2027年AI展望：双寡头格局定调ARR稳健放量，Agent-Infra成破局迭代主线

行业研究 · 行业专题

互联网 · 互联网II

投资评级：优于大市（维持）

证券分析师：张伦可
0755-81982651
zhanglunke@guosen.com.cn
S0980521120004

证券分析师：张昊晨
zhanghaochen1@guosen.com.cn
S0980525010001

请务必阅读正文之后的免责声明及其项下所有内容

报告摘要

- **写在 Anthropic（ANT）上市之际：**ANT 的挂牌上市标志着大模型行业迎来首个里程碑，“模型军备竞赛”告一段落，前沿智能的牌桌进一步缩圈。**长期来看，我们认为头部基础模型具备极其优异的商业模式与护城河，核心逻辑体现在两方面：第一，高算力与重工程构筑了极深的竞争壁垒。** 当前前沿模型的训练已演变为高度精密的“工程式流水线”，根据ANT招股材料披露，公司已签署算力及基建承诺超过9,000亿美元，并预计在未来12个月支付超过700亿美元的算力费用，其中我们预计大部分仍将投入于模型训练。近千亿美元的算力投入门槛，意味着后排跟随者已较难实现弯道超车。**第二，顶级智能具备市场支配地位，且衍生出极佳的单位经济效益（Unit Economics）。** 2026 年的产业趋势清晰印证了 Frontier Models 的市场份额持续向头部集中；同时，头部厂商可依托最强基座模型向下蒸馏、训练中低档模型，在几乎不增加额外研发边际成本的前提下覆盖全价格带需求，叠加软件级分发近乎为零的边际成本，模型层远期将展现出极强的边际利润率与盈利弹性。
- **复盘 2026 年，模型商业化呈现清晰的阶梯式破局：** Coding（代码生成）率先打开了 Anthropic 发展历程中首个 10 倍量级的 ARR 增长空间；随后，网络安全在上半年顺利跨越商业化可用性拐点；展望未来，Knowledge Work（通用知识工作）或成为驱动下一轮 ARR 爆发的核心主战场。但截至目前，该领域尚未迎来真正的商业化奇点，核心掣肘集中在三大痛点：① 企业私有与行业垂直数据沉淀不足；② 支撑长周期运行的 Agent-Infra 尚未成熟；③ 企业既有业务流程（Workflow）的重构仍需时间磨合。

请务必阅读正文之后的免责声明及其项下所有内容

报告摘要

- **展望 2027 年，我们判断 ANT 与 OpenAI 将步入ARR 稳健放量阶段，算力资源将继续向训练端倾斜，而推理侧的智能释放仍受制于基建短板：**
  - 训练侧（RSI加速拉大代际差）： 我们预计 RSI（递归自我改进）将进一步提速，成为驱动前沿实验室（Frontier Labs）持续加码训练算力的核心引擎——Anthropic 2026 年 Q3 内部研发使用的 Token 消耗量同比激增 45 倍、环比亦实现 4 倍跃升，便是明证。RSI 飞轮的高速运转，将在 2027 年进一步拉大头部厂商与后排追赶者之间的智能代差。
  - 推理与落地侧（基建补课正当时）：顶级智能向千行百业的快速渗透，正面临类似“早期微软操作系统/应用生态与 Intel顶级CPU 算力演进节奏错配”的产业瓶颈——“算力内核（模型智商）”已足够强大，但缺乏成熟的“系统总线与运行环境（Agent-Infra）”将其转化为稳定生产力。
- **本篇报告通过复盘 2026 年大模型商业化演进脉络，前瞻 2027 年技术趋势对二级市场投资的映射启示，并重点梳理当前头部模型厂与云巨头关于 Agent-Infra 的架构布局。**我们判断，2027 年将是 Agent-Infra 快速建设与积淀的关键之年，其底层架构的成熟将成为引爆下一轮模型 ARR 非线性增长的重要基石。建议重点关注 Agent-Infra 的建设节奏及产业链相关核心标的的投资机会。

请务必阅读正文之后的免责声明及其项下所有内容

目录

01 头部模型领先优势显著，Agent Infra 有望加速企业 AI 拐点到来

1.1 复盘 2026：Coding 率先引爆 ARR，推理降本推动份额向双寡头集中

1.2 展望 2027（训练侧）：RSI 开启研发自迭代闭环，拉大前沿智能代际差

1.3 展望 2027（应用侧）：Token 降价推动 AI 从编程向垂直知识工作与个人 Agent 渗透

1.4 破局核心（基建侧）：以云为鉴，Agent Infra 补齐生产短板，巨头与独角兽加速布局

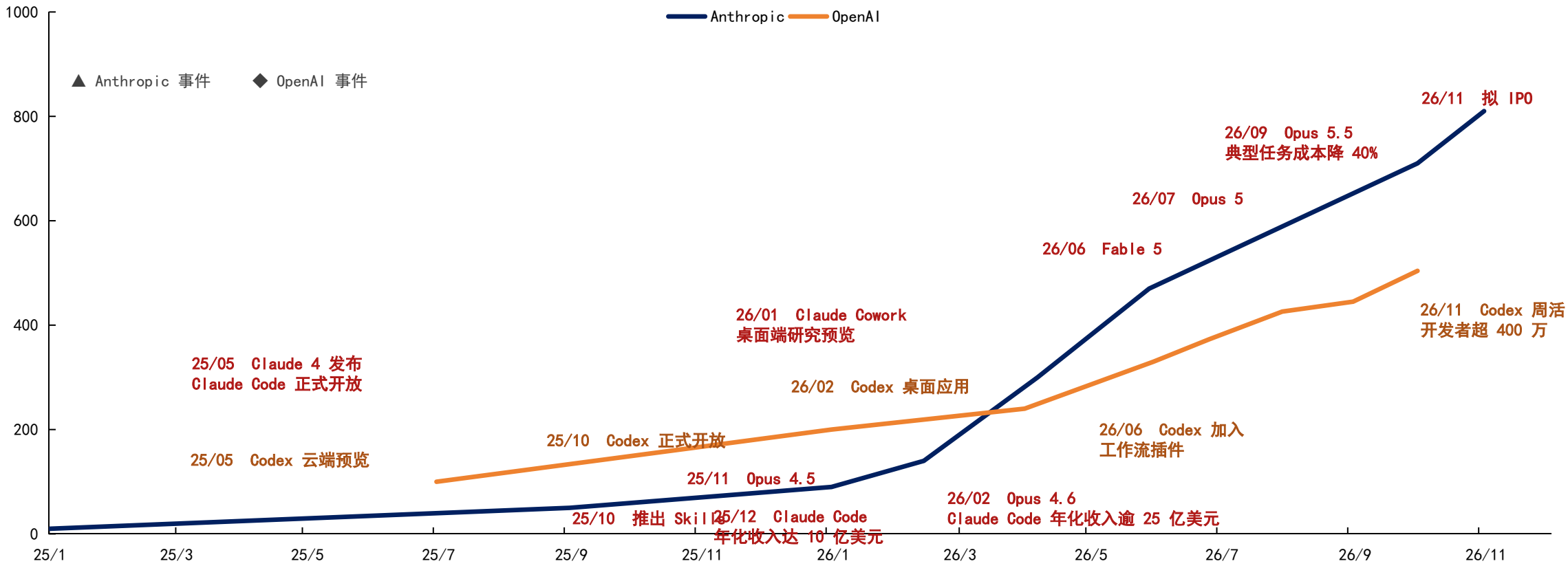
02 Anthropic：SOTA 模型构建商业化壁垒，核心收入来自企业客户

03 Anthropic：模型及产品矩阵丰富，覆盖全场景用户需求

复盘 2026 ：Coding 场景带动大模型 ARR 爆发式增长

- **Coding 场景带动大模型 ARR 爆发式增长。**编程 Agent 接手读仓库、改代码、跑测试和修复，每项任务需要反复调用模型，部署规模扩大后带动付费用量增长。到今年秋季，Anthropic 和 OpenAI 整体 ARR 已分别达到去年末的约 9 倍和 2.5 倍，截至26年10月，Anthropic ARR突破810亿美元。

图：Anthropic 与 OpenAI ARR 及关键事件（亿美元）



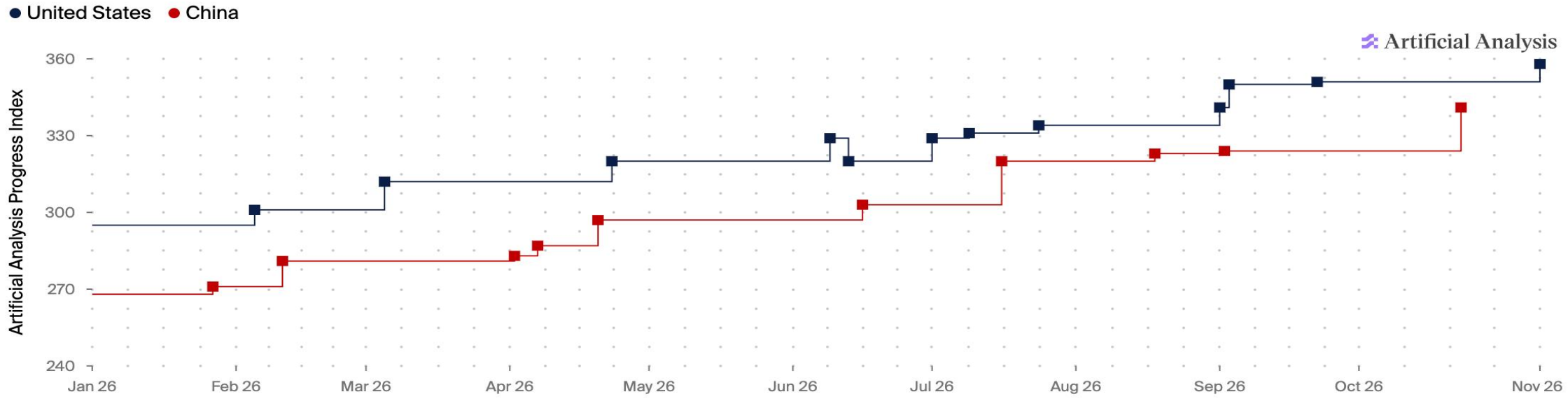
数据来源：Anthropic、SEC、OpenAI、TickerTrends，国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

复盘2026：国产模型年中追近，9月中美前沿差距再度扩大

- **6—7月中美模型差距收窄，9月反蒸馏力度加大后差距再度扩大。**AA进展指数显示，上半年，国内模型持续追近与海外闭源模型差距，但随着海外反蒸馏力度的加大，9月美国前沿模型进一步重新拉开跟国产模型的差距。
- **美国前沿模型公司目前主要设置了三层蒸馏防线：**1）利用专门用于检测对抗性数据提取的分类器，让模型内生地拒绝可疑请求；（2）在架构设计上隐藏或压缩思维链再输出；（3）用外部检测器识别数据提取，随后中断请求或封禁账号。
- OpenAI覆盖了蒸馏全生命周期的多层防御，包括在模型层隐藏原始 Chain-of-Thought，并专门训练模型避免泄露 reasoning traces等。Anthropic则在流量识别之外，更进一步将反蒸馏机制直接嵌入模型和 API 架构。 Fable 5开始引入专门的 distillation classifier，Fable 5.1 又引入 Preserved Thinking。

图：中美前沿模型能力对比（AA Progress Index v1，2026年1—9月）



数据来源：Artificial Analysis，国信证券经济研究所整理

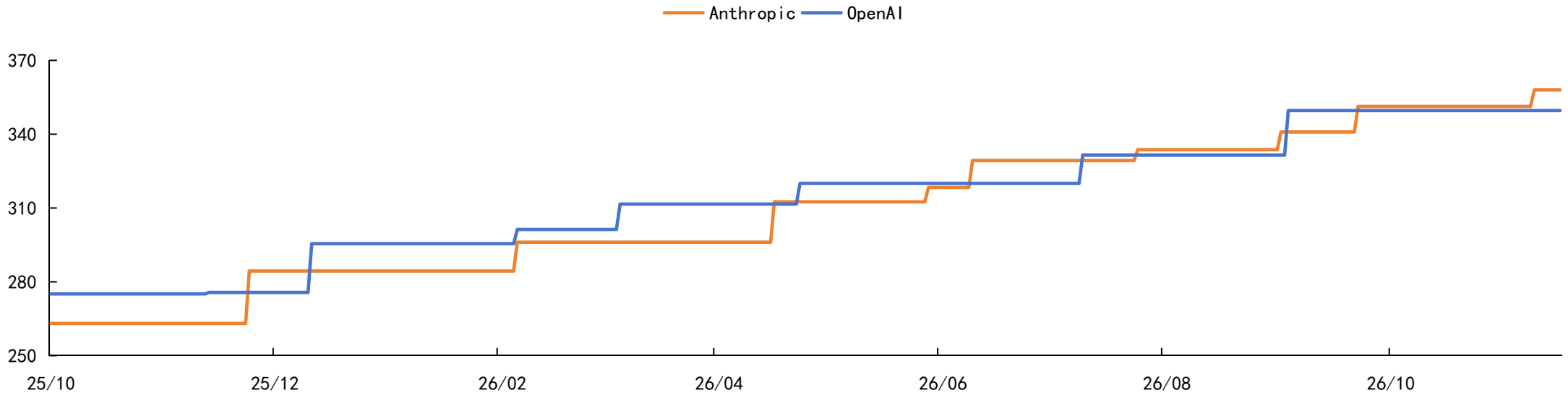
请务必阅读正文之后的免责声明及其项下所有内容

复盘2026：ANT与OAI错位竞争，预计走向寡头格局

- **两家模型多次交替领先。**9月初 GPT-6 Astra 超过 Claude Fable 5.1，9月下旬 Opus 5.5 反超。11月 Fable 5.2 将 Anthropic 最高分推至358，OpenAI 的 Astra 为350。
- **OpenAI 的投入范围更广，当前报表承受更大压力。**OpenAI 在模型和应用之外还投向芯片、消费硬件和机器人，较强的融资能力使其能同时推进多条业务线。Anthropic 的投入更集中在模型及企业应用。
- **如果模型能力难以拉开差距，Anthropic 很难长期一家独大。**若 OpenAI 收缩非核心投入，同时稳住模型能力和客户用量，利润率有望改善，并将更多经营回报投入下一轮研发。

图：OpenAI 与 Anthropic 历史最强模型评分（AA Progress Index v1）

按模型发布日期更新各公司已发布且获评分模型的历史最高分。数据截至 2026 年 11 月 16 日。



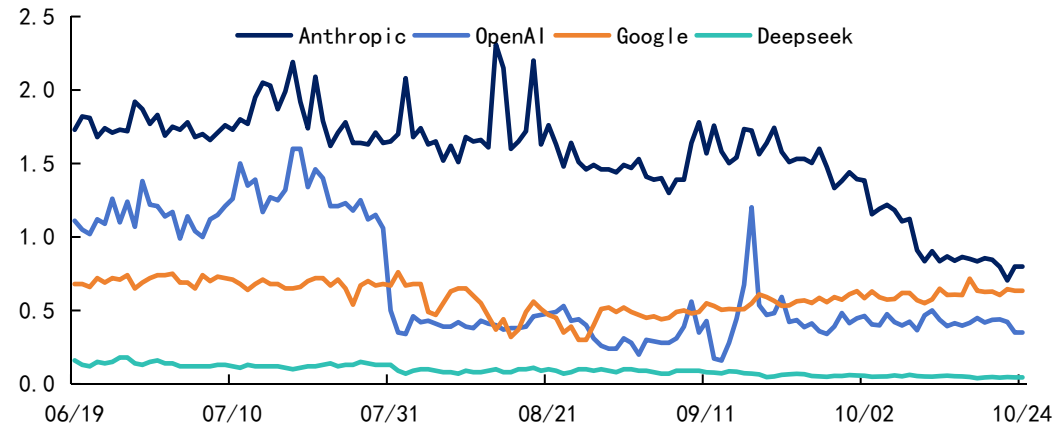
数据来源：Artificial Analysis、OpenAI、Anthropic，国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

复盘2026：混合Token价格持续下行

- **OpenAI率先大幅降价，由轻量档延伸至主力型号，Anthropic后续跟进。**7月底OpenAI将Luna、Terra价格分别下调80%、20%，9月GPT-6 Sol输入、输出价再降50%。Anthropic则将Opus 5.5输入、输出价下调20%，Haiku 5.5较前代降90%。
- **缓存命中是影响平均价格的主力。**目前在用户的实际使用中，缓存命中的token消耗占比很多情况下能达到90%以上，因此缓存命中的token价格下降对整体定价影响较为明显。Fable 5.1、Opus 5.5缓存价分别下调75%、60%；GPT-6.1 Sol与Sonnet 5.5在输入、输出价不变时，将缓存价减半。多轮Agent反复读取相同提示词和历史上下文，缓存降价直接减少这部分重复输入费用。

图：混合 Token 价格（美元/百万 Token，日度）



数据来源：Ornn，国信证券经济研究所整理

表：Claude 价格变化（美元/百万 Token）

日期	模型/事件	输入价	输出价	缓存命中价	主要变化
06/30	Sonnet 5 发布	3→2	15→10	0.30→0.20	输入/输出-33%
07/24	Opus 5 发布	5	25	0.50	定价不变
08/10	Sonnet 5 折扣转永久	2	10	0.20	取消计划涨价
09/01	Fable 5.1 发布	10	50	1→0.25	缓存-75%
09/22	Opus 5.5 发布	5→4	25→20	0.50→0.20	缓存-60%
09/28	Sonnet 5.5 发布	2	10	0.20	输入/输出持平
10/07	Sonnet 5.5 缓存降价	2	10	0.20→0.10	缓存-50%
10/07	Haiku 5.5 发布	1→0.10	5→0.50	0.10→0.01	各项-90%[1]

