的 token 膨胀效应，基本多消耗 40-50% token。AI 营销公司 Ploy 将部分生产系统中的 Claude 4.8 Opus 迁移至 GPT-5.6 sol13，在相同配置环境下的工作负载 token 消耗从 260 万下降至 170 万（-35%），考虑两个模型的单价差异，实际成本节约>40%。

其次是，Claude 系列模型的思考链路越来越长，即使剔除 tokenizer 的影响，Claude 4.7 系列后采用相同的 tokenizer，但 Fable 5→Opus 5→Sonnet 5→Fable 5.1 大体上思维链路和 token 消耗越来越长了，Fable 5.1 的输出 token 甚至是 Fable 5 的 2.3x（多出了 130%的 token）。下图是在 MacBook 绘图测试中，每个模型都被要求使用相同的 SVG 文件作为 Max effort 对应的输出 token。

图 19:Claude 系列模型的 token 消耗膨胀较快

Model	Output tokens, one drawing	vs Fable 5	Output ceiling
Claude Sonnet 5	121,713	2.9x	128,000
Claude Fable 5.1	99,072	**2.3x**	128,000
GPT-6 Astra	45,836	1.1x	128,000
Claude Opus 5	74,061	1.7x	128,000
Claude Fable 5	42,560	1.0x	128,000
GPT-5.6 Terra	29,834	0.7x	128,000
Gemini 3.7 Flash	13,863	0.3x	65,536

数据来源：Playcode[14]，中信建投

再次是 Tool Call/Function Calling 带来的消耗差距。AI 搜索引擎 You.com 26 年 5 月底[15]分享了对 Claude/GPT 模型用于搜索的工程博客，提到 Claude 在搜索查询中消耗的 token 数量是 GPT 的 4-5 倍（“我们提出了相同类型的问题，得到的答案质量也相差无几，但 token 使用量却相差甚远。”）。总体而言，成本差距缩小至约 3.3 倍，因为输出代币和固定的每次搜索费用削弱了输入代币之间的差异。但在搜索轮数较多的复杂查询中，累积的上下文信息会将成本比率推高至 5 倍以上。

[13] https://ploy.ai/blog/migrating-a-production-ai-agent-to-gpt-5-6
[14] https://playcode.io/blog/real-price-of-frontier-models
[15] https://www.linkedin.com/posts/youdotcom_we-ran-50-search-queries-through-claude-and-activity-7465774907902177282-92SM

下载日志已记录，仅供内部参考

图 20:Claude 在搜索查询中消耗的 token 数量是 GPT 的 4-5 倍

Tier	Claude Cost	GPT Cost	Cost Ratio	Claude Searches	GPT Searches
Simple	$0.08	$0.04	2.0x	1.3	1.4
Moderate	$0.12	$0.06	2.0x	1.6	1.7
Complex	$0.39	$0.13	2.9x	4.7	4.8
Deep Research	$0.84	$0.15	5.5x	8.4	5.0

数据来源：You.com，中信建投证券

需要注意这里对比的是 Claude 4.6 vs GPT-5.4 模型（不受 tokenizer 的影响），成本差距主要来自工具调用的效率，Claude 的架构在多轮迭代中将更多检索内容引入上下文。Anthropic 的搜索工具将搜索结果作为内容块返回到 API 响应中，这些内容块位于对话的消息数组中。当 Claude 决定需要再次搜索以回答同一查询时，下一次迭代会携带之前所有迭代积累的完整上下文，包括所有之前的搜索结果。GPT 的架构则将更多搜索处理工作保留在工具边界之后，GPT 模型消耗较少 token，但模型看到的是服务器端获取信息的精简版本。当然 Anthropic 后续[16]注意并且修复了搜索场景对于 HTML 内容的动态过滤。

Agent Labs[17]很对 Claude Code 和 Codex 等 Agent 框架的实证研究表明，虽然 Anthropic 和 OpenAI 都推出缓存命中，但在实际自主运行的 Agent 循环中，**实际上这些举措并不一定能带来成本节约，反而可能引入额外的成本。**团队横向评测了 6 款主流生产级 Agent 客户端/框架（包括 Claude Code CLI 2.1、Codex CLI 0.130），覆盖了 Anthropic 和 OpenAI 两条主流缓存链路。测试设计非常贴近真实工作流：连续进行 4 轮交互（每次附带约 5,186 字节对齐的 Token 负载），随后**强制引入 390 秒（6.5 分钟）的空闲停顿**（模拟人类写代码、思考或跑测试的真实间隙），再发起第 5 轮调用。390 秒恰好严格超过了 Anthropic 默认的 5 分钟（300 秒）TTL 缓存生存周期，但远低于其 1 小时（3600 秒）TTL，用以无死角检测各框架应对人类停顿的缓存保留能力。

表 2:Anthropic/OpenAI 的缓存体系对比

机制维度	Anthropic 缓存体系（以 Claude Code / AGNT 为代表）	OpenAI 缓存体系（以 Codex CLI 为代表）
客户端控制权	**显式控制（Explicit）：**需客户端手动打最多 4 个断点，显式声明 5m 或 1h TTL。	**隐式全自动（Implicit）：**超过 1,024 Tokens 的相同前缀自动享受折扣，无任何客户端标记。
写入成本	**有惩罚机制：**5m 写入收 1.25× 溢价，1h 写入收 2.0× 溢价。	**零写入费：**缓存写入完全免费，只对命中的读取打折。
容错率与门槛	**极低容错：**一旦客户端前缀发生 1 字节变动或停顿超期，不仅没省钱，反而因 1.25x/2.0x 写	**高容错但无保底：**客户端完全无法保障缓存留存时间（由服务器集群负载动态决定，多为数分钟级淘汰）。

