仅供内部参考，请勿外传

计算机

报告日期：2026 年 10 月 17 日

Anthropic 三季报解读

—— 2026.10.06—2026.10.16 大模型商业化周报

投资要点

- **本期重点关注 Anthropic 商业化进展、物联网设备上的模型部署与安全，以及个人购物和生活服务拓展。**
- **Anthropic 单位算力收入改善，长期基础设施承诺至少 5180 亿美元。**
  2026 年前三季度收入分别为 47 亿、116 亿和 168 亿美元，第三季度环比增长 44.8%。第三季度每 1 美元算力支出对应约 1.7 美元收入，高于一季度按披露比例换算的约 1.41 美元。本周披露毛利率 85%、算力利润率超过 40%，前三季度调整后经营利润率约 11%，自由现金流转正。Anthropic 年化收入运行率从 2025 年底 90 亿美元升至 7 月 650 亿美元，年化消费超过 100 万美元的客户从 2 月超过 500 家增至 4 月超过 1000 家。公司未来十年基础设施承诺至少 5180 亿美元，约 80%不可取消或需按约付款，长期合同扩大后续算力供给。部分投资者本周将 IPO 估值预期由约 2 万亿美元调整至约 1.5 万亿美元。
- **模型部署延伸至物联网设备，端侧适配与安全管理需求同步拓展。**
  Anthropic 用 15 台约 55 美元/台的电视盒子协同运行 260 亿参数模型，90 亿参数模型在 9 类硬件中的 8 类上成功搭建推理服务。新实例继承前序笔记后，部署成功率由 45%提高至后续各代的 61%—84%，表明环境识别、依赖配置和排错经验能够在设备部署中复用。低成本设备协同推理为利用分散算力提供实验路径，相关能力可延伸至模型部署、推理调度和设备运维。与此同时，分散运行的模型副本难以通过停用单一云端账户清除，防护需要覆盖模型加载、执行权限与跨设备通信，网络检测、终端防护及自动化安全运营的协同需求有望拓展。
- **个人 Agent 拓展购物与生活服务，消费者入口与商家工具进一步衔接。**
  Google 将超过 40 项应用接入 Gemini，包括生鲜杂货配送平台 Instacart、餐厅预订平台 OpenTable 等，并以 Universal Cart 连接不同商家的商品、库存、价格与支付。千问 Go 演示航班理赔、汽车养护和场馆预约，个人助理由提供建议向联系服务商、安排具体事务延伸。千问办公电商版则连接跨平台经营分析、投放优化与素材制作，补充商家侧的服务供给。购物与生活服务需要持续执行多个步骤，应用接入和后台任务能力有助于减少用户在平台间切换；消费者入口与商家工具的衔接，为个人 Agent 进入交易和服务流程提供更完整的产品路径。
- **风险提示**
  商业化增长及盈利改善不及预期；端侧部署及安全产品落地不及预期；个人 Agent 服务拓展不及预期；行业竞争加剧；数据安全与合规风险。

行业评级：看好(维持)

分析师：刘雯蜀
执业证书号：S1230523020002
liuwenshu03@stocke.com.cn

分析师：郑毅
执业证书号：S1230524070002
zhengyi@stocke.com.cn

相关报告

1 《端侧 AI，万众创新》 2026.10.10

2 《智谱接入 AWS 拓展企业渠道》 2026.10.07

3 《从 Muse 看端侧 AI 及核心元器件机遇》 2026.09.28

仅供内部参考，请勿外传

正文目录

1 本周模型发布、商业化数据总结 4
2 本周重点事件 5
2.1 Anthropic 商业化进展更新 5
2.1.1 第三季度收入达 168 亿美元，前三季度调整后经营利润率约 11% 5
2.1.2 历史收入运行率与编程业务年化收入 6
2.1.3 客户扩张与收入结构 6
2.1.4 历史损益 8
2.1.5 融资规模及长期算力合同 8
2.2 从自我复制研究看物联网端侧部署与安全需求 9
2.3 个人 Agent 拓展购物和生活服务 10
3 行业信息 11
4 各厂商其他动态 12
4.1 Anthropic 12
4.2 OpenAI 15
4.3 Gemini 16
4.4 字节跳动 18
4.5 阿里巴巴 19
4.6 智谱 19
4.7 MiniMax 20
4.8 Kimi 21
4.9 DeepSeek 21
5 风险提示 22