数据来源：Anthropic，国信证券经济研究所整理

表：OpenAI 价格变化（美元/百万 Token）

日期	模型/事件	输入价	输出价	缓存命中价	主要变化
07/30	GPT-5.6 Luna 降价	1→0.20	6→1.20	0.10→0.02	各项-80%
07/30	GPT-5.6 Terra 降价	2.50→2	15→12	0.25→0.20	各项-20%
08/21	GPT-5.6 Sol 促销	5→4	30→20	0.50→0.40	输入/缓存-20%
09/03	GPT-6 Astra 发布	10	50	1	高端定价
09/22	GPT-6 Sol 发布	4→2	20→10	0.40→0.20	各项-50%
09/22	GPT-6 Luna 发布	0.20→0.10	1.20→0.50	0.02→0.01	输出-58%
09/29	GPT-6.1 Sol 发布	2	10	0.20→0.10	缓存-50%

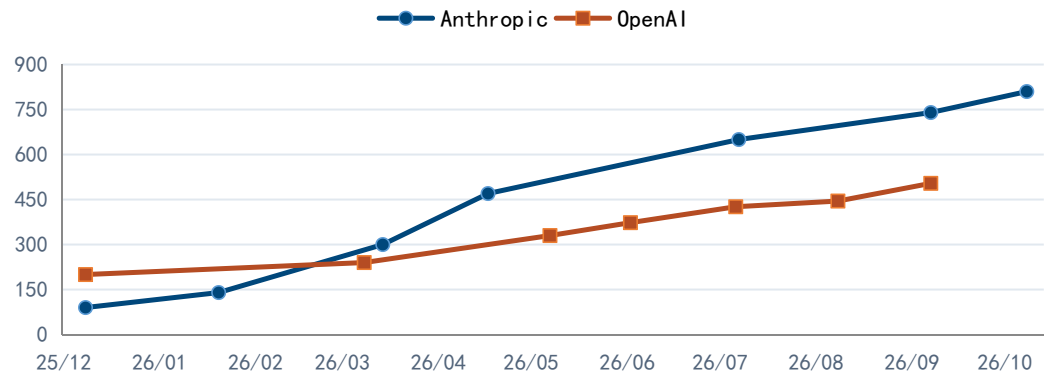
数据来源：OpenAI，国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

复盘2026：Anthropic降价后量增抵价，10月支出恢复增长

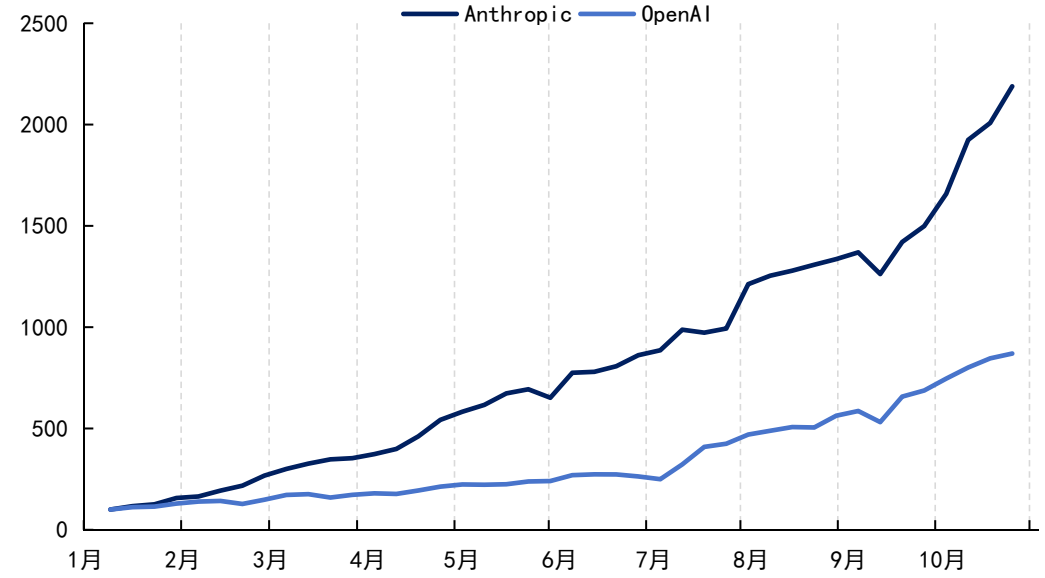
- **OpenAI降价加剧竞争，拖累Anthropic Q3增速。**根据官方数据，Anthropic26Q3收入环比+47%，较Q2回落明显，参考第三方数据ramp ai，样本企业9月较8月，Anthropic 的Token用量仅增7.2%、支出降13.5%，OpenAI则分别增19.3%、16.6%，使用量与支出的增量向OpenAI倾斜。
- **Anthropic跟进降价后Token用量加速，10月量增抵价、收入趋势改善。**根据Ramp ai数据，进入10月后Anthropic用量增40.1%、支出增17.0%，隐含单价降16.5%；OpenAI用量增32.5%、支出降1.7%，增长仍主要体现为消耗量。

图：Anthropic、OpenAI ARR



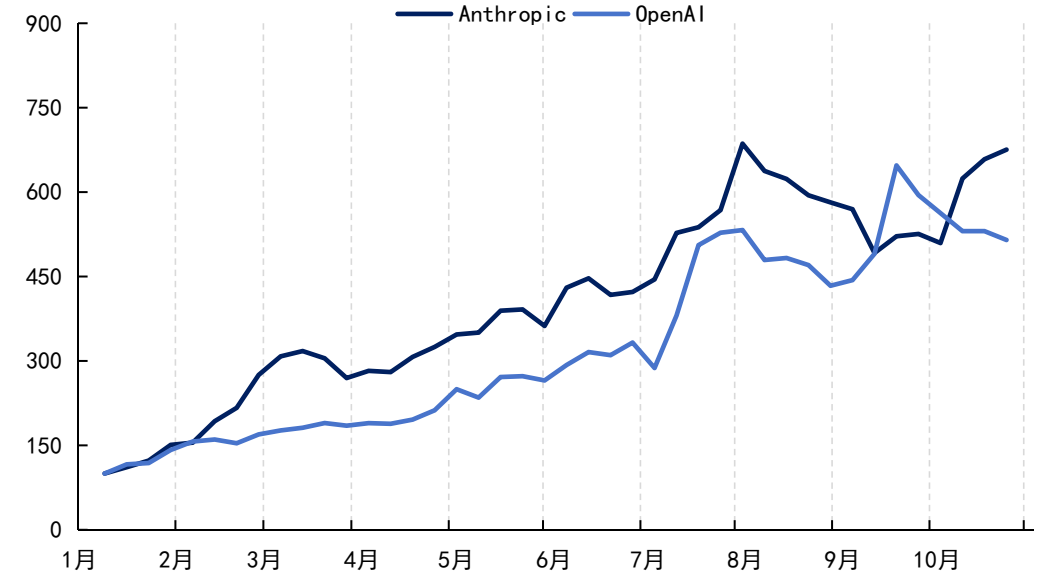
数据来源：Anthropic、OpenAI，国信证券经济研究所整理

图：Token消耗指数（2026/1/9=100，周度）



数据来源：Ramp AI Index，国信证券经济研究所整理

图：Token支出指数（2026/1/9=100，周度）



数据来源：Ramp AI Index，国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

复盘2026：降价进一步推动前沿模型用量份额翻倍

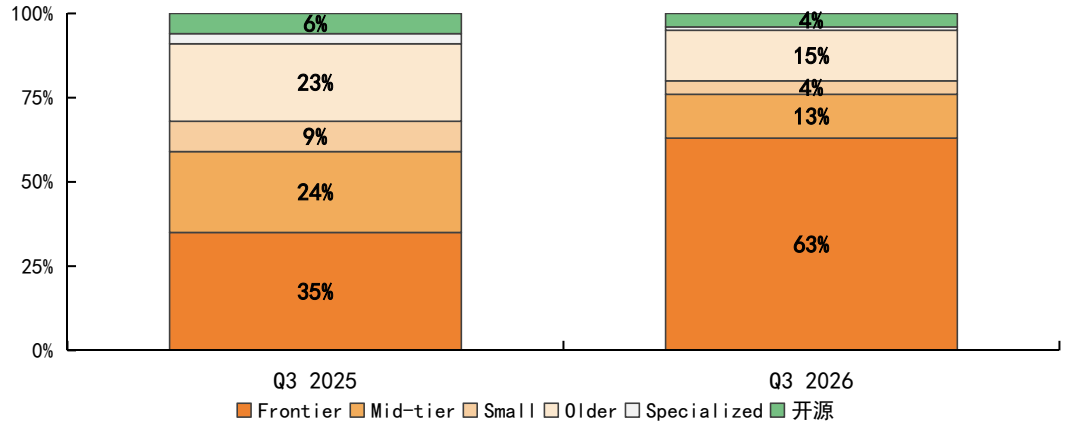
- **前沿模型用量份额超过四成，企业调用向高能力档位集中。**2025Q3至2026Q3，Frontier的Token占比由18%升至42%，支出占比由35%升至63%。Menlo按厂商当期产品线划分层级，同一系列的归类也可能变化，如Sonnet 4.5去年属Frontier、今年Sonnet系列归Mid-tier。
- **Opus两期均属Frontier，同系列升级后输入、输出价下降约73%。**Opus 4/4.1与5.5的缓存命中价也下降约87%；今年新增的Fable缺少去年同系列基期，不计算同比降幅。我们判断，能力升级拓展了模型可完成的任务，降价降低反复调用成本，共同支持用量向前沿模型迁移。

表：Menlo两期模型分层定义

层级	2025Q3	2026Q3
Frontier	GPT-5 Pro、Opus 4/4.1 Sonnet 4.5、Gemini 2.5 Pro	GPT-6 Astra、GPT-5.6 Sol、GPT-5.5 Fable系列、Opus 5/5.5
Mid-tier	GPT-5、Sonnet 4 Gemini 2.5 Flash	GPT-6.1 Sol、GPT-6 Sol GPT-5.6 Terra、Sonnet系列
Small	Mini/Nano、Haiku 3/3.5 Gemini 2.5 Flash-Lite	Luna、Haiku、Gemini Flash系列
Older	同系列落后最新版本至少两代	同系列落后最新版本至少两代
Specialized	图像、音频、网络安全、生物模型	图像、音频、网络安全、生物模型
开源	全部开源模型，涵盖所有档位	全部开源模型，涵盖所有档位

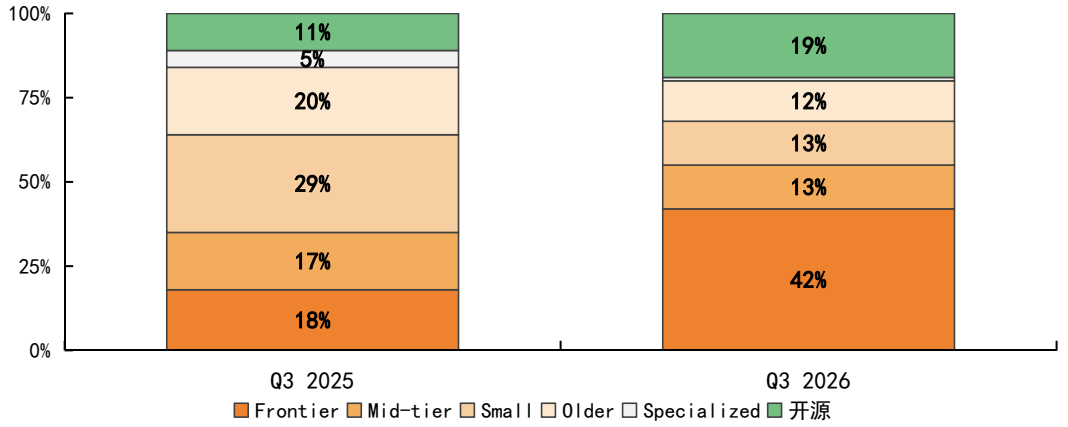
数据来源：Menlo Ventures，国信证券经济研究所整理

图：各档位模型支出份额



数据来源：Menlo Ventures，国信证券经济研究所整理

图：各档位模型Token用量份额



数据来源：Menlo Ventures，国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

降价的重要驱动因素：推理优化提升单卡吞吐

- **推理优化通过减少计算、降低数据搬运和提高GPU利用率，降低单位Token成本。**同样的GPU在相同时间内处理更多Token，才能在价格下降时缓解毛利率压力。具体应先判断瓶颈在计算、显存带宽还是调度，再选择对应方法。
- **不同方法解决的瓶颈不同，实际收益取决于模型和服务场景。**量化与服务系统优化通常可用于已有模型，MoE、稀疏注意力和蒸馏则涉及模型设计或再训练。各项优化可能作用于同一瓶颈，收益不能直接相乘，最终要比较同等能力和延迟要求下的单位Token成本。

表：推理优化的主要路径、代表方法与效果

优化方向	代表方法	对应效果与例子	主要约束
计算量与架构	1. MoE与稀疏注意力 2. 蒸馏与结构化剪枝	1. V4-Pro在1M上下文下，单Token推理FLOPs为V3.2的27% 2. R1蒸馏出1.5B—70B模型，较小模型承接部分推理任务	需要训练或微调验证能力 参数减少不等于同幅降本
计算精度	1. FP8权重与激活 2. FP4浮点量化 3. INT4权重量化	1. 相比FP16，FP8权重裸数据量减半，减少存储与传输 2. NVIDIA NVFP4模型内存约为FP16的1/3.5（含缩放开销） 3. AWQ配合TinyChat，论文测试较HF FP16实现提速超过3倍	依赖硬件和算子支持 需逐模型验证精度损失 内存节省不等于成本降幅
内存与访存	1. FlashAttention算子融合 2. PagedAttention分页管理	1. 通过分块和融合减少HBM读写，保持精确注意力计算 2. vLLM同延迟下吞吐为FasterTransformer/Orca的2—4倍（系统收益）	收益随序列长度与并发变化 系统收益包含多项优化
并行与调度	1. 连续批处理 2. 输入处理/生成分离（P/D）	1. 请求结束后及时补入新请求，减少批次内GPU空闲 2. DistServe在延迟约束下，可服务请求率最高为对照系统的7.4倍	需满足首Token和逐Token延迟 分离后增加KV传输开销
解码算法	1. 推测解码 2. 多Token预测（MTP）	1. 小模型起草、大模型并行验证，原论文T5-XXL测试提速2—3倍 2. V3用MTP辅助推测解码，第二Token接受率85%—90%，TPS为1.8倍	接受率影响收益 高并发下加速可能收窄
硬件升级	1. 更高显存带宽与计算能力	1. H100至B200的单GPU显存带宽由3.35增至8 TB/s，约为2.4倍	需同时考虑GPU租金和功耗 带宽提升不等于吞吐提升

数据来源：DeepSeek、arXiv、NVIDIA，国信证券经济研究所整理

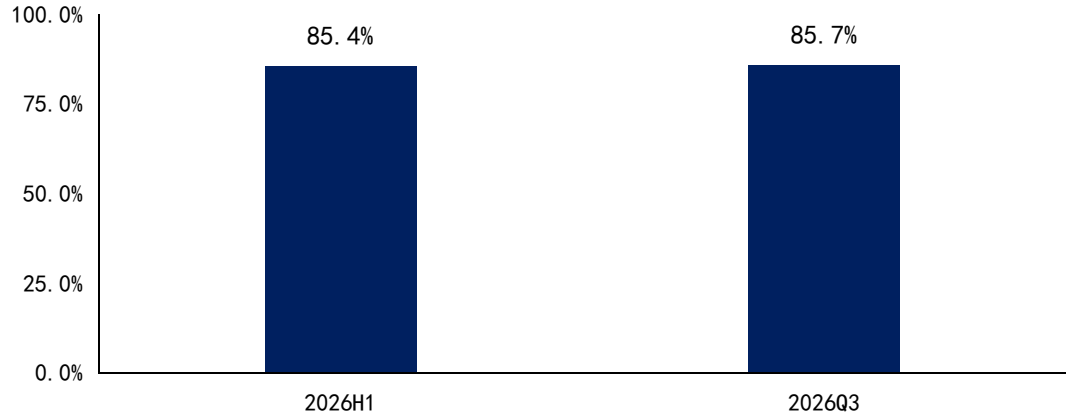
请务必阅读正文之后的免责声明及其项下所有内容

模型厂通过推理优化保证毛利率水平

- 根据公式，模型厂毛利率实际只取决于三个变量：Token 单价 P、有效单卡 TPS（q）、单卡租金 r。卡数 N 与使用时长 T 在分子分母同时出现后约去，因此单纯扩规模、靠摊薄并不改善毛利率。
- **当前存在算力租赁价格持续上涨，但Token价格持续下降的剪刀差，一方面来自行业竞争，另一方面，为了保证自身毛利率水平，只有推理优化做的足够好的模型厂商才能够尽量保证毛利率的稳定。**以Anthropic为例，26Q3较26H1毛利率基本稳定，但公司强调26Q3的99亿美元算力支出中，有约50亿来自高价的短期算力，且公司的训练推理算力比例大致与上半年持平，表明在算力价格上升的情况下，公司依然通过推理优化维持了良好的毛利率。但同时，10月开始Opus系列的明显降价效应开始显现，可能对Q4公司的毛利率水平产生短期扰动。