[16] https://claude.com/blog/improved-web-search-with-dynamic-filtering

[17] https://agnt.gg/whitepapers/the-cache-wars-2

下载日志已记录，仅供内部参考

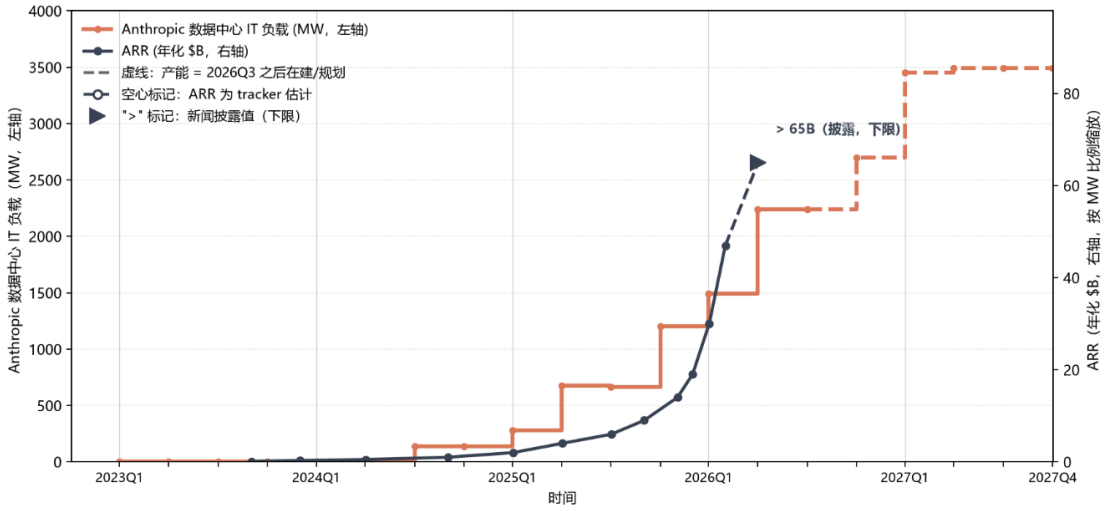
入溢价被双重惩罚。

资料来源：You.com，中信建投证券

评测发现，Anthropic 官方出品的 Claude Code CLI 在每轮交互中，会自动静默注入多达 6,817 个 Token 的系统级自生成脚手架（用于环境检测、工具声明与权限校验）。这导致即使在无网络分支的稳态下，开发者为脚手架支付的额外高溢价 Token 是实际工作载荷的 2.3 倍。

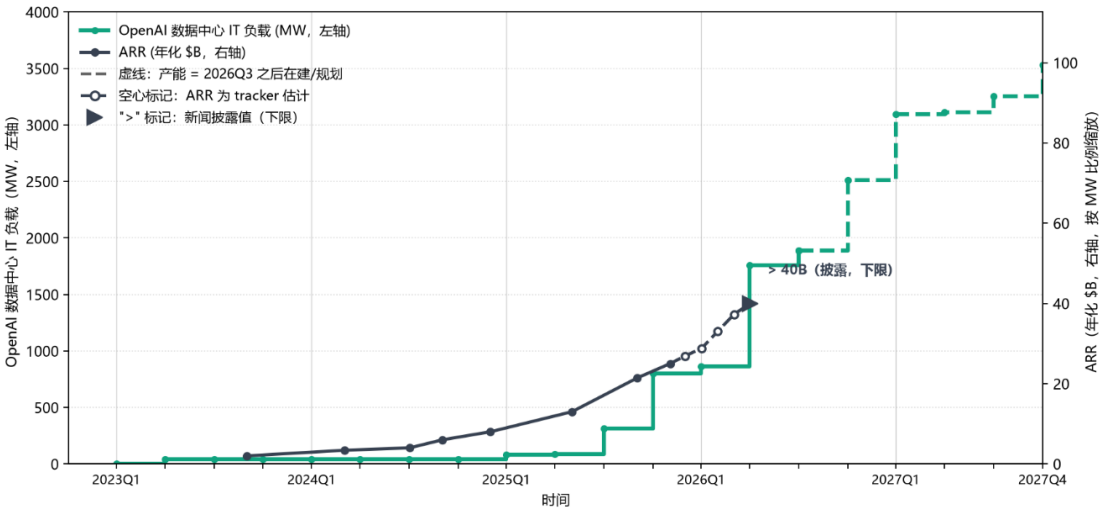
总而言之，2Q ARR 增长有所放缓更多是供应瓶颈+效率优化（纠正此前 token 膨胀的技术债），以及年初大量企业 FOMO 下采购 AI，2Q 后进入阶段性冷静期等因素共同导致的结果。

图 21:Anthropic ARR 与 IDC 供给趋势



数据来源：Epoch.AI，中信建投证券

图 22:OpenAI ARR 与 IDC 供给趋势



数据来源：Epoch.AI，中信建投证券

8 月 19 日，OpenAI Codex & ChatGPT 核心产品负责人 Tibo(Thibault Sottiaux) 直接回复 Anthropic 延长

下载日志已记录，仅供内部参考

Claude Code 限额但警告“capacity may be tight”的帖子：“Do not worry, we have compute”。 8 月下旬，Matthew Berman 访谈时 Tibo 强调 OpenAI 提前数年规划算力，两年前大规模投入曾被质疑，现在庆幸有储备。效率大幅提升（Luna 价格降 80%，整体速度比 3 个月前快约 60%），支持频繁 usage reset（“Tibo reset”成为社区梗）。模型已用于优化推理栈、CUDA 内核等（早期递归自我改进），大部分产能优先外部用户/企业。内部员工若无限制可“吞掉”所有生产 GPU。