仅供内部参考，请勿外传

图表目录

图 1： 收入规模与季度环比增速 5
图 2： IPO 估值讨论变化（亿美元） 6
图 3： 年化收入运行率及第三方估算（亿美元） 6
图 4： Claude Code 年化收入（亿美元） 6
图 5： Ramp 企业 AI 付费采用率 7
图 6： YipitData 客户支出集中度 7
图 7： 2025 年收入结构（亿美元） 7
图 8： 云渠道收入占比（%） 7
图 9： 订阅与 API 的收入和算力占比（机构估算） 7
图 10： 收入扩张与经营亏损（亿美元） 8
图 11： 2025 年经营费用结构（亿美元） 8
图 12： 融资规模与投后估值（亿美元） 8
图 13： 基础设施合同承诺（亿美元） 9
图 14： 自主复制研究的四个环节：寻找目标、获得执行、建立推理服务、继续传播 10
图 15： 继承前代部署经验后，Qwen3.5-9B 的推理服务部署成功率由 45%提高至 61%—84% 10
图 16： Gemini 应用连接超过 40 项服务，覆盖购物、预订、创作及经营工具 10
图 17： 千问 Go 演示界面 11
图 18： 千问办公电商版演示界面 11
图 19： Ramp 客户 Token 使用量与支出对比（原图指数口径） 12
图 20： Sonnet 5.5 与 Haiku 5.5 官方 API 价格（$/百万 Token） 12
图 21： Claude Haiku 5.5 与上一代及同类模型的能力对照 12
图 22： Claude Dashboards 演示界面 13
图 23： Claude Motion：将共享单车运营分析生成动态城市演示 13
图 24： Anthropic 每日代码变更合并数：9 月单日峰值超过 3 万次 14
图 25： 完整推理与推理摘要的蒸馏训练效果对比 14
图 26： 防护实施前后，被提取推理 Token 规模变化 14
图 27： Decisions API 官方功能示例：根据商品照片判断是否存在凹陷，并返回概率 15
图 28： Gemini 4 Argon 主要能力对照（Google 官方发布材料） 17
图 29： Vals Index：模型在编程、金融、法律和税务等专业任务中的综合表现 17
图 30： Argon 通宵优化 ADK 代码库的执行流程 17
图 31： ADK 留出测试集的优化前后表现 17
图 32： 豆包工作画布：将参考素材与多个设计方案集中展示并继续调整 18
图 33： GLM-Legal 在法律任务评测中的表现（Harvey LAB、LegalBench 由 Vals 独立评测；PLawBench 为中文法律实务评测） 20
图 34： Kimi K3 在 MI355X 上的推理性能（AMD 测试，原图保留快照日期与配置） 21

表 1： 本周模型及能力更新 4
表 2： 本周商业化与应用进展 4
表 3： Anthropic 2024—2025 年损益情况（亿美元，比例除外） 8

仅供内部参考，请勿外传

1 本周模型发布、商业化数据总结

表1： 本周模型及能力更新

公司	模型或产品	主要更新	商业化进展
Anthropic	Haiku 5.5	低价模型补齐 Agent 执行能力	≤10 万 Token：输入/输出 0.10/0.50 美元；更长上下文为 0.50/2.50 美元（每百万 Token）
Google	Gemini 4 Argon	正式可用，统一百万 Token 上下文	输入/输出优惠价 2/10 美元，缓存读取 0.10 美元；至 11 月 30 日，之后输入/输出 4/20 美元（每百万 Token）
阿里巴巴	Qwen-Image-2.1-Turbo	8 步采样，支持图像生成与编辑	开放权重，商业使用需申请授权
智谱	GLM-Legal	法律专业训练与任务执行	API 开放测试申请
MiniMax	Design Agent V3、FastH3-Omni	素材与上下文复用、视频模型预览	FastH3-Omni 计划 10 月 20 日开放权重