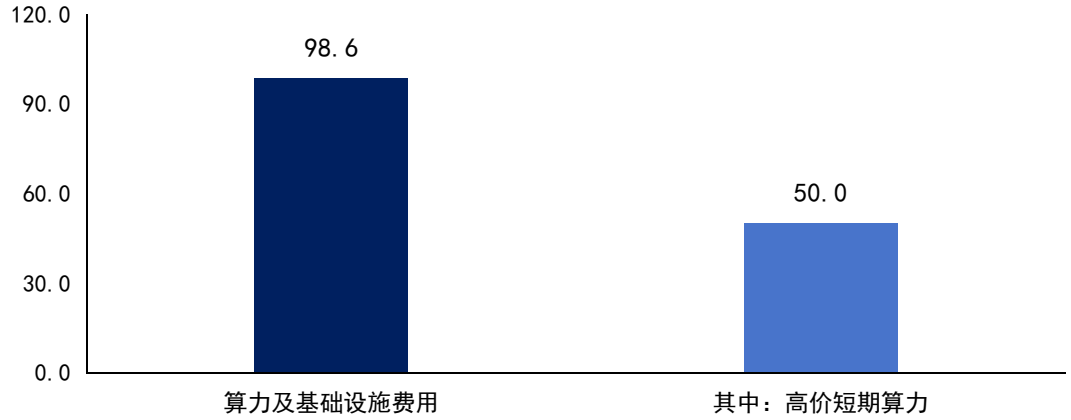
$$毛利率=1-\frac{营业成本}{收入}=1-\frac{GPU数量\times 单卡租金}{GPU数量\times 单卡TPS\times Token单价}=1-\frac{单卡租金}{单卡TPS\ x\ Token单价}$$

图：Anthropic 毛利率（%）



数据来源：Anthropic、SEC，国信证券经济研究所整理

图：Anthropic 2026Q3 算力及基础设施费用（亿美元）



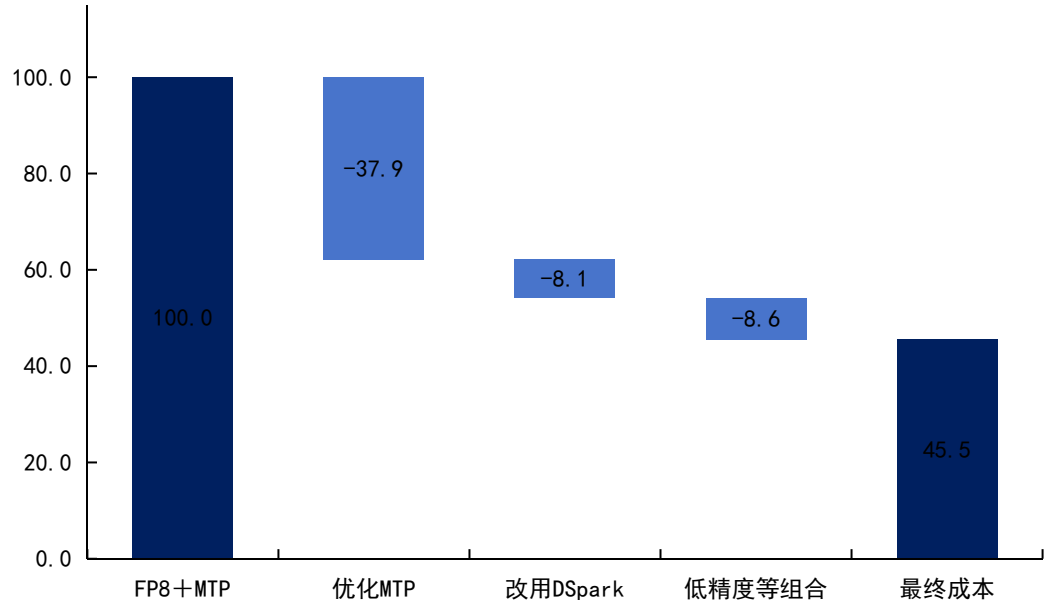
数据来源：Anthropic、SEC，国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

推理优化释放降本空间，支撑Token价格下行

- **LMSYS披露的V4-Pro部署优化案例，对应生成成本下降54.5%。**测试在同一套H20部署上，依次优化MTP执行、改用DSpark投机解码，再叠加低精度与缓存优化，将单卡吞吐从319.92提升至703.15 Token/s；按GPU小时费用不变折算，单位生成成本降低约一半。
- **DeepInfer通过计算、通信和调度协同优化，提高实际服务效率。**DeepSeek将输入处理与输出生成分开部署，通过专家并行和负载均衡减少跨卡等待，并结合投机解码、缓存复用及局部重算，减少生成开销和历史上下文的重复计算。
- **推理优化存在上限，但模型上线后仍可能释放较大降本空间。**我们判断，新模型从初期部署走向充分优化，仍有机会通过减少冗余计算、提高硬件利用率降低单位成本；模型与系统的后续迭代，有望继续支撑Token降价，幅度取决于部署起点和竞争。

图：DeepSeek-V4-Pro生成成本瀑布（基准＝100）



数据来源：LMSYS，国信证券经济研究所整理

表：DeepInfer如何优化推理效率

优化环节	DeepInfer及配套系统的做法	效率改善机制
分阶段部署	输入处理与输出生成分离 分别配置并行规模和资源	避免不同阶段争用资源 分别优化首字延迟与吞吐
专家并行与均衡	生成阶段采用Wide-EP 复制热点专家、动态调整分布	减少专家权重读取 降低忙闲不均及同步等待
投机解码	使用DSpark提出候选Token 按负载动态调整验证长度	一次验证多个候选 平衡接受收益与验证开销
缓存复用与重算	复用历史KV Cache 以SWA bounded replay重建局部状态	减少重复输入计算 兼顾缓存容量与重算成本
底层算子与通信	集成高效计算、通信算子 优化数据传输与执行路径	减少中间数据搬运 提高算力和带宽利用效率

数据来源：DeepSeek，国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

目录

01 头部模型领先优势显著，Agent Infra 有望加速企业 AI 拐点到来

1.1 复盘 2026：Coding 率先引爆 ARR，推理降本推动份额向双寡头集中

1.2 展望 2027（训练侧）：RSI 开启研发自迭代闭环，拉大前沿智能代际差

1.3 展望 2027（应用侧）：Token 降价推动 AI 从编程向垂直知识工作与个人 Agent 渗透

1.4 破局核心（基建侧）：以云为鉴，Agent Infra 补齐生产短板，巨头与独角兽加速布局

02 Anthropic：SOTA 模型构建商业化壁垒，核心收入来自企业客户

03 Anthropic：模型及产品矩阵丰富，覆盖全场景用户需求

展望2027：训练侧—RSI从研发辅助走向局部闭环，有望加快模型迭代

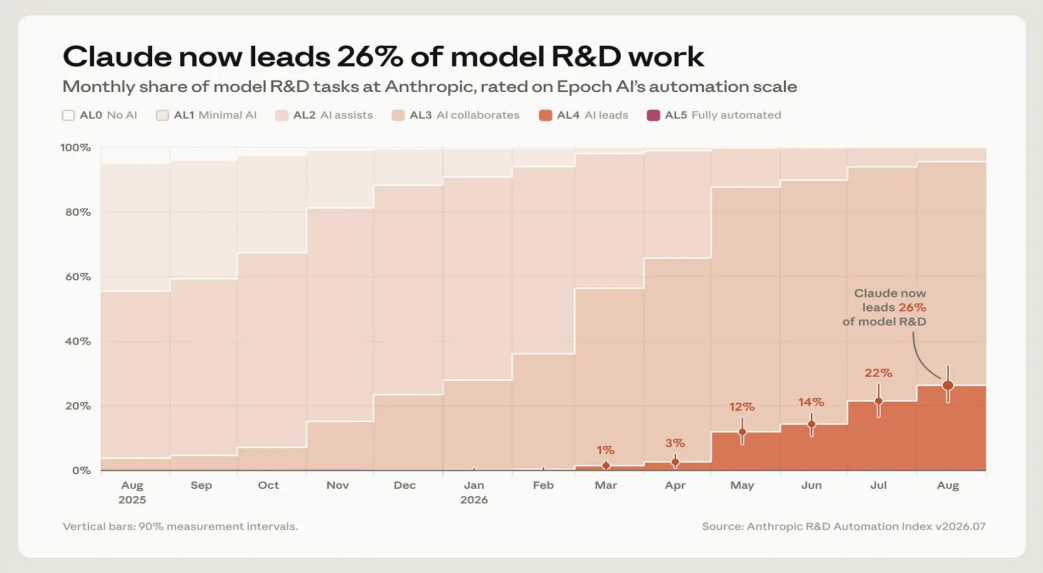
- **前沿数学与科学推理屡破上限，印证底层智能仍在加速迭代，尽管商业化节奏可能放缓。** 26 年以来，头部模型在前沿数学难题求解、复杂科学发现与高难逻辑推理（如 HLE-Diamond、Terminal-Bench-Science 等）上持续取得突破性进展。这表明尽管商业化收入放量在后续可能面临阶段性瓶颈，但从纯粹的“智能演进”维度看，RSI 飞轮驱动下的模型能力上限仍在加速突破。
- **RSI以“模型改进模型”形成递归反馈，中外厂商已在局部研发环节跑通闭环。**RSI指AI参与改进下一代模型，再由更强模型推进后续研发，完整RSI要求模型自主构建后继模型。千问已披露自动迭代训练，智谱从推理基础设施优化切入。
- **Anthropic预判最早2027年初可能迎来研发自动化突破。**Anthropic在《Frontier Safety Roadmap》中判断，最早2027年初，AI可能全面自动化或大幅加速顶级研究团队工作。公司披露，2026年8月Claude已主导26%的AI研发工作，这些任务仍接受人类监督。联合创始人Jack Clark在10月12日Import AI个人文章中，更倾向完整RSI在2027年成为可能，早于此前判断的2028年。

表：模型公司的RSI路径、进展与时间判断

公司	推进路径与进展	时间判断
Anthropic	完整RSI指模型自主构建后继模型；当前已能在人类监督下主导部分研发任务（AL4）。	公司在《Frontier Safety Roadmap》中判断，最早2027年初可能实现研发自动化或大幅加速。 联合创始人Jack Clark在10/12 Import AI个人文章中，更倾向2027年可实现完整RSI。
OpenAI	研究实习生已达成：可在人类指引下完成明确任务，涵盖熟练研究者需数天的工作。	OpenAI在9/6官方博客中提出，目标在2028年3月建成人类监督下的自动化AI研究员。
阿里千问	阿里集团9/25云栖大会回顾披露，Qwen3.8-Max一个月自动迭代33轮，AA评分由40升至45。	
智谱	智谱9/17技术博客披露，GLM-5.3驱动Infra Agent优化自身推理系统，不足两周上线，吞吐达初始基线3倍。	

数据来源：Anthropic、OpenAI、阿里巴巴、智谱，国信证券经济研究所整理

图：Claude研发工作自动化程度



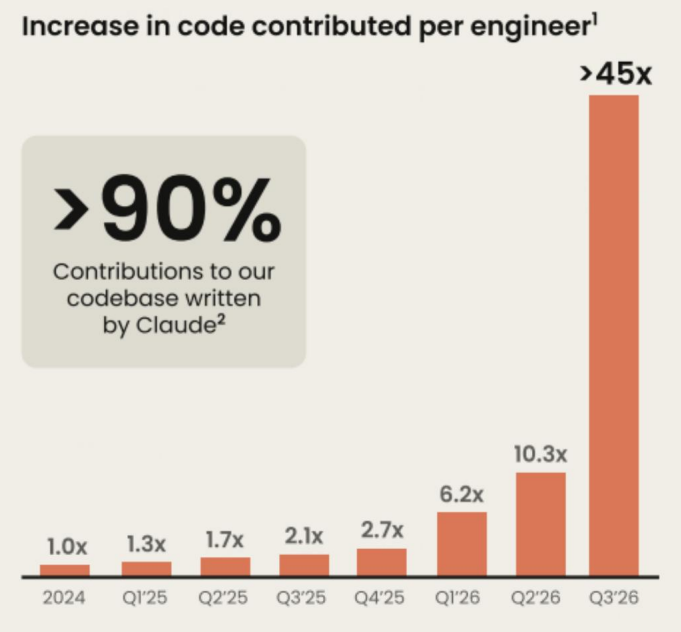
数据来源：Anthropic，国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

展望2027：训练侧—RSI 加速模型研发，领先厂商有望进一步构建壁垒

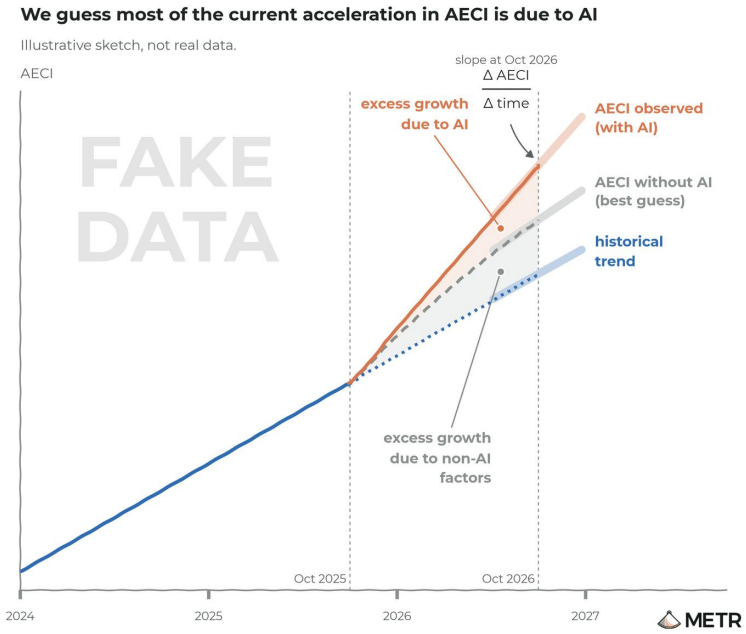
- **开放式任务跨过可用性拐点，研发代码产出提速。**Claude Code 开放式任务成功率从约26%升至91%。三季度，Anthropic 人均日代码贡献量从二季度的10.3倍跃升至45倍以上（2024年为基准）。
- **研发提速有望推升训练算力需求。根据**我们测算，Anthropic三季度研发费用约95亿美元，接近上半年。训练相关支出约74亿美元，超过上半年。前三季度训练支出同比增441%，快于推理支出的168%。
- **率先跑通训练闭环的厂商有望巩固领先。**更强的模型能并行验证更多想法，有效改进进入下一轮训练后，又能加快下一代的研发。

图：Anthropic 三季度人均代码贡献量加速



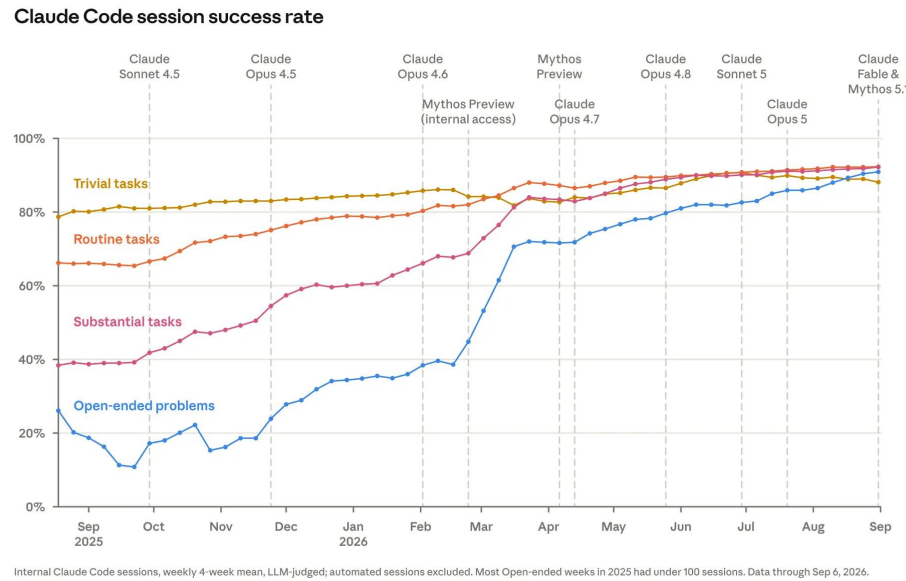
数据来源：Anthropic、SEC，国信证券经济研究所整理

图：METR 对 AECI 提速的归因（示意图）



数据来源：METR，国信证券经济研究所整理

图：Claude Code Q3开放式任务成功率升至91%



数据来源：Anthropic，国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

目录

01 头部模型领先优势显著，Agent Infra 有望加速企业 AI 拐点到来

1.1 复盘 2026：Coding 率先引爆 ARR，推理降本推动份额向双寡头集中

1.2 展望 2027（训练侧）：RSI 开启研发自迭代闭环，拉大前沿智能代际差

1.3 展望 2027（应用侧）：Token 降价推动 AI 从编程向垂直知识工作与个人 Agent 渗透

1.4 破局核心（基建侧）：以云为鉴，Agent Infra 补齐生产短板，巨头与独角兽加速布局

02 Anthropic：SOTA 模型构建商业化壁垒，核心收入来自企业客户

03 Anthropic：模型及产品矩阵丰富，覆盖全场景用户需求

展望2027：推理侧—Token降价推动AI应用走向规模化

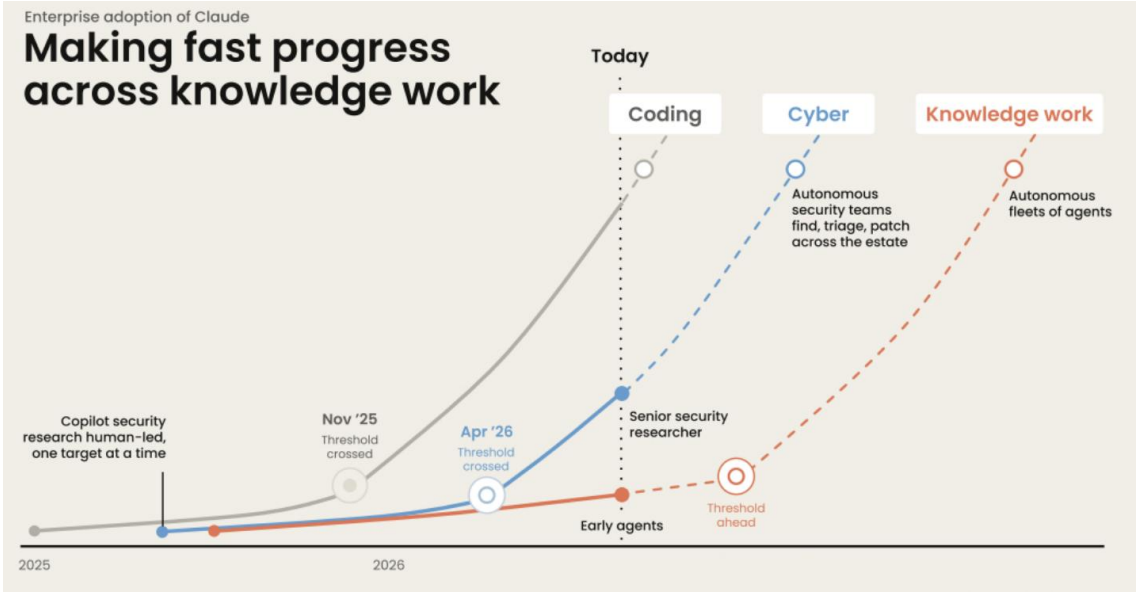
- **Token价格下降意味着应用的爆发成为可能。**推理价格下行降低了多步骤任务反复调用模型的成本，使更多场景达到可接受的单任务成本，为规模化使用创造前提。
- **从编程经验看，结果易验证、决策链路短、ROI可量化的场景落地最快。**Anthropic在路演中判断，网络安全继编程之后于2026年4月跨过能力门槛，知识工作有望成为下一处拐点。C端Personal Agent与物理AI也是后续的重要应用方向。