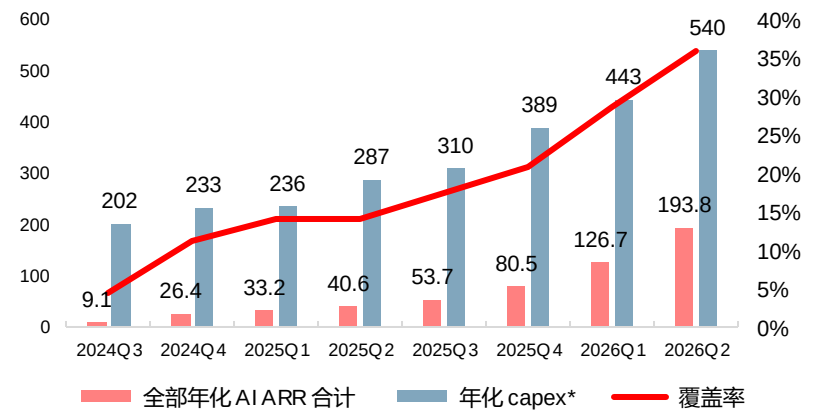
2026 年 9 月 GPT-6 Astra 发布后需求史无前例爆发，快速转紧。Tibo 公开：Astra 需求“truly unprecedented”，“I have never seen anything like this before”。公司可能暂停新 Pro 订阅（$200 档等）以保护现有用户质量，同时全力加产能。现有订阅正常续费。社区大量抱怨 capacity，Tibo 频繁 reset 但无法完全缓解。

Anthropic 2026 年 11 个月内签 $517B 协议（至少 14.8 GW 多年期），包括 Google/Broadcom（多 GW TPU，2027 起）、SpaceX Colossus（全 Colossus 1 + 部分 2，>220k-300k+ GPU / 300+ MW）、Lambda（$35B，~350 MW Texas）、Nscale（$45B，460 MW WV，Vera Rubin，late 2027）、Fluidstack、Amazon Trainium（up to 5 GW）、AMD、Volta、RUM 等。CFO：most significant commitment to keep pace with unprecedented growth。

行业长期订单改善，AI 收入/Capex 覆盖率快速提升

AI Run-Rate/Capex 或者累计折旧，比例都在稳步提升，尤其是当 AI 收入以倍数同比增长的情况下，毫无疑问 AI CapEx 的 ROI 正在兑现，投资者所谓对于 ROI 的担忧本质是 AI 收入增长放缓→线性外推 Terminal Value 恶化→DCF 估值收缩，这一过程伴随通胀对实际利率的挑战，高利率环境+上游供应链紧张，进一步放大了 AI 产业链的波动。

图 23:AI Labs ARR/CSP Capex 覆盖率

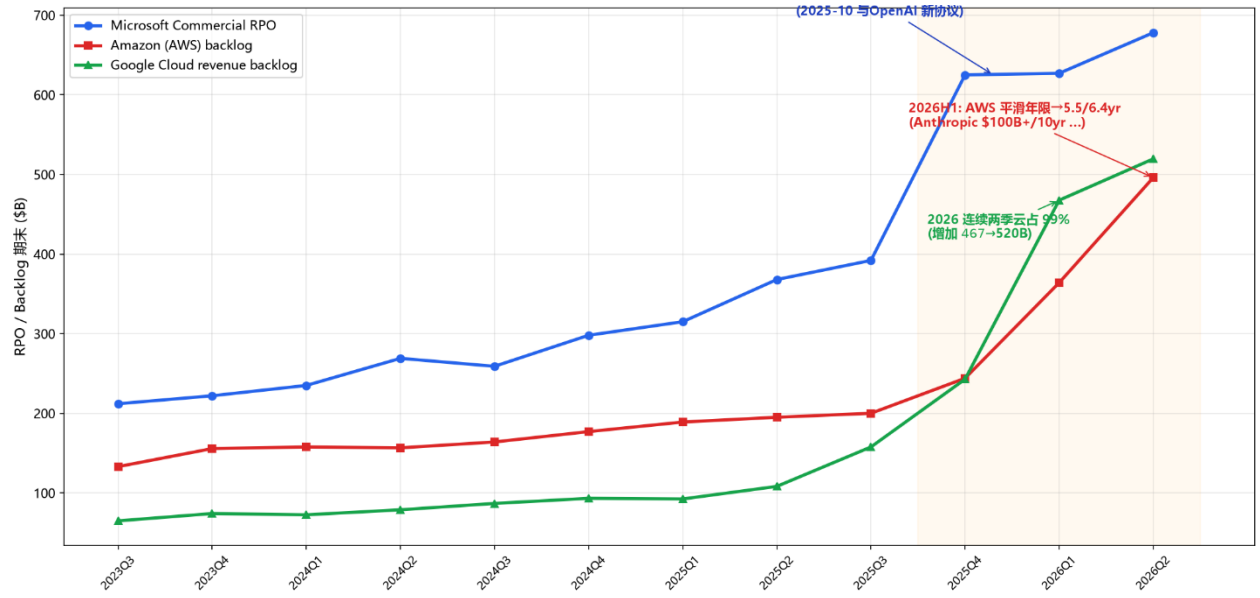


数据来源：微软/Google/AMZN 等财报，OpenAI/Anthropic，中信建投

RPO（待履约金额）是"未来收入可见度"的最佳代理。我们看到 4Q25 以来微软、Google、AWS 陆续收到 OpenAI/Anthropic 的大额长期合同，年化金额~100 亿美元，即使剔除这些 AI Labs 的大额合同，RPO 也持续加速增长，结合 The information/Bloomberg 等媒体披露的大合同订单平均期限都在 5-10 年，目前 AWS 长期合同贡献当期收入占比~46%，处于最近 3 年的峰值。