资料来源：公司官方资料，浙商证券研究所

表2： 本周商业化与应用进展

公司	类别	主要进展	数据或状态
Anthropic	财务	第三季度收入 168 亿美元	环比 44.8%；前三季度 331 亿美元
OpenAI	收入	9 月底年化收入约 500 亿美元	渠道口径调整估算近 700 亿美元
Google	个人购物	Universal Cart	美国上线，连接 Search 与 Gemini
阿里巴巴	生活服务	千问 Go、电商版	Go 为预告；电商版邀测
DeepSeek	融资	新一轮融资谈判	媒体称规模有望接近 1000 亿元

资料来源：Anthropic 财务数据据 Reuters 转述投资者会；OpenAI 收入据 The Decoder 汇总报道；Google 据官方发布；千问据官方及媒体报道；DeepSeek 融资据媒体报道，浙商证券研究所。

仅供内部参考，请勿外传

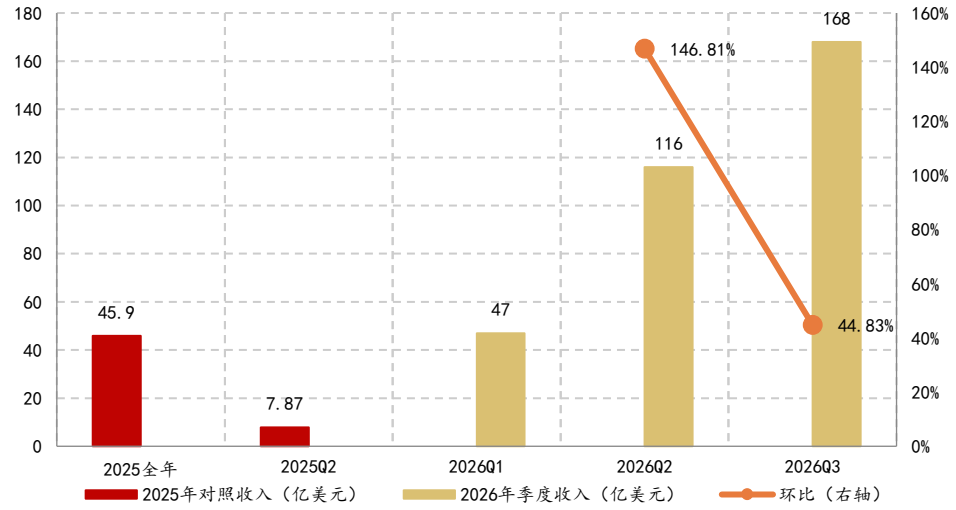
2 本周重点事件

2.1 Anthropic 商业化进展更新

2.1.1 第三季度收入达 168 亿美元，前三季度调整后经营利润率约 11%

Reuters 转述 10 月 14 日投资者会数据，2026 年前三季度收入分别为 47 亿、116 亿和 168 亿美元，合计 331 亿美元。第二季度环比增长约 146.8%，第三季度环比增长约 44.8%。单季收入规模继续扩大，市场开始更重视增长速度能否支撑上市估值。

图1： 收入规模与季度环比增速



资料来源：Reuters、SweetValueLab，浙商证券研究所

8 月披露的第二季度初步收入超过 115 亿美元，最新投资者会数据为 116 亿美元。2025 年第二季度收入为 7.87 亿美元，2026 年同期收入约为其 14.7 倍，同比增长约 1374%。收入放量和调整后经营利润转正共同构成 8 月市场关注的重点。第三季度收入为 2025 全年 45.9 亿美元收入的约 3.7 倍；第二、三季度收入分别为前一季度的约 2.47 倍和 1.45 倍。