表：五类推理应用场景的落地条件

应用场景	可验证程度	执行决策链路	付费意愿与预算
编程 缺陷修复	**高** 复现问题后，以单测和回归检查修复结果。	**高（短）** 开发者审查后即可合并，常规修复沿用现有流程。	**高** 研发团队可沿用开发工具预算，为高频使用按席位付费。
网络安全 告警与处置	**高** 告警可用日志、规则和历史事件交叉核验。	**中（中等）** 分析员可判级，隔离生产主机还需业务授权。	**高** 企业已有安全运营预算，买单看漏报率及响应提速。
通用知识工作 检索与起草	**中** 事实和引文可回查原始来源，观点仍需人工复核。	**中（中等）** 团队内初稿可直接协作，对外发布仍需业务审核。	**中** 企业可从办公软件预算付费，基础起草功能容易被套件覆盖。
垂直知识工作 法律、金融 医疗、医药	**中** 合同条款、财务数据可按规则核验。医疗预授权和药企临床资料需专业复核。	**低（长）** 初筛可辅助完成。合同定稿、审批或申报往往需专业人员签核。	**高** 律所、金融机构、医院和药企已有专业软件或人力预算，愿为减少高成本人工付费。
Personal Agent 预约与购物	**中** 订单可凭确认单验证，推荐是否合意靠用户判断。	**高（短）** 涉及支付时由本人确认，商户支持时即可执行。	**低** 偶发代办难支撑月费，付费取决于使用频率和交付结果。
物理 AI 固定工位操作	**低** 节拍、良率可测，需真机长期运行覆盖异常工况。	**低（长）** 现场部署需生产和安全负责人共同验收。	**中** 工厂有自动化预算，采购取决于节拍、可靠性和回收期。

数据来源：国信证券经济研究所

图：Anthropic对编程、网安和知识工作能力拐点的判断



数据来源：Anthropic，国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

展望2027：推理侧—知识工作从通用Agent走向垂直工作流

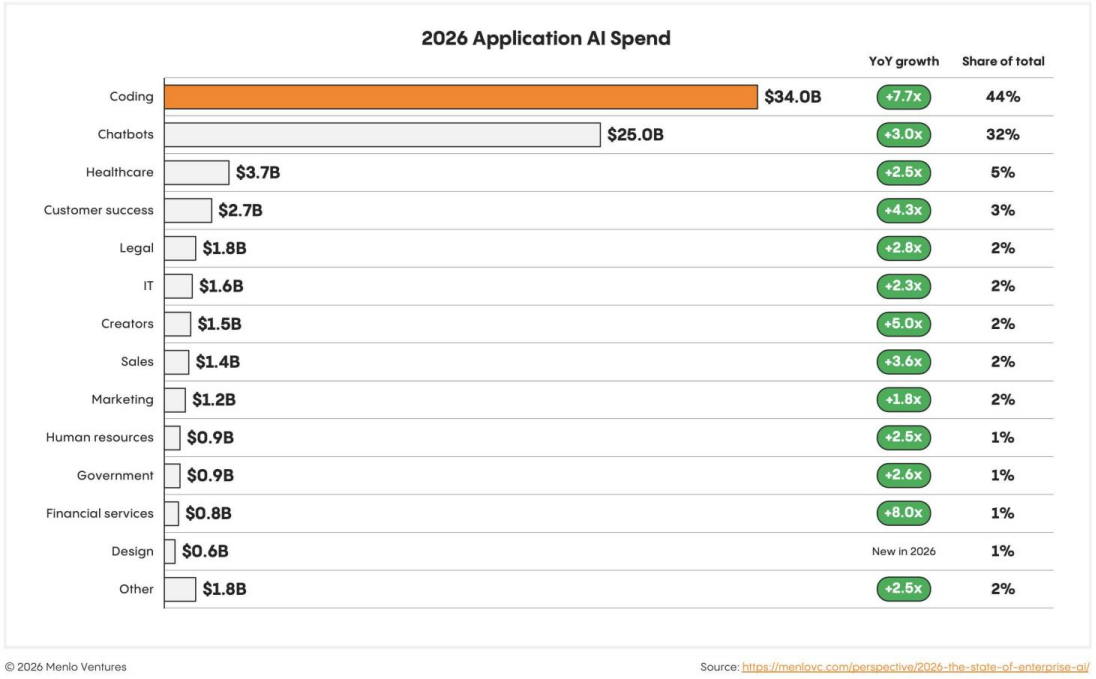
- **通用Agent作为办公工作台，覆盖面广。**用户可在同一工作台完成检索、表格处理和报告撰写等共性任务，基础功能的付费意愿仍有待验证。
- **垂直方案接入行业数据和工具，有望提升单客收入与Token用量。**模型能力从长程编程延伸到3D建模、交互式场景生成和科研资料分析，更复杂的任务需要更多轮调用，也更容易按业务价值定价。

表：知识工作Agent的产品形态与行业应用

产品 / 行业	代表产品与落地场景	重点能力与工具接入
通用Agent	Claude Cowork、ChatGPT Work、腾讯WorkBuddy	跨应用调用，完成检索、表格处理和报告撰写等多步任务
软件	Claude Code、Codex、Cursor	理解大型代码库，在长任务中连续修改代码并运行测试
工业设计	Claude接入Autodesk Fusion、Blender，辅助3D建模与改版	理解场景和参数，通过工具创建、修改三维模型
游戏	关卡与资产原型、脚本编写、自动化测试	Google Genie 4研究演示实时生成可交互场景
金融	Claude金融Agent接入FactSet等数据源，支持投研与尽调	分析财报和表格，调用行情、财务数据并保留核查依据
生物医药	Claude for Life Sciences	解读科研图表与专业文献，汇总和比对研究证据

数据来源：Anthropic、OpenAI、腾讯、Google，国信证券经济研究所整理

图：2026年企业AI应用支出分布（十亿美元）



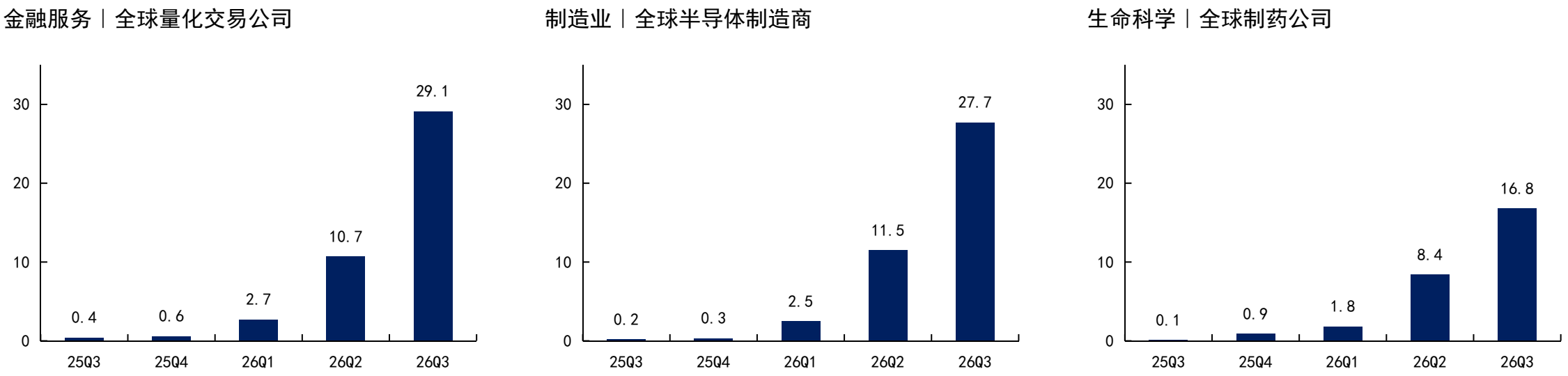
数据来源：Menlo Ventures，国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

展望2027：推理侧—Anthropic垂直行业客户支出增长加速

- **在编程行业之外，垂直行业的用量正处于快速增长，根据Anthropic，一年内，三家样本单季支出均由不足50万美元升至逾1,500万美元。**三个行业样本的支出分别为上年同期的69倍、168倍和115倍。
- **AIG把Claude用在保险核保。**Claude提取并整理投保材料，由核保人员审核。Anthropic称，AIG报价耗时缩短55%。
- **三星System LSI把Claude Code用于定制芯片验证。**据ChosunBiz，搭建验证环境及执行验证原预计需一个多月，实际2天完成。
- **百时美施贵宝把Claude推广至3万名员工。**Claude可跨库查询并生成流行病学报告。客户称，部分查询的答复时间从一两个月缩至约5分钟。

图：三家匿名客户的 Claude 季度支出（百万美元）



数据来源：Anthropic、ChosunBiz，国信证券经济研究所整理

展望2027：推理侧—小参数模型激发Personal Agent需求

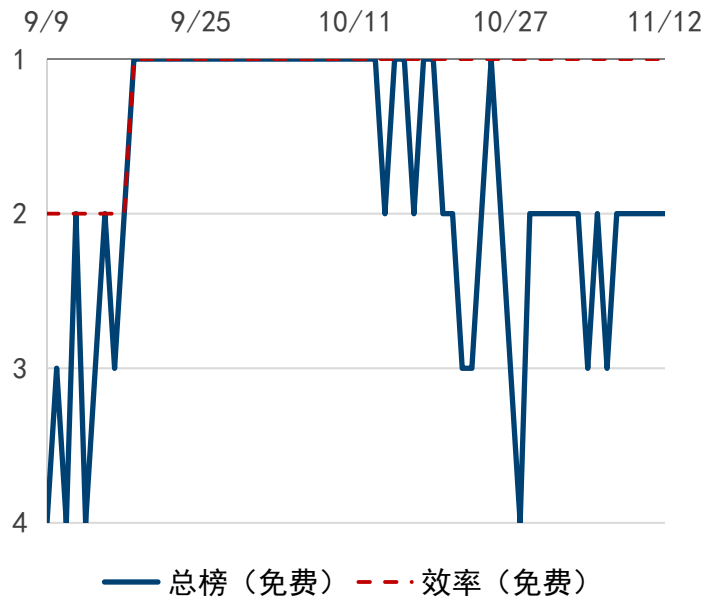
- **三款Agent承接长期任务的方式不同。**Muse把个人偏好和社交内容用作跟进生活事务的线索，用户反馈它能读懂IG转发视频和长对话。Dot延续个人项目的上下文并分派复杂任务。Grok Bot把做过的流程固化为团队可复用的技能。
- **变现取决于稳定交付，也受外部平台规则约束。**用户反馈Muse约一分钟订好纽约至伦敦机票，DoorDash登录却要本人验证。同一用户称用Muse购物后次日看到相关广告。Meta称对话与云机数据不进入广告系统，代访商家网站可能间接影响广告投放。

表：Personal Agent的任务编排、执行权限与商业模式

对比维度	Meta Muse	OpenAI Dot	Grok Bot
任务安排	从个人目标拆计划，关掉应用仍可盯价、等档期，变化时提醒。可把 IG 收藏做成购物看板	一个 Dot 并行推进多个项目。主动查已连应用，也可定时复查。复杂任务交给 Work 或 Codex	按岗位建多个 Bot。演示一次可存成 Skill，按时间或 Slack、GitHub 事件触发
权限与人工接管	连接器与云端浏览器并用。付款逐笔确认。用户实测 DoorDash 手机号验证需接管	授权后可用本地文件和浏览器。外发、购买等动作可分别设为预批准、逐次确认或本人处理	同账户多 Bot 共享浏览器登录和文件，各用独立屏幕并行。团队共享时按成员权限操作
收费方式	免费基础层加订阅。支持 Stripe Link 结账，未来交易变现取决于商户接入及费率	随 Pro 等高阶套餐提供。企业可按档位或用量付费，Work / Codex 任务另计	随 X Premium 及以上套餐或付费 Cursor 提供。X 套餐内与 Grok、Cursor 共用周额度
广告边界	Muse 未公布内置广告方案。对话与云机数据不进广告系统。商家网站活动可间接影响投放	Free / Go 可展示广告。Dot 随 Pro 等无广告套餐提供	按订阅和用量收费。Grok Bot 未公布内置广告方案

数据来源：Meta、OpenAI、Cursor，国信证券经济研究所整理

图：Muse免费榜排名（9/9—11/12）



数据来源：七麦数据，国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

估值增长较快的 15 家 AI 应用公司

- **编程与应用生成兼具规模和增速。**Cognition 4 个月升 85% 至 480 亿美元。Lovable、Replit 让非工程用户也能搭建应用，估值约为上次的 2 倍和 3 倍，分别达 133 亿和 90 亿美元。
- **语音、客服和法律也跑出百亿美元公司。**ElevenLabs 将语音生成拓展至实时客服，7 个月估值翻倍至 220 亿美元（员工老股）。Sierra、Harvey 分别切入订单处理和合同审查，估值增长 58% 和 41%，达 158 亿和 155 亿美元。

表：AI 应用公司上次与本次估值（亿美元，按本次估值降序）

公司	应用方向	上次估值时间	上次估值	本次估值时间	本次估值	间隔（月）	本次 / 上次	月均复合增速
Cognition	编程 Agent	2026.05	260	2026.09	480	4	1.85倍	16.6%
ElevenLabs	语音 AI	2026.02	110	2026.09 老股	220	7	2.00倍	10.4%
Sierra	客服 Agent	2025.09	100	2026.05	158	8	1.58倍	5.9%
Harvey	法律 AI	2026.03	110	2026.09	155	6	1.41倍	5.9%
Lovable	应用生成	2025.12	66	2026.08	133	8	2.02倍	9.2%
Replit	应用开发	2025.09	30	2026.03	90	6	3.00倍	20.1%
Legora	法律 Agent	2026.04	56	2026.10	88	6	1.57倍	7.8%
Higgsfield	视频生成	2026.01	约 13	2026.08	54	7	4.15倍	22.6%
Suno	音乐生成	2025.11	24.5	2026.06	54	7	2.20倍	12.0%
Decagon	客服 Agent	2025.06	15	2026.01	45	7	3.00倍	17.0%
Synthesia	数字人视频	2025.01	约 21	2026.01	40	12	1.90倍	5.5%
Genspark	办公 Agent	2026.03	约 16	2026.06	26	3	1.62倍	17.6%
Rogo	金融工作流	2026.01	7.5	2026.04	约 20	3	2.67倍	38.7%
Profound	AI 搜索营销	2026.02	10	2026.09	18	7	1.80倍	8.8%
Granola	会议助手	2025.05	2.5	2026.03	15	10	6.00倍	19.6%

数据来源：公司公告、Bloomberg、Reuters、TechCrunch、CNBC、Euronext、UCL Ventures，国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