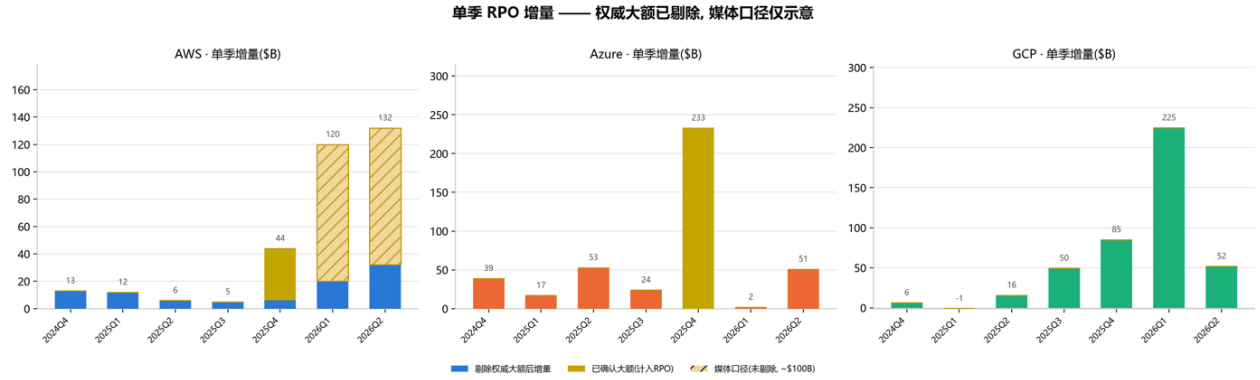
下载日志已记录，仅供内部参考

图 24:AZURE/AWS/GCP RPO/backlog 订单增长趋势



数据来源：公司公告，中信建投证券

图 25:即使剔除 AI Labs 订单，其他云订单的增长也好于长期趋势



数据来源：公司公告，中信建投证券

北美科技财报总结及前瞻

微软（Microsoft/MSFT）：Azure 43%提速破千亿，Copilot 加速转化带动叙事改善

业绩前市场普遍担忧 Azure 按量计费收入放缓及执行风险，但官方披露 Azure FY26Q4 增速达 43%（较 Q3 的 40% 进一步加速），打消降速顾虑，且宣布 Azure 全财年收入正式跨过 1000 亿美元大关。在手待履约订单（RPO）达到 $678B（+84% YoY），不仅反映 OpenAI 去年底签订的大额订单，更印证了 F500 强向 Azure

下载日志已记录，仅供内部参考

AI 迁移并锁定 3-5 年长期框架协议的真实落地。**核心亮点**是 Microsoft 365 Copilot 付费席位达到 3,000 万（较 Q3 的 2,000 万，环比+1,000 万，历史最高），拥有 50,000 席位以上的大企业客户数量同比增 7 倍，Purview 审计的 Copilot 企业交互次数单季突破 500 亿次。单季 CapEx 扩张至 $41.0B（+70% YoY），官方确认 Wisconsin（Mount Pleasant）Fairwater 超级数据中心数百 MW 算力集群于 2Q26 陆续并网，大幅缓解了上半年内部算力“紧平衡”的物理压力。

2H26 分歧：1）Bloomberg 一致预期反映的 FY27-28 RPO 低增长，即 OpenAI/Anthropic 大单外，不再有增长。1）然而，我们看到随着开源模型加速追赶，企业/主权政府在部署采用 AI 的意愿反而加强，带动非云的硬件需求/订单的提升。我们认为未来 2 年看到主权 AI 订单的机会并不低。2）即使不考虑 AI 带来的增长，目前市场对于 RPO 的增长预期也远低于传统 M365 Commercial/Non-AI Cloud 业务的长期增长率，似乎正在定价一种 AI 侵蚀企业预算，实质性侵蚀传统业务预算，未来 IT 预算冻结的情况，而这与企业 IT 支出趋势并不匹配。

2）Bloomberg 一致预期反映的 D&A/COGS 存在低估风险，虽然 FY26Q4 公司调整了 IDC 折旧年限（15 年→25 年），导致 ASC 432 准则下部分合同重新归类，原本计入折旧的转为当期租赁费用计入 COGS/R&D（大部分都是 COGS），同时冲减 FCF。目前市场预期调低了未来的 D&A 增长趋势，但是 COGS/FCF 也应该相应下修，我们认为目前市场下修的幅度和节奏没有体现这一财务变化。具体而言，市场预期在建工程几乎不增长，也就是即期转固，不太符合业务逻辑（一般没办法立刻通电，存在 3-4 季度的周期），因此 FY28 存在 COGS 上修/FCF 下修风险。

3）市场对于 OPEX 的预期偏于线性，实际上 FY24-26 公司的费用中包含了动视暴雪收购的一次性法律费用，多轮裁员的遣散福利费用，自研芯片 Maia 的流片，自研 MAI 模型的研发费用等，但是未来这些费用的趋势并不一定线性增长，尤其是芯片验证后量产规模效应下单片成本快速下降，以及前期收购/裁员费用也有下修空间。

总体而言，微软目前 Forward 25x P/E 低于历史交易区间中枢，一定程度上反映了投资者对于未来几年 EPS 增速放缓的担忧，也就是 D&A 上修，AI 投资的回报周期延长。然而，具体节奏上，我们看到 FY27 存在 RPO 订单超预期增长及转化的机会，且 FY27 OPEX 优化空间较大，因此 FY27 收入及 EPS 相较于预期上修的机会较大。

谷歌（Alphabet/GOOGL）：云持续加速，关注核心广告业务利润率

2Q26 Google Cloud 营收同比+82.0% 至 247.7 亿美元，Cloud 营业利润从去年同期的 28.3 亿美元增长至 88.1 亿美元，营业利润率增至 35.6%，单季新增云订单 Backlog 超过 500 亿美元，总额突破 5,140 亿美元。核心搜索广告（Search & other）实现 632.7 亿美元（+17.0%），YouTube 广告收入 110.6 亿美元（+13%），展现出强劲的防御性。资产负债表与现金流：单季 CapEx 翻倍至 449.2 亿美元（占总营收的 37.5%），直接导致季度自由现金流录得 -58.6 亿美元，自 2004 年上市 22 年以来首次出现单季负自由现金流。公司同时将 2026 全年 CapEx 指引上调至 $195B-$205B。