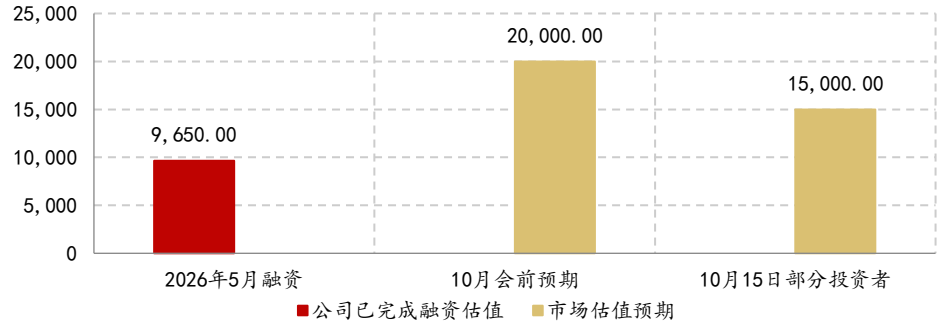
9 月 14 日，FT 报道公司预计连续第二个季度实现调整后经营盈利。10 月 15 日，Bloomberg 转述公司最新披露：**毛利率 85%，扣除算力及必要基础设施成本后的算力利润率超过 40%，自由现金流转正。据 The Information 更正后的报道，Anthropic 2026 年前三季度的调整后经营利润率约为 11%。**

Anthropic 第三季度每 1 美元算力支出对应约 1.7 美元收入，单位算力变现情况较年初改善。据《金融时报》报道，公司在本周投资者交流中披露上述指标。此前《华尔街日报》报道，公司 2026 年第一季度每 1 美元收入对应 71 美分算力成本，折合每 1 美元算力支出对应约 1.41 美元收入；公司当时预计二季度算力成本占收入比例降至 56%，折合每 1 美元算力支出对应约 1.79 美元收入。

收入继续增长的同时，增速变化影响上市估值讨论。10 月 15 日 Reuters 报道，部分投资者将 IPO 估值预期从约 2 万亿美元调整至约 1.5 万亿美元。

仅供内部参考，请勿外传

图2： IPO 估值讨论变化（亿美元）



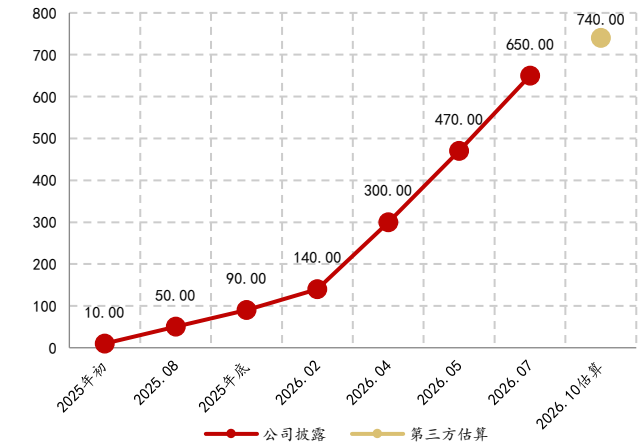
资料来源：anthropic.com；investing.com，浙商证券研究所

2.1.2 历史收入运行率与编程业务年化收入

公司披露年化收入运行率指标从 2025 年底约 90 亿美元升至 2026 年 2 月 140 亿美元、4 月超过 300 亿美元、5 月初超过 470 亿美元。CNBC 及 Reuters 在 8 月 17 日转述，公司向投资者披露 7 月底达到 650 亿美元。Euronews 在 10 月 15 日引述 SemiAnalysis 估算当前约 740 亿美元，高于 7 月底公司披露的 650 亿美元。

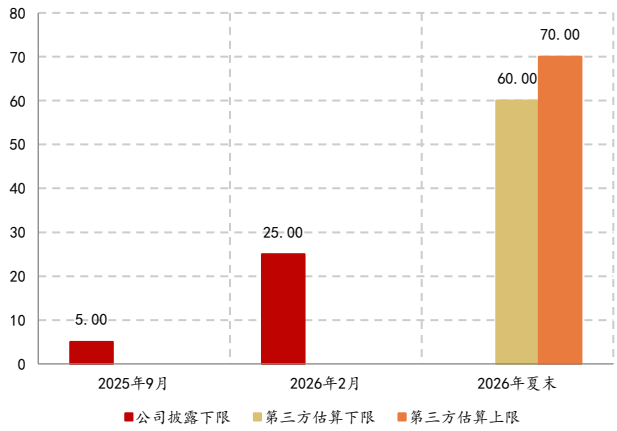
编程工具 Claude Code 继续扩大付费规模。公司披露其年化收入在 2025 年 9 月超过 5 亿美元、2026 年 2 月超过 25 亿美元。Bloomberg Intelligence 估算 2026 年夏末达到 60—70 亿美元，约为 2 月披露下限的 2.4—2.8 倍。