目录

01 头部模型领先优势显著，Agent Infra 有望加速企业 AI 拐点到来

1.1 复盘 2026：Coding 率先引爆 ARR，推理降本推动份额向双寡头集中

1.2 展望 2027（训练侧）：RSI 开启研发自迭代闭环，拉大前沿智能代际差

1.3 展望 2027（应用侧）：Token 降价推动 AI 从编程向垂直知识工作与个人 Agent 渗透

1.4 破局核心（基建侧）：以云为鉴，Agent Infra 补齐生产短板，巨头与独角兽加速布局

02 Anthropic：SOTA 模型构建商业化壁垒，核心收入来自企业客户

03 Anthropic：模型及产品矩阵丰富，覆盖全场景用户需求

展望2027：推理侧—企业 Agent 规模化落地面临三大卡点

- **模型能力决定 Agent 能否稳定完成复杂任务。** 以采购审核为例，Agent 要从不同格式的报价单和合同中提取含税价、币种和付款条款，再根据采购规则判断是否需要补充材料或提交审批。如果模型把未税价当成含税价，后续比价和预算判断都会偏离。若模型遇到条款冲突时仍直接给出结论，审核员就得回查原始文件、重做部分核对，整项任务能交给 Agent 独立完成的比例也会下降。
- **Agent Infra 决定 Agent 能否接进生产系统并稳定运行。**如果采购系统、合同库和审批记录没有打通，Agent 只能依赖人工上传的材料，可能拿不到最新报价或历史审批意见。若任务在工具调用超时后无法从中断处恢复，就需要人工重跑，还可能重复提交审批。若无法按岗位控制访问权限并留存操作记录，企业也难以开放敏感数据，将 Agent 推广到更多团队。
- **业务变革决定企业能否把效率兑现为 ROI。**即使 Agent 能生成准确的审核意见，如果每笔采购仍要人工逐项复核，或者遇到例外时没有明确的接手人，审核总耗时和积压可能变化有限。企业需要明确哪些任务可由 Agent 直接完成、哪些必须转人工，以及谁负责处理异常，再衡量节省的工时能否覆盖推理、复核和运营成本。
- **展望 2027 年，Agent Infra 成熟有望推动 Anthropic ARR 重新加速。** 随着系统接入、任务恢复和权限管理逐步完善，企业有望把 Agent 从少数团队的试点推广到更多日常生产流程，带动推理用量和企业付费规模增长。

图：企业 Agent 落地的三层框架



数据来源：Anthropic，国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

通过收购等方式获取垂直行业数据提升模型能力

国信证券 GUOSEN SECURITIES

- **垂直行业的瓶颈在于业务流程、行业/公司数据的获取难度较大，部分场景反馈周期较长，最终影响模型表现。**公开数据难覆盖完整工作样本和机构内部信息，投资、临床等任务的结果反馈慢，最终决策仍需专家复核和签核。
- **两家公司通过产品部署和投并购解决这些问题。**Anthropic用Skills固化流程，OpenAI用Frontier管理权限与评估。两者通过授权接入内部数据，用即时核对和专家复核缩短纠错，并借投并购获得行业团队和真实工作样本。

表：Anthropic投资与收购

时间	标的	性质	补充的能力或场景
26.10	OPAQUE	团队加入	机密计算与敏感数据访问控制
26.08	Sapiom	参投 A 轮	Agent 工具调用与支付基础设施
26.07	Ode	合资	为中型企业部署 Claude
26.07	Mithril Medicine	收购	直接参与在运行的医药研发项目
26.05	Stainless	收购	SDK、MCP 与业务系统接入
26.04	Coefficient Bio	收购	引入计算生物学团队和专业经验
26.04	Runhouse	团队加入	引入机器学习基础设施工程团队
26.02	Vercept	收购	提升 Claude 跨应用操作能力
25.12	Bun	收购	Claude Code 运行时及工具链
25.08	Humanloop	收购	引入模型评估与工作流迭代团队
24.07	Anthology Fund	联合发起	Menlo 出资 1 亿美元扶持应用

数据来源：Anthropic，国信证券经济研究所整理

表：OpenAI及OpenAI Startup Fund投资与收购

时间	标的	性质	补充的能力或场景
26.11	Odessia	收购	旅行规划和预订 Agent
26.11	Pierre	宣布拟收购	代码浏览和审查工具接入 Codex
26.10	Coherence Neuro	基金共同领投	神经技术与肿瘤医疗设备
26.10	Orchestra Bio	收购	药企研发协作平台接入 Rosalind
26.09	Biossil	基金领投	用 AI 评估药物项目的开发机会
26.07	Chai Discovery	参投 C 轮	生命科学模型和研发工具
26.07	Northslope	DeployCo 拟收购	补充企业实施和部署团队
26.06	Ona	宣布拟收购	支持在客户云环境内运行 Agent
26.05	Tomoro	宣布拟收购	为 DeployCo 引入约 150 名部署人员
26.04	Hiro	团队加入	引入个人理财 Agent 团队
26.03	Promptfoo	宣布拟收购	补充 Frontier 的评估与上线测试工具
26.01	Merge Labs	种子轮投资	布局脑机接口
26.01	Torch	收购	汇总分散的个人病历和健康记录
25.09	Statsig	宣布收购	引入 A/B 测试与发布监测能力
24.02	Ambience	基金共同领投	医院病历记录、编码与文档质控
22.12	Harvey	基金早期投资	后续合作定制判例模型，律师参与评估

数据来源：OpenAI、OpenAI Startup Fund，国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

Agent Infra 让 Agent 能持续、安全地执行

- **Agent Infra 是支撑智能体部署、运行和治理的基础设施。**它让 Agent 能在生产环境中持续、安全地执行多步任务。
- **广义的 Agent Infra 包含六个模块。**编排、运行与隔离、记忆与上下文、工具与网关、身份与安全、可观测性与评测。
- **Harness 编排模型和工具调用，Runtime 调度进程，沙箱隔离执行。**记忆服务保存进度和上下文，网关接入企业系统，身份与安全管理权限和审批，可观测性与评测用于定位故障、检验结果。

表：Agent Infra 的六个模块及其在报销流程中的作用

模块	负责什么 / 以报销 Agent 为例	代表产品
Harness / 编排	组织模型与工具的循环。把发票和制度交给模型，按校验结果写草稿或转人工。	LangGraph、Claude Agent SDK
Runtime / 沙箱	托管 Runtime 复用或启动进程，确认后重新派发原任务。若调用生成的脚本，沙箱限制其文件和网络权限。	AWS AgentCore Runtime、Cloudflare Sandbox
记忆 / 上下文	状态存储保存任务编号、草稿和待确认状态，恢复时按编号读取。上下文或记忆服务提供相关历史和规则。	AWS AgentCore Memory、Google Memory Bank
工具 / 网关	把 ERP、制度库和报销系统接成可调用工具，规范查制度、写草稿和正式提交的接口。	MCP、AWS AgentCore Gateway
身份 / 安全	限制可访问的公司主体和可提交金额。确认前只允许查询和写草稿，正式提交须审批并留痕。	AWS AgentCore Identity、Google Agent Gateway
可观测性 / 评测	记录失败位置、模型费用与接口结果。用历史发票回放，检验分类准确率、越权拦截和重复提交。	LangSmith

数据来源：Anthropic、OpenAI、AWS、Google Cloud、Cloudflare、LangChain，国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

复盘传统云计算时代，跨越“模型可用”到“生产可靠”，Agent Infra 是爆发前提

- **算力/模型够用仅能跑通早期尝鲜，基础设施缺失才是阻碍企业规模化落地的核心症结。** 复盘云计算早期（2006－2008 年），EC2 与 S3 虽已解决底层算力与存储的可用性，但因缺乏细粒度权限与生产级监控，传统大企业渗透率近乎为零。这与当前 AI 落地困境高度相似：基座模型“智商”已基本够用，但支撑长周期运行的 Agent Infra 尚未成熟，导致企业难以将试点推向核心生产环节。
- **AWS 补齐监控、权限与审计“铁三角”后，才真正打通企业信任壁垒，迎来二次加速。** 2009－2014 年间，随着 CloudWatch（监控）、IAM（权限隔离）与 CloudTrail（合规审计）相继商用，AWS 彻底补齐生产级基建短板，促成传统巨头与政务客户核心系统大规模迁云，推动 AWS 营收增速在 2015 年（百亿美元体量下）从 49.4% 逆势跃升至 69.7%。
- **对标当下，Agent Infra 的成熟将是解锁通用知识工作、引爆下一轮 ARR 非线性增长的关键。** 当前企业在 Coding 之外的推广受阻，本质是缺乏可控的读写权限、中断恢复机制与全链路审计。随着 2027 年权限治理、持久化 Runtime 与可观测性等 Agent Infra 模块走向标准化，企业部署的合规风险与试错成本将大幅收敛，驱动 Agent 从局部试点走向全面生产爆发。

表：AWS 云计算发展复盘

发展阶段	代表年份	AWS 年收入规模	同比增长率（YoY）	行业核心特征与客户结构
第一阶段：基建缺失期（仅有 EC2/S3，无 IAM 与成熟监控）	2006－2008 年	2006 年仅 $2,100 万美元 2008 年不足 $5 亿美元	早期基数极小，年增速约 60%－80%（但绝对增量每年仅数千万至二亿美元）	“极客与初创公司玩具”：因为没有细粒度权限和监控，客户主要是互联网初创公司（如早期的 Netflix、Dropbox），传统大型企业渗透率几乎为零。
第二阶段：铁三角合体与破局期（2010－2011 年 IAM/CloudWatch 商用）	2010－2013 年	2010 年约 $10－15 亿美元 2012 年突破 $15－20 亿美元 2013 年达 $31.1 亿美元（首次实现全年盈利）	绝对增量开始起飞 （2013 年首次拿下 CIA 6 亿美元政务云大单，标志政企信任破冰）	“企业级信任破冰”：IAM 解决了数据隔离与权限合规，CloudWatch 解决了生产级 SLA 监控。传统企业（Capital One、GE、美国中央情报局 CIA）开始将核心系统迁入云端。
第三阶段：企业全面上云爆发期（基建成熟后的规模化收割）	2014－2018 年	2014 年：$46.4 亿美元 2015 年：$78.8 亿美元 2016 年：$122.2 亿美元 2018 年：$256.6 亿美元	2014 年：+49.4% 2015 年：+69.7%（百亿级体量下的二次加速！） 2016 年：+55.1% 2018 年：+46.9%	“大象起舞与 ARR 指数爆发”：在基建完全成熟后，AWS 不仅没有因为体量变大而失速，反而在 2015 年百亿美金体量上打出近 70% 的恐怖增速，一年净增营收高达 32 亿－82 亿美元（单年增量已超过前 7 年的总和）。
第四阶段：全球超大规模成熟期	2020－2025 年	2020 年：$453.7 亿美元 2025 年：$1,287.3 亿美元（公有云全行业逼近 8,000 亿美元）	维持在 18%－37% 稳健高增（2025 年单年净增营收达 $211.7 亿美元）	云计算彻底成为全球商业的“水电煤”。

数据来源：Amazon，国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

Agent Infra 融资增加，头部估值加速上行

- **头部估值在半年左右就出现数倍上调。**Modal 从2025年9月的11亿美元升到2026年10月的157.5亿美元，约13个月达到14.3倍。其中最近一轮较上轮约5个月达到3.4倍，Temporal 约7个月达到2.5倍。
- **融资金额增长集中在大额轮次。**过去12个月已披露融资约47亿美元，对比前12个月约20亿美元，交易笔数由72笔增至74笔。剔除单轮超过5000万美元的交易后，融资额增幅约7%。

表：Agent Infra 样本头部公司的融资估值节点（亿美元）

序号	厂商	主营业务	上上节点（时间 / 金额）	上一节点（时间 / 金额）	最新节点（时间 / 金额）	近两次估值变化
1	Modal	弹性计算与 Agent 沙箱	25/09 B 轮 · 11.0	26/05 C 轮 · 46.5	26/10 D 轮 · 157.5	约5个月至3.4倍
2	Temporal	持久化工作流执行	25/03 C 轮 · 17.2	26/02 D 轮 · 50.0	26/09 E 轮 · 125.5	约7个月至2.5倍
3	OpenRouter	多模型路由与统一计费	25/06 A 轮 · 约5.47	26/05 B 轮 · 13.0	26/08 拟并购 · 75－85	融资约11个月至2.4倍
4	n8n	可视化工作流编排	25/03 B 轮 · 约2.7	25/10 C 轮 · 25.0	26/05 战略入股 · 52.0	约7个月至2.1倍
5	LangChain	Agent 编排与评测	24/02 A 轮 · 约2.0	25/10 B 轮 · 12.5	26/11 C 轮 · 30.0	约13个月至2.4倍
6	Exa	面向 Agent 的网页搜索	24/07 A 轮 · 约0.7	25/09 B 轮 · 7.0	26/05 C 轮 · 22.0	约8个月至3.1倍
7	Parallel Web Systems	网页搜索与研究 API		25/11 A 轮 · 7.4	26/04 B 轮 · 20.0	约5个月至2.7倍
8	LiveKit	实时音视频与语音 Agent	25/04 B 轮 · 3.45	26/01 C 轮 · 10.0	26/08 C-1 · 约14.5	约7个月至1.5倍
9	Zenity	Agent 安全与运行治理	23/09 A 轮	24/10 B 轮	26/08 C 轮 · 8.33－16	
10	Daytona	隔离云环境与代码执行	24/06 种子轮 · 约0.25	26/02 A 轮 · 约1.23	26/10 B 轮 · 10.0	约8个月至8.1倍

数据来源：New Market Pitch、Forge、TechCrunch，国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

模型厂组合交付自有模型与 Agent Infra

- **模型厂以自有模型为核心，组合交付 Agent Infra。**OpenAI 将 Codex Harness、会话和沙箱纳入 Agents API。Anthropic 将 Claude Harness 与持久会话、工具接入和执行环境纳入 Managed Agents。两家也通过 SDK 支持客户自建。
- **模型调用是核心收费项，部分配套服务另外计量。**OpenAI 按容器计量托管沙箱，Anthropic 的会话运行费为每小时 0.08 美元。组合交付便于同时迭代模型与 Harness，联合优化工具调用、上下文管理和任务恢复。

表：OpenAI 与 Anthropic 的 Agent Infra 产品、主要功能和商业模式

厂商	产品环节	产品	主要功能	商业模式
OpenAI	开发框架	Agents SDK	在客户应用中编排模型、工具和多个 Agent，提供交接、审批与执行追踪。	开源框架，自建部署。模型与所用托管工具分别计量。
	托管执行	Agents API	托管 Codex Harness、会话和任务恢复。支持 OpenAI 托管沙箱或接入客户执行环境。	模型按 API 用量计费，OpenAI 托管沙箱按容器标准计费。
	企业平台	Frontier	接入共享业务数据和工具，为企业 Agent 配置身份、访问权限、评测和审计。	企业解决方案，联系销售。
	部署服务	Presence	协助将企业工作流接入生产系统，负责实施、评测和持续监控。	按客户及部署范围定价。
Anthropic	开发框架	Claude Agent SDK	在客户应用中嵌入 Claude Code 的执行循环、内置工具、权限控制和上下文管理。	自建部署，模型按 Claude API 用量计费。
	托管执行	Claude Managed Agents	托管 Claude Harness、持久会话和沙箱。可接入客户自管沙箱，支持长任务恢复。	模型 Token 按量。会话运行时间另计，标价 0.08 美元/小时。

数据来源：OpenAI、Anthropic，国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

头部模型厂近期研究均指向如何让 Agent 可靠承接复杂任务

- **近期头部模型厂商围绕Agent发布了相关研究，共同问题是如何把复杂任务稳定地交给 Agent。**任务可能跨系统、跨会话，并在执行中等待审批或调整目标。研究因此聚焦授权与审计、状态与经验的保存，以及多任务调度。常驻 Agent 是这些能力组合后的一类应用，可以在业务事件到来时启动并接续工作。
- **OpenAI：**内部已有逾 **500** 个 **Agent** 在后台主动执行任务，但**当 Agent 可以在后台主动做事、运行代码和调用企业工具时，怎样把它的行动限制在企业能授权、能追踪、也能叫停的范围内是重要议题。**文章提强调将安全控制与自主执行分开。可信控制服务负责身份、网络和审批，模型生成的操作在隔离环境中执行。Agent 使用独立身份，企业可以追溯动作并撤销权限。自动复核把需要人工判断的操作留给人审。
- **Anthropic 分别研究任务如何接续、经验如何保留。**11 月 13 日的 Always-on Agents 指南让业务事件或定时器唤醒 Agent，每次处理后保存进度并休眠。11 月 6 日的 Agent Memory 研究提出了Dreaming的机制，用分层记忆让 Claude 在新会话里复用过去的经验，再用 Dreaming 定期整理和纠错。CL-Bench 案例中，Opus 5 能把失败经验提炼为经验证的规则，供新会话复用。
- **DeepSeek 关注如何更高效实现多Agent任务执行。**DeepSeek观察到，人通常只能同时跟踪约 3－5 个活跃会话，把几项工作都放进一个会话，长工具调用会阻塞交互，不同任务的上下文也容易混在一起。给每项工作各开一个会话，又要由人记住进度、转达变更和汇总结果。Super-agent 想把这部分协调工作交给系统，让人主要负责定目标和做关键决策。
    - Super-agent 让 Router 维护跨渠道互动与任务关系，Worker 在隔离会话中并行执行。长任务运行期间，用户仍可通过同一入口提交或调整任务。合成测试中，任务完成时间中位数从 20.4 分钟降至 12.4 分钟，缓存读取 Token 同时增加。