2H26 乃至 27-28 年的中期分歧：1）TAC 高估倾向：投资者仍然机械性地假设 TAC 受 Network 轻微负增长带动，TAC/收入比例下降，但这忽略了结构变化。TAC 中主要还是 Google 搜索的默认搜索协议带来的分成费用，例如与 Apple/Samsung/Firefox 等，2024 年与美国司法部的诉讼于 2025 年 9 月完成初审裁决，法官认定后续 Google 禁止进行溢价分成换取排他搜索引擎分发，并于 2025 年 12 月的最终判决（2026 年 2 月正式生效）中明确确立了“完全非排他”与“合同最长 1 年期”的强制性制度红线。过去 Google 向 Apple 支付高达 36%的

下载日志已记录，仅供内部参考

Safari 搜索广告净分成（年化逾 200 亿美元），本质是为了防御性封杀微软 Bing 等竞对并遏制苹果自研；但在法律严禁排他的新规下，竞对已无法通过竞价买断入口，Google 继续支付高昂排他溢价的合规与商业基础已不复存在，名义分成率有机会向公允的非排他渠道技术服务费（如 15%以内）。叠加 iOS 生态中约 35%的流量早已由免税的原生入口（Google Search App 独立端、主屏小组件等）承接，全 Apple 渠道的综合有效 TAC 率具备显著的下行弹性。

2）Youtube 等重资产购入体育赛事等内容版权，对于广告主家庭预算渗透率提升有帮助，目前 Youtube 的增长似乎更多反映线性趋势。

总体而言，Google 目前 Forward 21x P/E 处于历史交易区间中枢，一定程度上反映了投资者对于未来几年 EPS 增速放缓的担忧，以及 AI 投资的风险。然而，搜索广告业务的反转我们认为并未结束，2025 年 9 月的裁决消除了极端下行风险，但对于利润率提升也留出空间，不过过去 1 年市场聚焦于 Apple 几乎不受损，Google 避免尾部风险进行了“双赢”交易，但这更多是合同尚未到期+新一轮上诉周期导致的市场钝感，实际上我们看到 Apple/Samsung 等多个重要合作方集中于 26/27 年到期，根据目前裁决，最多签署 1 年合同且不得以其他形式间接分成以换取排他协议（例如通过 AI 助手绑定），我们认为这是一个有分量且未被充分定价的催化剂，叠加 Gemini 4 系列模型的回归，Google 的叙事有望重新改善。

Meta（META）：广告推荐算法变现效率持续验证，关注 C 端 AI 变现的边际 ROI

Q2 营业收入 608.0 亿美元，同比增长 28.0%，广告收入达到 594 亿美元（+27%），在极高基数下依旧维持强劲动能。营业利润实现 188.0 亿美元，同比下降 8%（OPM 降至 30.9%），主要由于计提了 24 亿美元的法务和解支出与 12 亿美元重组遣散费；调整后经营利润同比增长 9%。单季 CapEx 达 311.0 亿美元，将自由现金流大幅压缩至 7.84 亿美元（较去年同期的百亿美元级在 CapEx 挤压下显著收窄）。全年 CapEx 指引收$130B－$145B 的区间。

2H26 分歧：1）变现闭环高效，但缺乏外部风险对冲：Meta 将算力全量灌入 Llama 广告推荐引擎，广告曝光量与转化率持续上行。然而与微软、亚马逊、谷歌不同，Meta 完全没有外部企业级公有云客户的长期订单来为其兜底。2）尚无法看到 AI 回报的长期空间：一方面公司自研 MTIA 芯片渗透率仍处于初期，仍需大量向外部采购英伟达高价算力；另一方面，Meta 本身也是 Anthropic 等前沿实验室的大客户（贡献年化数十亿美元算力租用），基建资本开支正全面演化为报表端的硬性折旧支出。

市场对 Meta 的估值框架正在从“广告变现优等生”向“高固定成本重资产科技股”重估。若 2H26 宏观广告需求出现周期性波动，刚性折旧对利润率的侵蚀程度远大于其他三家 CSP。Muse 的长期留存率和广告、电商变现值得关注。

亚马逊（Amazon / AMZN）：AWS 37% 创 18 季度新高，自研芯片造就最强单位经济模型

2Q26 AWS 营收达到 422 亿美元，同比增长 36.7%，实现连续 5 个季度环比加速，创下过去 4 年半（18 个季度）以来的最快增速，年化 Run-rate 达 1,690 亿美元。AWS 营业利润录得 166 亿美元，营业利润率达

下载日志已记录，仅供内部参考

到 39.0%（即使剔除 6 亿美元能源衍生品公允价值变动收益，核心经营利润率仍高达 38.0%，同比大幅扩张 520 个基点）。管理层首次披露两大重量级里程碑：AWS AI 收入年化 Run-rate 突破 250 亿美元，且自研芯片（Trainium 与 Graviton）年化业务规模同样跨过 250 亿美元，两者均保持三位数以上的同比爆发。2026 全年 CapEx 指引进一步上调至 约 2,200 亿美元（此前为 2,000 亿美元）。

2H26 分歧：1）零售业务受供应链成本冲击，广告业务诉讼风险下，Non-Cloud 业务的利润率修复趋势能否持续。YTD 的涨跌幅与零售股呈现较高相关度；2）云业务方面，AWS 利润率能够顶住高投入扩张至 38%+，关键在于其芯片自研替代战略。Anthropic 签署的最高达 5 GW 的 Trainium 合作协议，使 AWS 在模型训练与批量推理中极大地规避了英伟达平台的高昂硬件溢价与供应链抽成。AWS Backlog 期末达 4,960 亿美元，且长期合同（平均年限 5.5–6.4 年）贡献当期收入比例达到 46%，CEO Andy Jassy 所称“2026 年算力配额全部售罄，客户正在锁定 2028 年产能”，呈现显著的供不应求格局。AWS 的 AI 订单能否带动 non-AI 负载的消耗，对应 AWS 的长期竞争力能否回归。