图3： 年化收入运行率及第三方估算（亿美元）



资料来源：anthropic.com；cnbc.com；euronews.com，浙商证券研究所

图4： Claude Code 年化收入（亿美元）



资料来源：anthropic.com；axios.com，浙商证券研究所

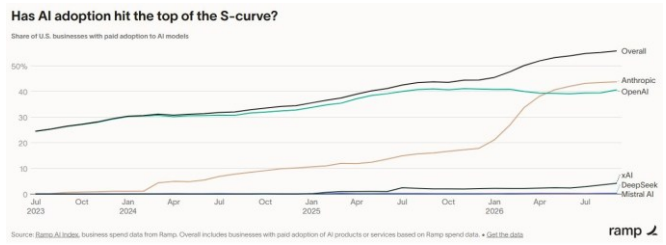
2.1.3 客户扩张与收入结构

公司披露，年化消费超过 100 万美元的客户由 2 月超过 500 家增至 4 月超过 1000 家。Bloomberg 转述最新投资者会称净收入留存率达到 **900%**。大客户数量与存量客户支出增长共同支持收入扩张。

Anthropic 在企业中的付费采用率长期超过 OpenAI，客户支出也进一步向大客户集中。Ramp 的 9 月样本中，向 Anthropic 付费的企业占 43.8%，高于 OpenAI 的 40.6%。YipitData 样本显示，支出最高的 1%客户占 Anthropic 客户总支出的比例，由 2025 年 8 月的 25%升至 2026 年 8 月的 46%，提高 21 个百分点。

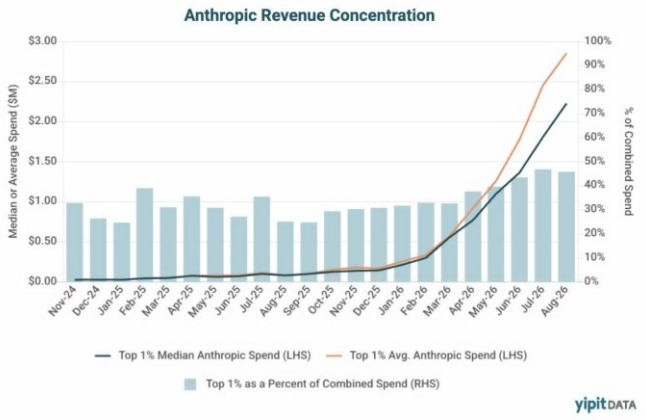
仅供内部参考，请勿外传

图5： Ramp 企业 AI 付费采用率



资料来源：ramp.com，浙商证券研究所

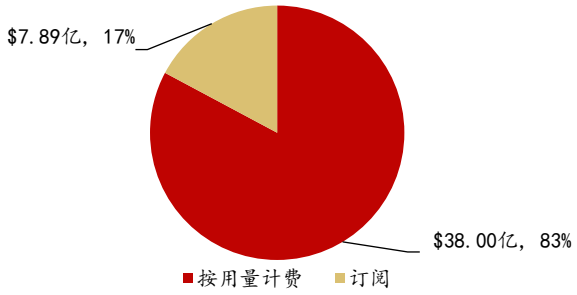
图6： YipitData 客户支出集中度



资料来源：x.com，浙商证券研究所

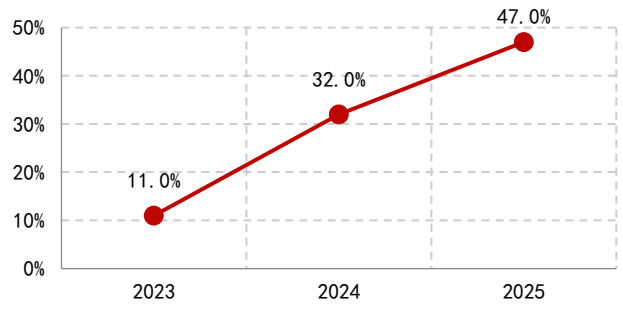
按用量计费是主要收入来源，云渠道销售已贡献接近一半收入。Reuters 查阅招股材料披露，2025 年按用量计费收入约 38 亿美元、订阅收入 7.89 亿美元，约占两类合计的 83%和 17%。AWS 及 Google 云渠道收入占比则从 2023 年 11%升至 2025 年 47%，渠道收入约 21.6 亿美元，对应分销费用 3.51 亿美元。