请务必阅读正文之后的免责声明及其项下所有内容

云厂商Agent Infra产品功能接近，微软在办公场景具备优势

国信证券 GUOSEN SECURITIES

- **三家云厂商的 Agent Infra 底层功能较为接近。**AWS AgentCore、微软 Foundry 和 Google Cloud 的 Gemini Enterprise Agent Platform 均覆盖托管运行、工具接入、会话管理和评测。
- **微软将跨平台治理Agent 365单独作为重要产品线推出，并按席位订阅收费。**Agent 365 集中纳管微软及第三方 Agent，并将 Entra 的身份验证、访问控制和微软的安全合规策略延伸到 Agent，美国独立标价为每用户每月 15 美元。AWS 和 Google Cloud 将相近的治理能力纳入各自的 Agent 平台，相关组件按用量计费。商业模式的差异核心是基于微软在IAM市场的领先地位，FY26Q1 披露 Entra 月活用户已达 10 亿，未来微软有望依托过去在SaaS领域优势，快速取得Agent infra市场份额。

表：AWS、Microsoft、Google Cloud 的 Agent Infra 产品与商业模式

厂商	产品系列	产品 / 组件	主要功能	商业模式
AWS	Amazon Bedrock AgentCore	Runtime / Harness	Runtime 托管自建 Agent 的部署、隔离与扩缩容。Harness 托管模型与工具的执行循环。	按组件用量计费。Runtime 按计算、内存时长，Registry 按登记记录数，Policy 按授权请求。
		Gateway / Identity / Policy	将 API 接为 MCP 工具，管理身份及委托授权，并按业务规则限制工具调用。	
		Memory	保存会话上下文，提取可跨会话复用的长期信息。	
		Browser / Code Interpreter Payments	提供托管浏览器和代码沙箱。Payments 支持限额内调用付费 API 与内容。	
		Observability / Evaluations Optimization / Registry	追踪执行并评估质量，通过试验优化配置。Registry 登记和发现 Agent、工具及技能。	
Microsoft	Microsoft Foundry	Foundry Agent Service Foundry Control Plane	构建、部署和托管 Agent，接入模型、工具和企业数据，并做追踪、在线评测和运行监控。	模型、工具和容器按用量计费。提示词或工作流型 Agent 无额外服务费。
	Microsoft Agent 365	Agent 目录与治理	跨平台登记、纳管 Agent，联动 Entra、Defender 与 Purview 实施身份、安全与合规策略。	美国独立标价 15 美元/用户/月，也纳入 Microsoft 365 E7。
Google Cloud	Gemini Enterprise Agent Platform	Agent Runtime / Sessions Memory Bank	托管多框架 Agent，保存会话交互历史，并在不同会话间检索长期记忆。	按组件用量计费。Runtime 按 vCPU、内存时长，Agent Gateway 按调用或授权请求。
		Agent Gateway / Agent Identity 评测与可观测性	统一连接和授权，执行安全策略，并对测试或生产流量做追踪与效果评估。	

数据来源：AWS、Microsoft、Google Cloud，国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

独立厂商的Agent Infra产品与云厂差异主要体现在部署方式上

国信证券 GUOSEN SECURITIES

- **独立厂商与云厂商的 Agent Infra 产品功能已有较大重叠，均覆盖编排、运行、状态管理和评测。差异主要体现在部署方式。**云厂商将这些能力与自身的计算、身份权限和数据服务集成，便于承接已在本云运行的业务。独立厂商可以接入企业现有的模型、框架和运行环境，客户可按需采购其中一项能力。

表：LangChain、Cloudflare、Modal、Temporal 的 Agent Infra 产品与商业模式

厂商	产品	主要功能	商业模式
LangChain	LangGraph / Deep Agents	LangGraph 提供有状态的分支、暂停与恢复。Deep Agents 提供可复用的规划、工具调用及子 Agent 编排。	开源框架，可自托管。
	LangSmith Observability & Evaluation / Agent Server	对接不同框架的执行轨迹，管理评测集与回归评测。Agent Server 托管部署、并发与运行状态。	LangSmith Plus 美国标价 39 美元/席/月，轨迹和托管运行等按量计费。
Cloudflare	Agents SDK / Workers / Durable Objects	在边缘运行具备持久状态的 Agent，支持消息与定时触发、实时连接及会话恢复。	SDK 开源。Workers 与 Durable Objects 按请求、计算、状态存储等用量计费。
	Sandbox / Browser Run	隔离运行生成的脚本，并提供托管浏览器执行网页操作。两项可与外部 Harness 集成。	沙箱按计算资源计量，Browser Run 按浏览器运行时长及适用的并发计量。
	AI Gateway	对多模型 API 提供统一入口、缓存、速率限制和用量监控。	网关核心功能免费，日志导出等扩展能力按用量计费。
Modal	Sandboxes	按任务创建隔离环境，支持自定义镜像和 CPU/GPU，适用于 Agent 生成代码或运行测试。	计算资源按秒计费，按需扩缩容。
	Functions / GPU Jobs	将 Python 函数、服务和批处理任务部署到弹性 CPU/GPU 环境，处理任务高峰及并行作业。	
Temporal	Temporal SDK / Service	为跨系统流程保留事件历史，提供重试、等待外部信号及故障恢复。Worker 执行业务逻辑。	SDK 和 Service 开源，可自托管。
	Temporal Cloud	托管工作流状态、任务队列和调度服务，应用可自行部署 Worker。	按工作流 Actions（如启动、调度）与事件历史存储等用量计费。

数据来源：LangChain、Cloudflare、Modal、Temporal，国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

目录

01 头部模型领先优势显著，Agent Infra 有望加速企业 AI 拐点到来

1.1 复盘 2026：Coding 率先引爆 ARR，推理降本推动份额向双寡头集中

1.2 展望 2027（训练侧）：RSI 开启研发自迭代闭环，拉大前沿智能代际差

1.3 展望 2027（应用侧）：Token 降价推动 AI 从编程向垂直知识工作与个人 Agent 渗透

1.4 破局核心（基建侧）：以云为鉴，Agent Infra 补齐生产短板，巨头与独角兽加速布局

02 Anthropic：SOTA 模型构建商业化壁垒，核心收入来自企业客户

03 Anthropic：模型及产品矩阵丰富，覆盖全场景用户需求

七位联创持股12.63%，Anthropic公益基金持股6.47%

- **七位联创合计持有 16.53 亿股 A 类股。**Dario 和 Daniela 分别持 2.34 亿股和 2.35 亿股，其余五位各持约 2.36－2.38 亿股。亚马逊、谷歌及 Anthropic 全资公益基金分别持有 19.21 亿股、17.90 亿股和 8.47 亿股 N 类股。外部慈善机构的持股归入其他股东。联创各自承诺将 80% 的个人股权用于慈善，已转让的股份按实际持有人归类。

图：Anthropic 拟发行后股东持股（股数为亿股，占比为总股本比例）

Anthropic 拟发行后总股本 130.87 亿股 100.00%

七位联创（A 类） 12.63%	主要机构及公益基金（N 类） 34.82%	其他股东 52.55%	特殊股份 2 股 0.00%
Dario Amodei 2.34｜1.79%	亚马逊 19.21｜14.68%	其他 A 类 44.13%	Founder LLC（F 类）
Daniela Amodei 2.35｜1.80%	谷歌 17.90｜13.67%	其他 N 类 8.42%	1 股｜一般事项表决权 50.1%
Tom Brown 2.38｜1.82%	Anthropic 全资公益基金 8.47｜6.47%	其中外部慈善基金	长期利益信托（T 类）
Jack Clark 2.36｜1.80%		462.11 万股 A 类	1 股｜通常可选任 4 席
Jared Kaplan 2.36｜1.80%			
Sam McCandlish 2.37｜1.81%			
Chris Olah 2.37｜1.81%			

按股类汇总 A 类 56.76%，N 类 43.24%，F、T 类各 1 股

主要职能与负责人

模型研究与安全	算力与工程	产品与商业	公共事务与运营
Jared Kaplan 技术研究	Tom Brown 算力采购与扩容	Ami Vora Claude 产品	Jack Clark 公共利益研究
Sam McCandlish 大模型训练	Rahul Patil 推理、基础设施及安全工程	Paul Smith 商业战略与全球市场拓展	Tino Cuéllar 全球事务
Chris Olah 模型可解释性		Sasha de Marigny 品牌与传播	Brian Israel 法律与合规
			Krishna Rao 财务
			Hannah Pritchett 人事

数据来源：Anthropic、SEC，国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

七位联创在创办前已于OpenAI共事

- **早期合作集中在模型规模化研究和 GPT-3。**Dario 和 Daniela 是兄妹，分别负责研究和安全政策。Kaplan 与 McCandlish 牵头 Scaling Laws 研究，Brown 与 Dario 参与其中，四人后来又合作开发 GPT-3。Clark 负责政策，Olah 领导可解释性研究。七人于 2021 年 1 月共同创立 Anthropic。

表：Anthropic 联合创始人的分工与代表性履历

姓名	现任职务	在 Anthropic 的分工	代表性履历
Dario Amodei	首席执行官	主导研究方向与公司战略	曾任 OpenAI 研究副总裁，参与领导 GPT-2、GPT-3 研发。 此前在 Google Brain 从事研究，获普林斯顿大学生物物理学博士。
Daniela Amodei	总裁兼董事会主席	统筹研发、产品与经营，管理执行团队	曾任 OpenAI 安全与政策副总裁，也曾管理人事和工程团队。 加入 OpenAI 前，在 Stripe 从事风险管理。
Tom Brown	首席计算官	负责算力获取、扩容与利用效率	在 OpenAI 主导 GPT-3 工程，与 Dario 等合作研究 RLHF。 获麻省理工学院计算机科学方向工学硕士。
Jared Kaplan	首席科学官	负责技术研究及安全方向的基础研究	在 OpenAI 共同牵头 Scaling Laws 研究，参与 GPT-3。 曾任约翰斯·霍普金斯大学物理学副教授，获哈佛大学物理学博士。
Sam McCandlish	首席架构师	负责预训练、研发效率和强化学习基建	在 OpenAI 组建模型规模化研究团队，参与 GPT-3 研发。 曾任 Anthropic 首席技术官，获斯坦福大学物理学博士。
Chris Olah	可解释性研究负责人	研究模型内部机制	曾领导 OpenAI 的可解释性研究，此前在 Google Brain 任职。 参与创办机器学习期刊《Distill》，并担任编辑。
Jack Clark	公共利益负责人	领导 Anthropic Institute， 研究 AI 的社会影响	曾任 OpenAI 政策总监，负责与政府和监管机构沟通。 早年从事科技报道，后创办 AI 研究通讯《Import AI》。

数据来源：Anthropic、SEC、OpenAI，国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

执行团队引入企业软件与大型平台高管

- **Anthropic 将模型训练与规模化部署分开管理。**2025 年 10 月，前 Stripe 首席技术官 Rahul Patil 接手工程团队，联创 Sam McCandlish 转任首席架构师，继续负责大模型训练。2026 年 1 月，Ami Vora 开始领导产品团队，与 Patil 配合推动 Claude 扩大部署。曾负责 ServiceNow 全球客户业务的 Paul Smith 则负责市场拓展。

表：Anthropic 执行团队的分工与代表性履历

姓名	现任职务	主要职责	代表性履历
Rahul Patil	首席技术官	统筹推理、产品工程、基建与安全	曾任 Stripe 首席技术官，负责支付平台的工程体系。 此前在 AWS、微软和甲骨文云担任高级工程管理职务。
Ami Vora	首席产品官	领导 Claude 产品团队	曾任 Faire 首席产品官，负责面向零售商的产品。 在 Meta 任职超过 15 年，曾参与管理 WhatsApp 等产品。
Paul Smith	首席商务官	负责商业战略和全球市场拓展	曾任 ServiceNow 首席商务官，兼管全球客户与现场运营。 此前曾在 Salesforce 和微软担任高级管理职务。
Krishna Rao	首席财务官	负责财务战略与运营	曾任 Fanatics Commerce 和 Cedar 首席财务官。 此前在 Airbnb 负责公司发展和财务规划分析。
Mariano-Florentino (Tino) Cuéllar	首席全球事务官	负责全球政策、国际合作与政府关系	曾任卡内基国际和平基金会主席、加州最高法院法官。 2026 年曾担任 Anthropic 长期利益信托受托人。
Brian Israel	首席法律官	负责法律事务、诚信与合规	曾在美国国务院和 NASA 处理国际法、国家安全法等事务。 曾任航天科技公司 Planetary Resources 总法律顾问。
Sasha de Marigny	首席品牌与传播官	负责传播、品牌和产品营销	曾任 Thrive Capital 品牌与传播副总裁，为投资企业提供咨询。 此前负责 Stripe 全球品牌传播，并参与创办 Stripe Press。
Hannah Pritchett	首席人事官	负责招聘、组织发展与员工体验	曾任 Binti 人事副总裁，负责人才和业务运营。 此前在 Stripe 和 Facebook 担任运营及人事管理职务。

数据来源：Anthropic、SEC，国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

预计IPO后稀释后估值1.33万亿美元

- **E 至 H 轮融资额上升。**F 轮募资 130 亿美元，较 E 轮增 271.43%。H 轮投后估值较 G 轮增 5,850 亿美元。
- **IPO 拟发售 8.25 亿股，定价每股 70－78 美元。**按中值 74 美元，IPO 规模 610.5 亿美元，其中公司募资 593 亿美元。另有 260 亿美元同步私募。**两项发行完成后，稀释前市值（按已发行股份）约 9,684 亿美元，完全稀释后约 1.33 万亿美元。**

表：Anthropic 历轮融资

轮次	时间	融资金额（亿美元）	投后估值（亿美元）	领投方 / 主要投资者
Series A	2021年5月	1.24	—	Jaan Tallinn领投，James McClave、Dustin Moskovitz、新兴风险研究中心、Eric Schmidt 等参与
Series B	2022年4月	5.8	—	Sam Bankman-Fried领投，Caroline Ellison、Jim McClave、Nishad Singh、Jaan Tallinn以及新兴风险研究中心等参与
Series C	2023年5月	4.5	—	Spark Capital 领投，Google、Salesforce Ventures、Sound Ventures、Zoom Ventures等参与
Series D				
Series E	2025年3月	35	615	Lightspeed Venture Partners领投，Bessemer Venture Partners、Cisco Investments、D1 Capital Partners、Fidelity Management & Research Company、General Catalyst、Jane Street、Menlo Ventures和Salesforce Ventures等参与
Series F	2025年9月	130	1,830	ICONIQ 领投，Fidelity、Lightspeed 联合领投
Series G	2026年2月	300	3,800	GIC、Coatue 领投，D. E. Shaw、Dragoneer、Founders Fund、ICONIQ、MGX 联合领投
Series H	2026年5月	650	9,650	Altimeter、Dragoneer、Greenoaks、Sequoia Capital领投，Capital Group、Coatue、D1 Capital Partners、GIC、ICONIQ和XN联合领投

数据来源：Anthropic、SEC、Crunchbase、The Information，国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

单季经营利润转正

- **收入 5 个季度放大 30 倍：**营业收入从 2023 年 0.41 亿美元增至 2025 年 45.85 亿，2026H1 单半年即达 162.86 亿、同比 +1,300%；季度口径 25Q1 的 3.76 亿到 26Q2 的 115.55 亿，五个季度放大 30 倍。
- **26Q2 单季经营利润首次转正：**GAAP 经营亏损从 2025 年的 80.60 亿收窄至 2026H1 的 6.93 亿，26Q2 单季经营利润 9.74 亿、利润率 +8.4%；Non-GAAP 口径 2026H1 已盈利 16.18 亿（9.9%）。
- **净亏损主要是账面损失：**2026H1 净亏损 807.48 亿中，可转债、衍生负债与 SAFE 公允价值变动 -723.88 亿、其他收益（费用）净额 -87.57 亿，合计 811.45 亿，不构成经营性现金流出。

表：Anthropic 年度、半年度与季度损益（GAAP / Non-GAAP）

项目（百万美元）	2023	2024	2025	2025H1	2026H1	25Q1	25Q2	25Q3	25Q4	26Q1	26Q2
营业收入	41	386	4,585	1,163	16,286	376	787	1,382	2,040	4,731	11,555
同比增速	—	+841%	+1,088%	—	+1,300%	—	—	—	—	+1,158%	+1,368%
环比增速	—	—	—	—	—	—	+109%	+76%	+48%	+132%	+144%
算力及基础设施	433	2,522	7,326	2,334	8,945	944	1,390	2,047	2,945	3,387	5,558
人员费用	95	454	1,839	538	3,152	218	320	489	812	1,295	1,857
Bartz 和解费用	—	—	1,500	—	—	—	—	1,500	—	—	—
销售、营销及合作费用	2	43	568	120	3,168	24	96	202	246	1,033	2,135
其他运营及管理费用	62	347	1,412	306	1,714	125	181	233	873	683	1,031
总运营费用	592	3,366	12,645	3,298	16,979	1,311	1,987	4,471	4,876	6,398	10,581
经营利润（亏损）	(551)	(2,980)	(8,060)	(2,135)	(693)	(935)	(1,200)	(3,089)	(2,836)	(1,667)	974
经营利润率	-1,344%	-772%	-176%	-184%	-4%	-249%	-153%	-224%	-139%	-35%	+8%
经调整经营利润（Non-GAAP）	(522)	(2,788)	(4,854)	(1,849)	1,618	(830)	(1,019)	(1,284)	(1,721)	(617)	2,235
经调整经营利润率	-1,273%	-722%	-106%	-159%	+10%	-221%	-130%	-93%	-84%	-13%	+19%
净利润（亏损）	(1,531)	(8,309)	(41,965)	(12,969)	(80,748)	(6,592)	(6,377)	(12,138)	(16,858)	(14,439)	(66,309)