2H26 前瞻与投资论点：1）核心优势：北美零售经营利润率提升至 7.9%，国际业务持续转正盈利，电商与履约效率释放的充沛现金流能够充分覆盖并支持 2,200 亿美元的年化 CapEx，完全无需稀释股权或举债。

软件层：分化与 AI 转型仍在早期阶段

1. 系统记录层（System of Record）的不可替代性。LLM 是“概率型推理引擎”，无法凭空绕过企业账务、人事、订单与客户权限底座。真正的壁垒在于专有业务图谱 + 合规审计链条 + 跨系统编排权限。AI Agent 执行动作时，必须经由 SoR/Action 系统做权限校验、审计与数据落地，否则无法生产级部署，这强化了 NOW、SAP、WDAY 等的护城河，而非颠覆。

2. 商业模式大迁徙（Seat-based → Usage/Outcome-based 混合）。头部 SaaS 正积极加速商业模型换轨，而是推动“席位费保底 + 智能体交互/成果抽成（Credit / Consumption / Per-Resolution）”混合架构。NOW 通过 Pro Plus/Enterprise Plus 席位溢价（约 30-60% ASP uplift，叠加 Assist credits）最顺畅；CRM 用 Agentforce 探索 $2/conversation 或$2/resolution + Flex Credits；SNOW/MDB 纯用量；安全厂商平台化捆绑。这降低纯席位逆风，同时捕捉 AI 用量爆发。

3. IT 预算的“冷思考与回流”。上半年 FOMO 式大模型 API 采购后，进入消化/降本期（纠正 Token 膨胀与幻觉）。企业发现无高质私域数据清洗 + 工作流编排，前沿模型难落地生产。预算回流至现代数据栈（Data Infrastructure，SNOW/MDB）与业务工作流枢纽（System of Action，NOW 等），以及刚性安全。AI 成为现有平台的加速器，而非替代。

具体而言，ServiceNow (NOW) 定位企业级 AI 的最佳“行动系统”，Pro Plus 溢价变现最顺畅。极少数无需颠覆自身商业模式即可无缝吞吐 AI 红利的标的，定位“AI Control Tower / 企业 AI 路由器”。业绩核实（Q2 2026）：订阅收入$3.877B，+24.5% YoY（23% CC），符合/略超 22.5%-23%区间；cRPO $13.20B，+21% YoY（21.5% CC）；RPO $29B +21-22% CC。续约率约 98%。ServiceNow AI ACV 已超$1B（历史最快爬坡模块之一），agentic 部署 9 个月增 9 倍。FCF margin 指引强劲（FY 约 35%）。

商业化核心：席位提价至 Pro Plus/Enterprise Plus，内嵌代码生成、IT 服务台自动分流、HR 自动化等，客户愿意支付显著 ASP 溢价（市场情报约 30-60% uplift，叠加 Assist credits/conversation 计量）。避开纯消耗波动。

下载日志已记录，仅供内部参考

预期差与 2H26/前瞻：企业自研 Agent 难绕过其跨系统 API 编排器（权限/审计必需）。碎片化 IT 架构使其成为“企业 AI 路由器”。跟踪 Pro Plus 在非 IT（CSM、HR）渗透；AI ACV 目标继续超预期。有效对冲传统席位收缩的负面冲击。

SAP (SAP) 2027 停服截至期倒计时，Clean Core + Joule 构筑最强护城河。欧洲领头羊，业绩可见度与防御属性最高。业绩核实（Q2 2026）：云业务+22%（+24% CC），Cloud ERP Suite +25%（+27% CC），接近/略低于原稿 30%+高增区间（前期更高）；Current Cloud Backlog (CCB) €22.9B，+27%（+26% CC），稳定在 26-28%区间。FCF 进入强劲释放。RISE with SAP 迁徙确定性高。

RISE/2027 确定性：ECC 主流支持 2027 年底终止（可延长至 2030 付费），超大规模跨国企业必须加速 S/4HANA Cloud 迁徙，合同周期常 5 年+，免疫短期宏观。Clean Core 标准化是运行商业 AI（Joule）的前提。

预期差与前瞻：市场低估 Clean Core + Joule 深度结合。锁定全球 2000 强核心经营事实数据库。催化剂：欧洲主权云需求、Cloud 毛利率向 75%迈进、2027 倒计时加速迁徙。2H26/2027 总营收有望加速。

Salesforce 全力推进 Agentforce 的商业化验证，新旧计费交替期阵痛缓解中。Q2 FY27 总营收约$11.3B +11% YoY/CC（含 Informatica 贡献，有机约改善至高单位数-低双位数）；cRPO 加速至约+14% CC（高于原稿~10%）。非 GAAP op margin 33%+甚至更高，伴随大规模回购。核心 Sales/Service Cloud 仍受席位/人头逆风（外包/初级团队收紧）。

Agentforce：ARR 超$1.5B（+240%+ YoY，含相关），Data 360 等合计接近$3.9B。定价：$2/conversation 或$2/resolution（Help Agent outcome-based），叠加 Flex Credits（约$0.10/action）。试图从“辅助人类”跃升为“数字员工”。若 Agent 效率极高可能反噬基础席位；但新模式规模化前，老引擎失速压制倍数。2H 跟踪 Agentforce 商业化验证与有机加速。多空焦点仍在转型执行。