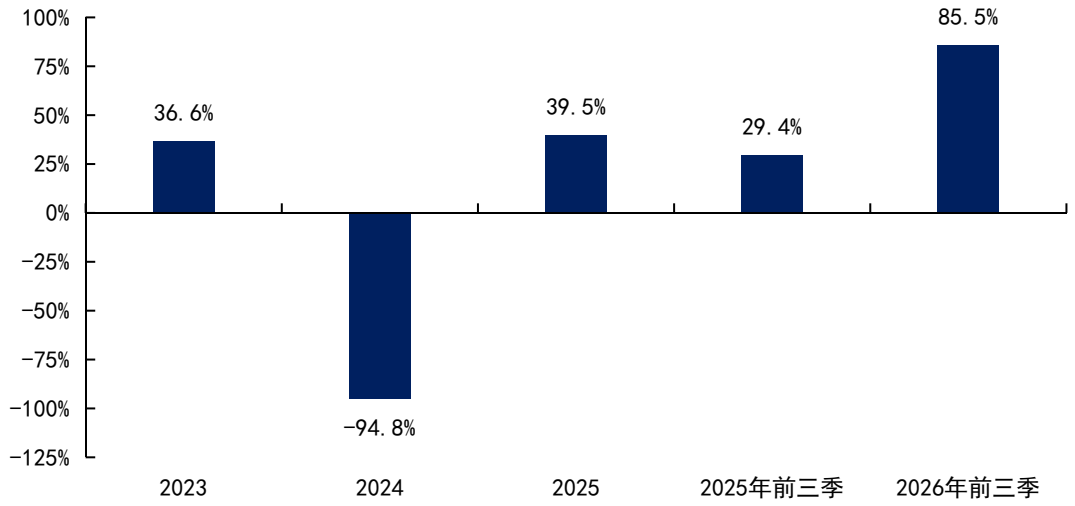
数据来源：Anthropic、SEC，国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

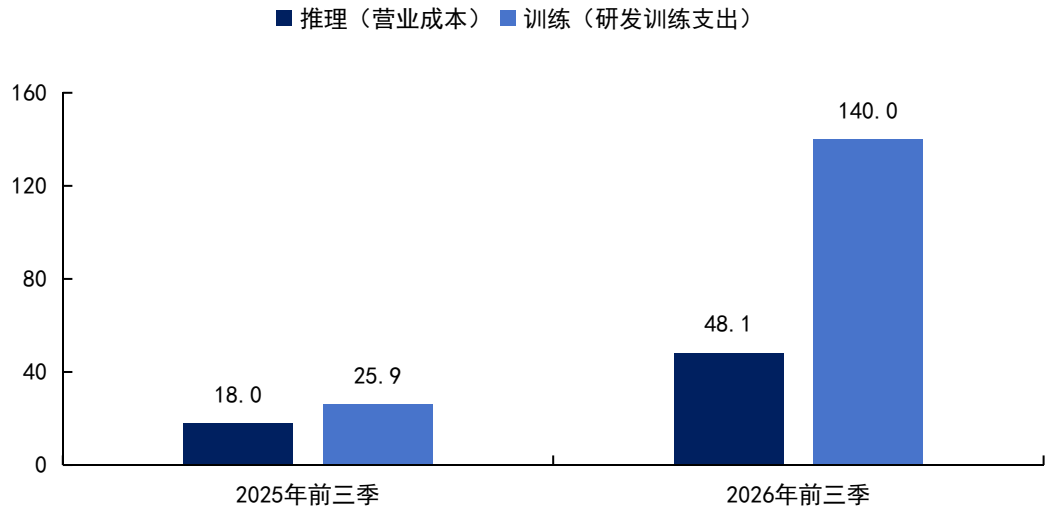
前三季度毛利率升至85.5%，算力增量主要投向训练

- **收入增长快于营业成本，毛利率同比提升56.1个百分点。**2026年前三季度，收入由上年同期的25.45亿增至332.42亿美元，而营业成本由17.96亿增至48.12亿美元。
- **算力投入进一步向训练倾斜。**2026年前三季度训练支出增至139.96亿美元，占算力及基础设施费用的74.4%，同比提升15.4个百分点。以营业成本近似推理支出，训练与推理的支出之比接近3：1。

图：毛利率（%）



图：前三季度推理与训练算力支出（亿美元）



数据来源：Anthropic、SEC，国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

云厂商借既有客户和采购体系帮Claude获客

- **云渠道减少了企业采购和接入Claude的工作。**企业直连官方API，由Anthropic开通并结算；走云渠道，可沿用AWS、Google Cloud或Azure的账户、权限和账单。云厂商也能向既有客户推荐Claude，扩大Anthropic的获客范围。
- **云市场贡献44%的收入，渠道分成是这一路径的成本。**2026年前三季度，云市场收入约146亿美元，按模型估算，约35%的云渠道收入用于分成。Anthropic按总额确认云渠道收入，向合作伙伴支付的分成计入运营费用。

表：云渠道收入及分成

指标	2023	2024	2025	2026年前三季度
营业收入（亿美元）	0.41	3.86	45.85	332.42
云渠道收入占比	11%	32%	47%	44%
云渠道收入（亿美元，推算）	0.05	1.24	21.55	146.26
渠道分成（亿美元，估算）			3.45	50.59
渠道分成 / 云渠道收入			16.0%	34.6%
渠道分成 / 营业收入			7.5%	15.2%

数据来源：Anthropic、SEC、AWS，国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

大客户超1.19万家，前五大客户占比降至21%

- **大客户数量上升，存量客户也在增加用量。**2026年9月末，大客户超过1.19万家，上年同期超过950家；其中逾1,900家过去12个月收入达100万美元，约250家达1,000万美元。大客户前三季度贡献245亿美元，占总收入74%。截至9月底，净收入留存率约900%。
- **前五大客户占比下降，收入仍在增长。**2026年一季度至三季度，前五大客户的收入占比由约36%降至约21%，三季度前五大客户各自贡献的收入均创新高。前三季度前五大客户合计占约25%，单一客户均未超过10%。

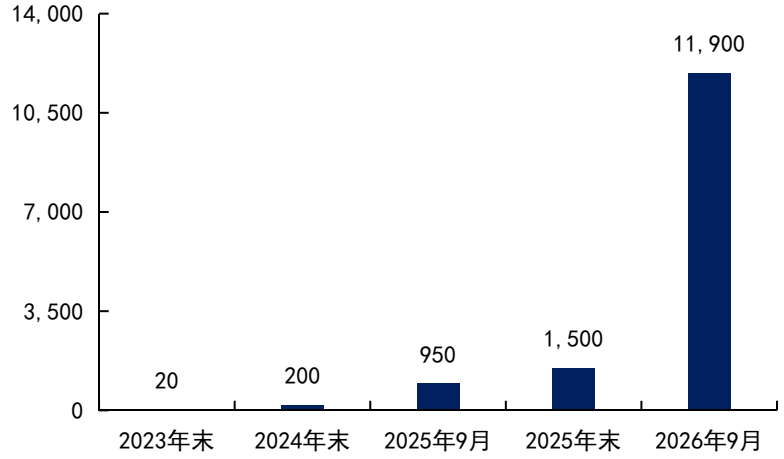
表：大客户贡献与收入结构

指标	2023	2024	2025	2025年前三季度	2026年前三季度
大客户收入（亿美元）	0.37	2.20	35.00	19.00	245.00
大客户收入占比	90%	57%	76%	75%	74%
按量付费收入占比	95%	75%	83%	83%	79%
订阅收入占比	5%	25%	17%	17%	21%
前五大客户收入占比					**约25%**

大客户按期末过去12个月收入不少于10万美元认定；年度客户数为约数，9月末客户数为披露下限。

前五大客户单季占比按100%减去其余客户的收入占比计算。

图：期末大客户数（家）



数据来源：Anthropic、SEC，国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

咨询与IT服务商把Claude带进客户系统

- **云市场提供采购入口。**AWS、Google Cloud和Azure提供原生访问，企业可沿用现有云采购流程。2026年前三季度，约44%的收入来自云市场销售。
- **咨询和IT服务商负责集成落地。**2026年3月推出Claude Partner Network，首期承诺投入1亿美元用于培训、认证和联合拓客。截至9月底，逾4,000家机构的15万余名从业人员获得认证。

代表性咨询与IT服务伙伴的部署与获客方式

合作伙伴	内部部署与人员培养	如何触达外部企业客户
Accenture	成立Anthropic Business Group，计划培训约3万人。	两家共同向CIO提供咨询及交付，覆盖金融、医疗和公共部门等受监管行业。
Deloitte	计划向全球47万名员工开放Claude。	以内部部署积累经验，与Anthropic联合为受监管行业提供实施和合规支持。
PwC	在美国扩大Claude Code和Cowork使用，计划培训和认证约3万人。	成立联合卓越中心，共同销售及交付企业技术改造项目。
KPMG	向全球逾27.6万名员工开放Claude，并接入Digital Gateway。	通过既有平台触达税务与法律客户，并向私募股权机构及被投企业推广。
DXC	计划培养数万名Claude认证的前线部署工程师。	利用现有托管和应用运维关系，将Claude接入保险及其他客户的核心系统。
TCS	计划在56个国家为5万名员工提供Claude相关培训。	共同为金融、医疗和公共部门客户开发和部署行业应用。
Cognizant	已培训逾3万人，并将Claude集成到自有交付平台。	以全球高级合作伙伴身份，面向制造、保险等客户提供行业方案。

员工开放范围、培训计划和已培训人数分别列示。伙伴公告统计时点不同。

数据来源：Anthropic、SEC，国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

Anthropic用FDE推动重点客户从试点走向生产

- **自有工程师与客户共建。**Anthropic将前线部署工程师（FDE）放在Applied AI团队，工程师与客户一起选择高价值场景、接入系统、建立评估并通过安全审查，完成生产上线后把维护和扩展交给客户。
- **用培训和合资交付扩大覆盖。**Frontier Academy培养客户和伙伴自己的部署工程师，Ode则承接需要长期实施支持的中型企业项目。

Anthropic的FDE团队、人才培养和交付布局

布局	如何执行	进展或实例
自有团队 共同部署重点项目	与客户共同确定场景和成功标准，搭建原型和评估，将Claude接入实际业务系统。上线后完成知识转移，并把部署中遇到的问题反馈给产品团队。	FIS与Anthropic共同开发银行反洗钱调查助手。FIS在自有私有云上交付，并于2026年10月面向美国银行商业化。
Frontier Academy 培养客户与伙伴	客户和伙伴提名有实际项目的工程师。通过模拟部署考核后，学员回原单位完成12周实战，在Anthropic工程师支持下推动真实场景上线。	2026年10月宣布投入1亿美元，目标在2027年底前培养1万名客户及伙伴工程师。
Ode 为中型企业交付	Anthropic与Blackstone、Hellman & Friedman等共同发起企业AI服务公司，由实施团队嵌入客户流程，承接项目上线及后续扩展。	2026年7月正式推出。部署团队包括来自Fractional AI及Anthropic的工程师。

数据来源：Anthropic、FIS、Ode，国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

Anthropic算力及基建承诺9,009亿美元，长期自建算力占比持续提升

国信证券 GUOSEN SECURITIES

- **2030年底前预计支出约4,516亿美元。** 2026年9月末起12个月内预计支付约707亿美元。总承诺较11月2日版增加55亿美元。
- **Anthropic Direct 把机房和芯片分开签约。** 公司先锁定交付周期更长的机房和电力，待需求明确后再采购或融资租赁芯片，部署在由合作方运营的专用机房。机房投产后，长期算力可逐步替换高峰期采购的高价短租算力。
- **管理层预计未来几年 Direct 在整体算力中的占比提高。此外，**随着效率提高、Direct 扩张和更多算力转向推理，长期算力及基建费用有望趋近收入的45%－50%。

表：截至2026年9月末已固定、可确定的付款义务（亿美元）

项目	2026年余下	2027年	2028年	2029年	2030年	2031年及以后	合计
经营租赁	0.57	3.03	4.62	4.45	4.66	24.91	42.24
融资租赁	1.21	4.38	4.22	4.06	3.90	1.73	19.50
采购承诺	50.53	284.24	434.73	469.01	442.14	1,457.54	3,138.19
逐年付款合计	52.31	291.65	443.57	477.52	450.70	1,484.18	3,199.93

表：主要具名算力与基建合同（亿美元）

交易对手	供给方式	期限或交付	金额	付款条件与可取消性
Microsoft	云算力服务	2026/11—2033/06	472－492	无论用量均须支付。重大违约或部分批次未交付可解除。
Google Cloud	云算力服务	2026/04—2033/07	≥1,213	用量不足补差，重大违约除外。
Amazon AWS	云算力服务	2026/05—2036/04	≥1,100	用量不足补差，重大违约除外。
SpaceXAI	GPU算力服务	至2029年	最高1,205	满产约31亿/月。初始锁定期结束后可提前90天通知解除。
Nscale	GPU算力服务	至2033年	最高446	验收后计费。未达交付或可用性要求可解除。
Lambda	GPU算力服务	至2034年	约413	验收后计费。其中349亿的协议目前可解除，双方尚未发出通知。
博通及融资方	**Ant Direct：设备**	五年期设备租赁	约1,612	TPU融资租赁，按季付款。金额不含利息。
Fluidstack	**Ant Direct：机房**	开发及运营	未单列	负责专用数据中心开发与代运营。
额外TPU	云服务或直购/直租	2028－2031年交付	预计1,250－1,500	未约定固定金额。交付路径待定，可采用 Ant Direct。
AMD Helios	直购或第三方服务	首批容量2027年	预计超200	无固定金额，按首批容量全量采购估算。交付路径待定。
表列合同金额小计			**6,461－6,481**	按表列金额相加，含下限、上限及约定额，剔除未单列及无固定金额项目。

数据来源：Anthropic、SEC，国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

目录

01 头部模型领先优势显著，Agent Infra 有望加速企业 AI 拐点到来

1.1 复盘 2026：Coding 率先引爆 ARR，推理降本推动份额向双寡头集中

1.2 展望 2027（训练侧）：RSI 开启研发自迭代闭环，拉大前沿智能代际差

1.3 展望 2027（应用侧）：Token 降价推动 AI 从编程向垂直知识工作与个人 Agent 渗透

1.4 破局核心（基建侧）：以云为鉴，Agent Infra 补齐生产短板，巨头与独角兽加速布局

02 Anthropic：SOTA 模型构建商业化壁垒，核心收入来自企业客户

03 **Anthropic：模型及产品矩阵丰富，覆盖全场景用户需求**

产品矩阵：覆盖从编程扩展到办公与专业工作

- **Claude 已有三条主要交付路径。**员工可在 Claude 里委托工作，也可直接在 Office、Slack 等现有工具中使用 Claude。开发团队则通过 API 或托管 Agent，把 Claude 嵌入自己的业务系统。
- **办公场景的增量来自可编辑产物。**11 月 10 日推出 Sheets、Whiteboards 公测，Claude Data 同日正式开放。用户可以在同一对话里继续改表格、看板和演示文稿。

表：Anthropic 产品矩阵（截至 2026 年 11 月 12 日）

层级	代表产品	定位和交付形态
通用与办公	Claude Chat、Cowork、Claude Duet	Chat 用于问答研究。Cowork 跨工具委托任务。Duet 主动给出简报并处理待办，目前在美国邀请预览。
编程与专业	Claude Code、Claude Science、Claude Security	从代码开发扩展到科研工作台与代码安全审查。Science 为公开 Beta，Security 为 Enterprise Beta。
内置产物	Docs、Slides、Design、Data、Sheets、Whiteboards、Motion	在 Claude 中生成、直接编辑并共享。Sheets、Whiteboards、Motion 处于 Beta。
嵌入式入口	Microsoft 365、Google Workspace、Chrome、Slack（Claude Tag）	在办公套件原文件、浏览器和 Slack 线程中使用 Claude。Google Workspace 与 Tag 处于 Beta。
开发与扩展	Claude Platform、Managed Agents、Marketplace	API 提供模型调用，Managed Agents（Beta）托管任务运行。Connectors、Skills、Plugins 接入数据和流程。
组织采购	Team、Enterprise	企业席位覆盖 Chat、Code、Cowork 等，支持集中管理。Enterprise 官网自助方案按席位加用量计费。

数据来源：Anthropic、Claude Help Center，国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

行业解决方案：把 Claude 接进行业数据和流程

- **行业数据和验收规则进入现有工具。**金融团队可将行情和财报带入 Excel，沿用原有财务口径和公式。法务可在 Word 里按内部规则逐条审阅合同，并把修订记录留给同事复核。
- **受监管场景配有专用部署。**医疗可采用符合条件的 HIPAA-ready 配置，美国政府可使用 FedRAMP High 环境。企业还可通过 API 和托管 Agent 把工作流接入自己的系统。

表：Claude 主要行业和部门解决方案

行业 / 部门	代表任务	数据、工具与交付
金融 / 财富管理	财报与市场研究、尽调、对账与月结、投顾会前准备	金融数据和托管平台连接器，Excel、PowerPoint 与预构建金融 Agent
法律	合同条款审阅、尽调、法规跟踪和请求分流	法律 Artifacts 与插件，接入 iManage 等，Word 保留修订痕迹
医疗 / 生命科学	预授权材料、理赔申诉、临床文书、研究数据分析	医疗提供 HIPAA-ready 配置。科研接入 Claude Science 和专业数据库
销售 / 客服	客户研究、销售预测、工单分流、知识库问答	连接 CRM 和支持系统，支持模板插件或 API 自建支持 Agent
网络安全	自有代码库安全审查、漏洞修复建议与复核	Claude Security（Enterprise Beta），接入代码审查和修复流程
政府 / 教育	公文与案例整理、软件现代化、教学与校务	美国政府有 FedRAMP High 环境，高校和 K－12 提供不同方案
中小企 / 非营利	账务、销售和运营、筹款、市场及招聘	预置插件及工作流，连接 QuickBooks、Shopify 等现有工具

数据来源：Anthropic，国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

三款编程 Agent 的差异在任务组织方式

- **三款产品都能在本地或云端改代码和跑测试。**Claude Code 便于把团队规则嵌入仓库工作流，Codex 便于隔离并行任务和逐个审查，Cursor 把手写、补全和委托 Agent 放在同一编辑器。
- **Harness 的侧重点可从工具接口和任务状态看出。**Claude Code 以仓库指令和 Hooks 约束执行，Codex 开放线程与审批接口，Cursor 将编辑器上下文带入本地及云端 Agent。

表：Claude Code、Codex、Cursor 的功能和 Harness 对比

维度	Claude Code	OpenAI Codex	Cursor
主入口	终端 CLI、IDE 扩展、桌面及网页。以对话交代仓库级任务。	桌面 Codex 视图、CLI、IDE 与云端。桌面集中管理多个任务和差异。	AI 编辑器与 Agents Window 并列。Tab 补全、内联修改和 Agent 连续使用。
上下文	按需读文件、搜索和执行命令。CLAUDE.md 注入规则，长会话自动压缩。	AGENTS.md、Skills 和 MCP 注入项目上下文。支持恢复线程和压缩上下文。	当前文件、选区及项目 Rules 提供上下文。Agent 通过 Instant Grep 和文件读取查找代码。
执行与验收	搜索、修改、测试循环。Hooks 绑定工具事件，子 Agent 可分拆工作。	本地 worktree 隔离改动，云端可并行跑任务。通过 diff、测试结果与 PR 审查验收。	可在本地、worktree、SSH 或云端运行。浏览器 Design Mode 可标注页面元素再迭代。
Harness 侧重	把仓库规则、命令行工具和可配置权限嵌入开发流程。Agent SDK 可复用执行能力。	开源 CLI、SDK 和 App Server 提供任务状态、事件流、工具、沙箱及审批接口。	编辑器提供上下文，支持自研 Composer 与第三方模型。Cloud Projects 负责分派并汇总子任务。

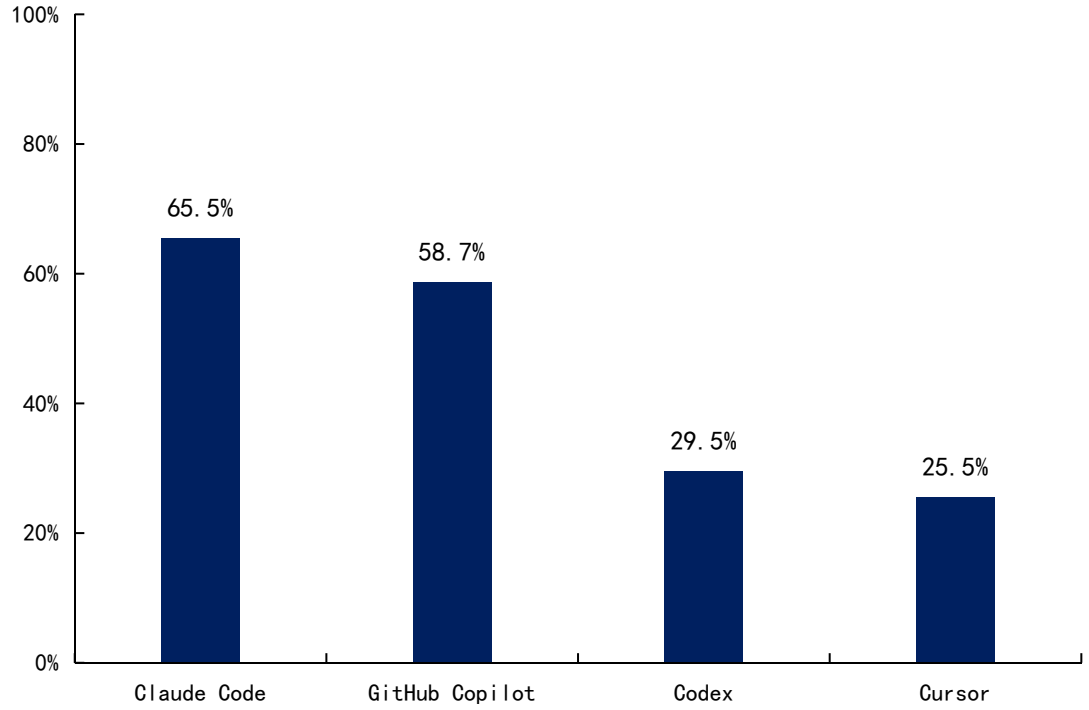
数据来源：Claude Code Docs、OpenAI Developers、ChatGPT Learn、Cursor Docs，国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

Claude Code 在两组采用率样本中领先

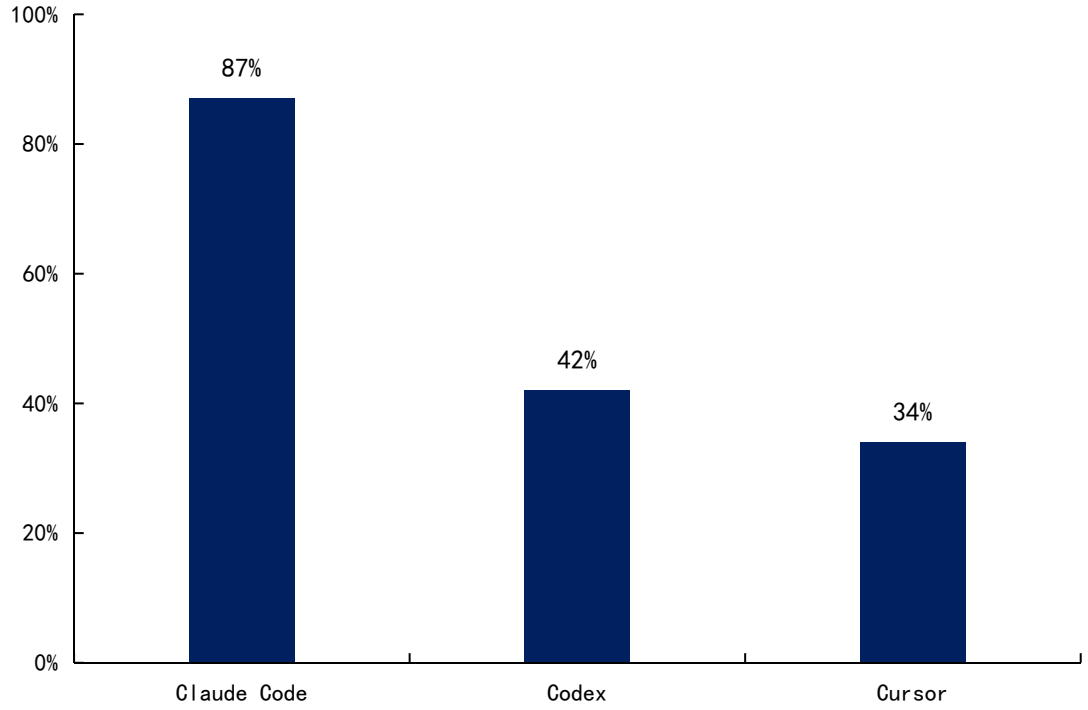
- **Claude Code 的开发者触达与企业采购均领先。**Stack Overflow 样本中，Claude Code 过去一年使用率为 65.5%。Ramp 的编程工具采购企业中，Claude Code 渗透率达 87%。
- **Codex 的企业采购覆盖在追赶。**Ramp 样本中 Codex 达 42%，高于 Cursor 的 34%，月均增 2.1 个百分点。

图：开发者过去一年使用率（%，可多选）



数据来源：Stack Overflow，国信证券经济研究所整理

图：Ramp 编程 Agent 采购企业渗透率（%）



数据来源：Ramp，国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

Cowork 与 Chatgpt Work 的差异落在交付和触发

- **两者都能跨文件和工具把任务做完。**Cowork 正把任务委托并入 Claude 对话，同时扩展原生 Artifacts 和 Office 插件。ChatGPT Work 保留独立工作模式，并可把持续更新的产物沉淀在 Space。
- **两者都可云端运行并调用获授权的本地资源。**Cowork 文档明确优先走连接器，其次浏览器，最后使用桌面 UI。ChatGPT Work 公布了云端、桌面及混合执行的权限边界，并支持应用事件触发。

表：Claude Cowork 与 ChatGPT Work 的功能和 Harness 对比

维度	Claude Cowork	ChatGPT Work
入口和交互	Chat 与 Cowork 逐步合并，直接在 Claude 对话交代任务，跨桌面、网页、手机继续。	在 ChatGPT 选 Work 模式委托任务，跟进执行、调整方向和审批。可将任务和素材放进 Project。
编辑与交付	内置 Docs、Slides、Sheets、Data 等 Artifacts。通过 Microsoft 365 插件在原文件中继续工作。	交付可审阅的文档、表格、演示等。Space 可组织 Pages、Slides、Spreadsheets、文件和共享内容。
任务触发	一次性委托和云端定时任务。可把常用工作流程保存为 Skill 或装成插件。	支持一次性、定时、监测及合资格的应用事件触发，如新邮件、Slack 消息、GitHub PR 活动。
执行与权限	云端每会话独立沙箱。连接器授权留在服务端，本地授权目录和浏览器经在线桌面端接入。	云容器或桌面执行，获授权时可由云端协调并调用在线电脑。连接器、浏览器、文件和网络按策略授权。
Harness 侧重	支持子 Agent 并行。优先连接器，再走浏览器，必要时用桌面 UI。将技能和产物放回 Claude 对话。	与 Codex 共享核心执行、隔离和权限机制。围绕插件、技能、子 Agent 和任务触发组织办公交付。

数据来源：Claude Help Center、ChatGPT Learn，国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

最前沿模型：Fable 5.2 的领先集中在科学与高难推理

- **旗舰溢价主要对应更难的科学和推理任务。**Fable 5.2 在 AA Intelligence Index v5.0.1 得 65 分，位列其可比榜单第 1。较 Fable 5.1 同档提升 16 分，科学终端任务得分由 43.3% 升至 73.3%。
- **常规编程与知识工作仍可交给 Opus。**Fable 5.2 与 Opus 5.5 的 Terminal-Bench 4.0 同为 59.6%，AA-Briefcase 接近。Fable 与 Mythos 共享底层模型，Mythos 仅向获批机构开放。

表：Fable 5.2 与前代旗舰及 Opus 的可比评测

评测	任务类型	Fable 5.2	Fable 5.1	Opus 5.5
AA Intelligence Index	综合能力（分）	65	49	59
Terminal-Bench-Science 0.1	科学研究任务	73.3%	43.3%	59.0%
HLE-Diamond	高难知识和推理	80.2%	60.1%	67.9%
FrontierCode-AA 2.0	真实代码库工程	48.8%	31.6%	46.4%
Terminal-Bench 4.0	终端编程任务	59.6%	52.0%	59.6%
AA-Briefcase v1.1	知识工作（Elo）	1785	1669	1801

数据来源：Artificial Analysis、Anthropic，国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

模型矩阵针对不同任务需求覆盖多价格带

- **Opus 的综合评分接近旗舰，单任务成本低 62%。**AA 样本中，Opus 5.5 Max 平均 10.04 美元、得 59 分，可作为跨文件重构、专业分析和文档交付的主力模型。
- **Sonnet 的成本取决于推理档位，Haiku 适合高频任务。**Sonnet 5.5 Xhigh 为 4.09 美元/任务，Max 升至 10.02 美元。Haiku 5.5 Low 平均只需 0.04 美元、2.1 分钟，可承接分流和高频摘要。

表：标准 API 单价、AA 评分与任务成本

模型	推理档位	输入/输出价	AA 分数	AA 美元/任务	适用场景
Fable 5.2	Max	8/40	65	26.54	高难科学推理、长程研究及复杂迁移
Opus 5.5	Max	4/20	59	10.04	复杂编程和专业知识工作主力
Sonnet 5.5	Max	2/10	55	10.02	提高推理强度后接近 Opus 的任务成本
Sonnet 5.5	Xhigh	2/10	50	4.09	在成本受限时承接常规编程与办公 Agent
Haiku 5.5	Max	0.10/0.50	36	0.41	低单价下为简单任务提供更强推理档位
Haiku 5.5	Low	0.10/0.50	19	0.04	低延迟分流、摘要和高频子任务

数据来源：Anthropic、Artificial Analysis，国信证券经济研究所整理

请务必阅读正文之后的免责声明及其项下所有内容

风险提示

第一，宏观经济波动。若宏观经济波动，公司业务、产业变革及新技术的落地节奏或将受到影响。

第二，下游需求不及预期。若下游AI需求不及预期，相关的AI研发投入增长或慢于预期，致使行业增长不及预期。

第三，核心技术水平升级不及预期的风险。AI大模型研发进度落后，AIGC相关产业技术壁垒较高，核心技术难以突破，影响整体进度。

第四，AI快速迭代、平权化下竞争加剧，影响云业务利润率。

请务必阅读正文之后的免责声明及其项下所有内容

免责声明

国信证券 GUOSEN SECURITIES

国信证券投资评级

投资评级标准	类别	级别	说明
报告中投资建议所涉及的评级（如有）分为股票评级和行业评级（另有说明的除外）。评级标准为报告发布日后6到12个月内的相对市场表现，也即报告发布日后的6到12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（000300.SH）作为基准；新三板市场以三板成指（899001.CSI）为基准；香港市场以恒生指数(HSI.HI)作为基准；美国市场以标普500指数(SPX.GI)或纳斯达克指数（IXIC.GI）为基准。	股票投资评级	优于大市	股价表现优于市场代表性指数10%以上
		中性	股价表现介于市场代表性指数±10%之间
		弱于大市	股价表现弱于市场代表性指数10%以上
		无评级	股价与市场代表性指数相比无明确观点
	行业投资评级	优于大市	行业指数表现优于市场代表性指数10%以上
		中性	行业指数表现介于市场代表性指数±10%之间
		弱于大市	行业指数表现弱于市场代表性指数10%以上

分析师承诺

作者保证报告所采用的数据均来自合规渠道；分析逻辑基于作者的职业理解，通过合理判断并得出结论，力求独立、客观、公正，结论不受任何第三方的授意或影响；作者在过去、现在或未来未就其研究报告所提供的具体建议或所表述的意见直接或间接收取任何报酬，特此声明。

重要声明

本报告由国信证券股份有限公司（已具备中国证监会许可的证券投资咨询业务资格）制作；报告版权归国信证券股份有限公司（以下简称“我公司”）所有。本报告仅供我公司客户使用，本公司不会因接收人收到本报告而视其为客户。未经书面许可，任何机构和个人不得以任何形式使用、复制或传播。任何有关本报告的摘要或节选都不代表本报告正式完整的观点，一切须以我公司向客户发布的本报告完整版本为准。

本报告基于已公开的资料或信息撰写，但我公司不保证该资料及信息的完整性、准确性。本报告所载的信息、资料、建议及推测仅反映我公司于本报告公开发布当日的判断，在不同时期，我公司可能撰写并发布与本报告所载资料、建议及推测不一致的报告。我公司不保证本报告所含信息及资料处于最新状态；我公司可能随时补充、更新和修订有关信息及资料，投资者应当自行关注相关更新和修订内容。我公司或关联机构可能会持有本报告中所提到的公司所发行的证券并进行交易，还可能为这些公司提供或争取提供投资银行、财务顾问或金融产品等相关服务。本公司的资产管理部门、自营部门以及其他投资业务部门可能独立做出与本报告中意见或建议不一致的投资决策。

本报告仅供参考之用，不构成出售或购买证券或其他投资标的要约或邀请。在任何情况下，本报告中的信息和意见均不构成对任何个人的投资建议。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。投资者应结合自己的投资目标和财务状况自行判断是否采用本报告所载内容和信息并自行承担风险，我公司及雇员对投资者使用本报告及其内容而造成的一切后果不承担任何法律责任。

证券投资咨询业务的说明

本公司具备中国证监会核准的证券投资咨询业务资格。证券投资咨询，是指从事证券投资咨询业务的机构及其投资咨询人员以下列形式为证券投资人或者客户提供证券投资分析、预测或者建议等直接或者间接有偿咨询服务的活动：接受投资人或者客户委托，提供证券投资咨询服务；举办有关证券投资咨询的讲座、报告会、分析会等；在报刊上发表证券投资咨询的文章、评论、报告，以及通过电台、电视台等公众传播媒体提供证券投资咨询服务；通过电话、传真、电脑网络等电信设备系统，提供证券投资咨询服务；中国证监会认定的其他形式。

发布证券研究报告是证券投资咨询业务的一种基本形式，指证券公司、证券投资咨询机构对证券及证券相关产品的价值、市场走势或者相关影响因素进行分析，形成证券估值、投资评级等投资分析意见，制作证券研究报告，并向客户发布的行为。

国信证券
GUOSEN SECURITIES

国信证券经济研究所

深圳

深圳市福田区福华一路125号国信金融大厦36层

邮编：518046　　总机：0755-82130833

上海

上海浦东民生路1199弄证大五道口广场1号楼12楼

邮编：200135

北京

北京西城区金融大街兴盛街6号国信证券9层

邮编：100032