Workday 受益于 3-5 年不可撤销长约支撑，AI 贡献仍处于早期信号。业绩核实：订阅收入约+14%（近期 Q 约 13.9%，FY 约 14.5%），接近 15-16%区间；12 个月/cRPO backlog 约+14-16%；op margin 向 25-30%+扩张。合同多为 3-5 年不可撤销，按天/月直线确认，报表韧性强。Net New Seats 受大型企业白领招聘放缓/科技裁员余波拖累。Illuminate AI 等以捆绑为主，缺乏 NOW 式显著溢价。2H 维持稳健但缺 Alpha。AI 平台强化人才调度/薪酬，但商业化渐进。

Snowflake 走出 Iceberg 冲击，企业级私域数据“模型化”核心受益者（动能加速）。Q2 FY27 产品收入$1.49B +37% YoY（加速超原稿 28-30%）；NRR 126%；RPO $9B +30%。客户>$1M 828 家 +27%。Iceberg 存储冲击已被计算消耗对冲；管理层更迭出清。Cortex AI 等原位运行（Zero Data Movement）爆发，银行/医药/制造将敏感数据留在 VPC 内微调/检索，Credit 消耗加速。2H/FY 指引大幅上修（产品收入约+36%）。AI 从概念到真实用量。

MongoDB 向量搜索渗透迅速，但宏观/消费波动敏感。总营收增速约 20-27%区间（近期更强时达高 20s）；Atlas +27-29%左右，占比突破 70-73%。开发者心智稳固，Atlas Vector Search 成 AI 原生应用（Agent、代码辅助）上下文存储首选。长期 GenAI 卡位优，但中小/初创预算周期与大企业立项拉长导致波动高于合同锁定型 SaaS。2H 宏观不确定性下，收入确认波动显著。

CrowdStrike 危机后验证极强客户粘性，平台整合领头羊。ARR 向$5.5B+迈进（+24-25%+ YoY，净新增强

下载日志已记录，仅供内部参考

劲，早期更高增速已基数化）；FCF margin 26-30%+。24 年 7 月 Outage 事件后客户净流失率维持极低水位，验证粘性。Falcon Flex 灵活承诺模式成功，客户将二线预算整合，模块数持续上升。GenAI 打破攻防平衡（自动化 Agent、多模态钓鱼、自适应恶意代码），企业无法靠加人应对。Charlotte AI 全自动分析/修复成“生存必需”。2H 安全预算整合潮深化，继续抢份额。

Palo Alto Networks 平台战略取得积极进展，短期挤压出清后加速。FY26Q4：NGS ARR $9.1B +63% YoY（含并购，有机约高 20s-30%）；总营收 FY +24%，Q4 +34%；非 GAAP op margin 约 27-29%；Adj. FCF margin 约 38%。平台化客户/Net New Platformizations 加速（远超早期 1100 家量级，目标 FY30 4000+驱动$20B NGS ARR）。前期“免费过渡期”换取 3-5 年多产品大单的阵痛已过，2H/FY27 正是免租期结束、全额高毛利结转窗口，RPO 与现金流二次加速。平台化 NRR 高（>120%）。

投资评价和建议

进入 2H26，投资者的关注从“算力军备竞赛想象力”全面转向“订单可见度、自研芯片单位成本优势与现金流自洽度”。在大型科技巨头方面，各标的因基建路径与商业模式差异呈现清晰的分化特征：亚马逊（AMZN）依托较长的 Backlog 锁定期、自研 Trainium 闭环以及零售业务充沛的现金流支撑，展现出较强的单位经济学与抗风险优势；谷歌（GOOGL）凭借全栈 TPU 构筑的算力成本壁垒以及搜索 TAC 结构性下行的潜在契机，具备稳固的利润安全垫；微软（MSFT）后续的关键变量在于 Copilot 及其他 AI 应用的推进商业化节奏；Meta（META）在缺乏外部企业级云订单对冲的背景下，高额 CapEx 带来的刚性折旧对利润率的潜在侵蚀仍需持续观察。在软件板块方面，行业正由前期的估值收缩转向业务实质分化：ServiceNow（NOW）依托 Pro Plus 提价通道与企业 AI 编排中枢地位展现出较强韧性，CrowdStrike（CRWD）受益于自动化威胁升级下的安全预算平台化整合，SAP 享有 2027 年 ECC 支持终止带来的长周期刚性迁徙红利，而 Snowflake（SNOW）则逐步体现出企业私域数据原位计算带来的用量回升。整体而言，板块后续的分化主线主要围绕长期订单锁定度、自研硬件单位经济学以及资本支出与现金流的自洽度展开。维持“强于大市”评级。

下载日志已记录，仅供内部参考

风险分析

宏观经济与企业 IT 支出波动风险：全球宏观经济增长放缓与高利率环境可能抑制企业整体数字化支出意愿，导致企业级大单的审批决策链条明显拉长。同时，白领雇员招聘的持续平稳可能制约传统 SaaS 厂商席位净增的弹性，并对电商零售及数字化广告需求带来周期性压制。

行业监管与反垄断司法审查风险：反垄断司法审查及合规红线正在深刻重塑产业分发格局，例如针对默认搜索排他协议的强制性裁决落地，可能重构核心分发入口的商业合作条款与利益分成机制。此外，针对数据安全隐私、跨境数据流动以及前沿模型技术标准的监管趋严，也可能带来业务合规成本上升与估值承压的风险。

物理电网负荷与先进硬件交付延宕风险：超大规模数据中心扩建高度受制于区域电力基础设施与并网审批进度，能源供应紧张与变压器等长周期设备的交付摩擦可能导致算力集群上电投产晚于预期。同时，新一代计算平台（如 Vera Rubin 系列及自研 ASIC）的大规模量产、先进封装产能与机柜散热调优若发生延期，将直接制约下游有效算力供给的释放。

客户结构集中度偏高与商业模式换轨执行风险：前沿实验室企业端营收展现出极高的头部集中特征（Top 1%客户贡献绝大部分支出），单一大客户的采购调整或自研替代易引发高频收入指标剧烈波动。与此同时，传统软件厂商加速由“纯席位制”向“用量/智能体成果抽成”混合模式转型，若新定价策略面临客户接受度阻力或反噬存量基础席位，将给收入兑现带来执行层面的不确定性。

下载日志已记录，仅供内部参考

分析师介绍

于伯韬

海外研究首席分析师，FRM，香港大学金融学硕士，2020 年加入中信建投海外研究团队，2018-2020 年就职于长江证券研究所海外团队，2025/2024/2021/2020 年新财富港股及海外方向第五名，2022 年新浪金麒麟港股及海外市场最佳分析师第三名，2020 年新浪金麒麟港股及海外市场新锐分析师第一名。

许悦

海外研究员，南洋理工大学硕士，专注于港股互联网及美股软件研究，2022 年加入中信建投海外前瞻组，2023 年新浪金麒麟港股及海外市场菁英分析师第二名，2023-24 第十七届、十八届水晶球最佳分析师海外行业入围，2024-25 年度证券时报·新财富杂志海外市场研究团队第五名。

下载日志已记录，仅供内部参考

评级说明

投资评级标准		评级	说明
报告中投资建议涉及的评级标准为报告发布日后 6 个月内的相对市场表现，也即报告发布日后的 6 个月内公司股价（或行业指数）相对同期相关证券市场代表性指数的涨跌幅作为基准。A 股市场以沪深 300 指数作为基准；新三板市场以三板成指为基准；香港市场以恒生指数作为基准；美国市场以标普 500 指数为基准。	股票评级	买入	相对涨幅 15％以上
		增持	相对涨幅 5%—15%
		中性	相对涨幅-5%—5%之间
		减持	相对跌幅 5%—15%
		卖出	相对跌幅 15％以上
	行业评级	强于大市	相对涨幅 10%以上
		中性	相对涨幅-10-10%之间
		弱于大市	相对跌幅 10%以上

分析师声明

本报告署名分析师在此声明：（i）以勤勉的职业态度、专业审慎的研究方法，使用合法合规的信息，独立、客观地出具本报告，结论不受任何第三方的授意或影响。（ii）本人不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接收到任何形式的补偿。

法律主体说明

本报告由中信建投证券股份有限公司及/或其附属机构（以下合称“中信建投”）制作，由中信建投证券股份有限公司在中华人民共和国（仅为本报告目的，不包括香港、澳门、台湾）提供。中信建投证券股份有限公司具有中国证监会许可的投资咨询业务资格，本报告署名分析师所持中国证券业协会授予的证券投资咨询执业资格证书编号已披露在报告首页。

在遵守适用的法律法规情况下，本报告亦可能由中信建投（国际）证券有限公司在香港提供。本报告作者所持香港证监会牌照的中央编号已披露在报告首页。

一般性声明

本报告由中信建投制作。发送本报告不构成任何合同或承诺的基础，不因接收者收到本报告而视其为中信建投客户。

本报告的信息均来源于中信建投认为可靠的公开资料，但中信建投对这些信息的准确性及完整性不作任何保证。本报告所载观点、评估和预测仅反映本报告出具日该分析师的判断，该等观点、评估和预测可能在不发出通知的情况下有所变更，亦有可能因使用不同假设和标准或者采用不同分析方法而与中信建投其他部门、人员口头或书面表达的意见不同或相反。本报告所引证券或其他金融工具的过往业绩不代表其未来表现。报告中所含任何具有预测性质的内容皆基于相应的假设条件，而任何假设条件都可能随时发生变化并影响实际投资收益。中信建投不承诺、不保证本报告所含具有预测性质的内容必然得以实现。

本报告内容的全部或部分均不构成投资建议。本报告所包含的观点、建议并未考虑报告接收人在财务状况、投资目的、风险偏好等方面的具体情况，报告接收者应当独立评估本报告所含信息，基于自身投资目标、需求、市场机会、风险及其他因素自主做出决策并自行承担投资风险。中信建投建议所有投资者应就任何潜在投资向其税务、会计或法律顾问咨询。不论报告接收者是否根据本报告做出投资决策，中信建投都不对该等投资决策提供任何形式的担保，亦不以任何形式分享投资收益或者分担投资损失。中信建投不对使用本报告所产生的任何直接或间接损失承担责任。

在法律法规及监管规定允许的范围内，中信建投可能持有并交易本报告中所提公司的股份或其他财产权益，也可能在过去 12 个月、目前或者将来为本报告中所提公司提供或者争取为其提供投资银行、做市交易、财务顾问或其他金融服务。本报告内容真实、准确、完整地反映了署名分析师的观点，分析师的薪酬无论过去、现在或未来都不会直接或间接与其所撰写报告中的具体观点相联系，分析师亦不会因撰写本报告而获取不当利益。

本报告为中信建投所有。未经中信建投事先书面许可，任何机构和/或个人不得以任何形式转发、翻版、复制、发布或引用本报告全部或部分内容，亦不得从未经中信建投书面授权的任何机构、个人或其运营的媒体平台接收、翻版、复制或引用本报告全部或部分内容。版权所有，违者必究。

中信建投证券研究发展部

北京
朝阳区景辉街 16 号院 1 号楼 18 层
电话：（8610） 56135088
联系人：李祉瑶
邮箱：lizhiyao@csc.com.cn

上海
上海浦东新区浦东南路 528 号南塔 2103 室
电话：（8621） 6882-1600
联系人：翁起帆
邮箱：wengqifan@csc.com.cn

深圳
福田区福中三路与鹏程一路交汇处广电金融中心 35 楼
电话：（86755）8252-1369
联系人：曹莹
邮箱：caoying@csc.com.cn

信建投（国际）
香港
中环交易广场 2 期 18 楼
电话：（852）3465-5600
联系人：刘泓麟
邮箱：charleneliu@csci.hk