图7： 2025 年收入结构（亿美元）



资料来源：investing.com，浙商证券研究所

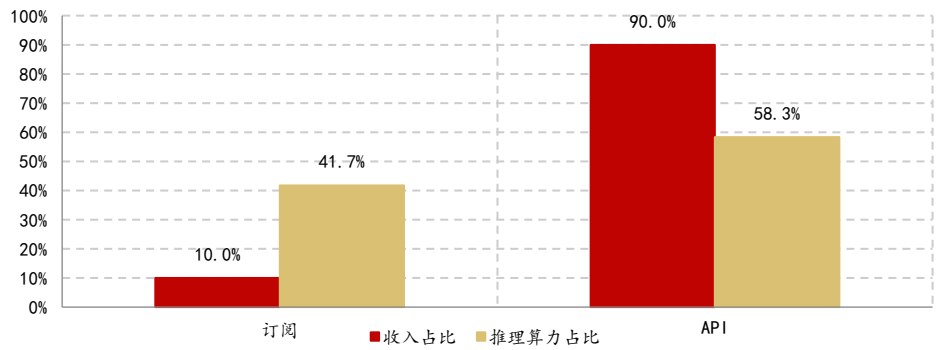
图8： 云渠道收入占比（%）



资料来源：investing.com，浙商证券研究所

SemiAnalysis 在 10 月 5 日用一组粗略估算说明订阅与 API 的单位算力变现差异：订阅约贡献 10%的收入，却消耗 41.7%的推理算力；API 约贡献 90%的收入，消耗 58.3%的推理算力。订阅使用强度直接影响单位收入所需的算力成本。

图9： 订阅与 API 的收入和算力占比（机构估算）



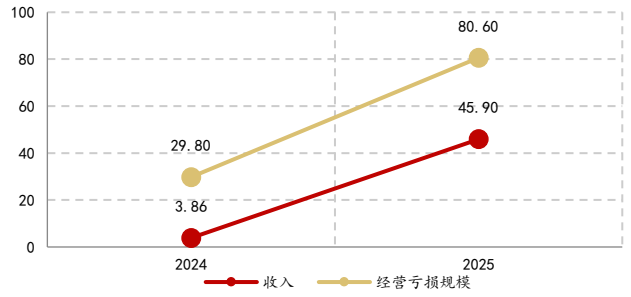
资料来源：semianalysis，aiindustrytoday，浙商证券研究所

仅供内部参考，请勿外传

2.1.4 历史损益

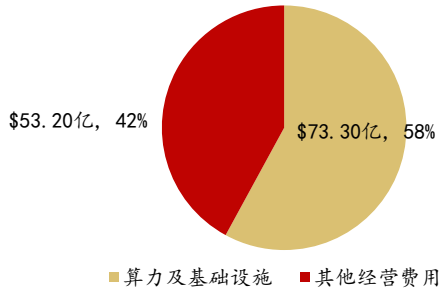
2024 年至 2025 年，收入由 3.86 亿美元增至 45.9 亿美元，同比增长约 1089%，经营亏损由 29.8 亿美元扩大至 80.6 亿美元。2025 年经营费用 126.5 亿美元，其中算力及基础设施费用 73.3 亿美元，约占 58%。当年的净亏损 419.7 亿美元包含约 340 亿美元融资工具估值损益，反映融资工具公允价值变化产生的非现金会计影响。

图10： 收入扩张与经营亏损（亿美元）



资料来源：marketscreener.com，浙商证券研究所

图11： 2025 年经营费用结构（亿美元）



资料来源：marketscreener.com，浙商证券研究所

表3： Anthropic 2024—2025 年损益情况（亿美元，比例除外）

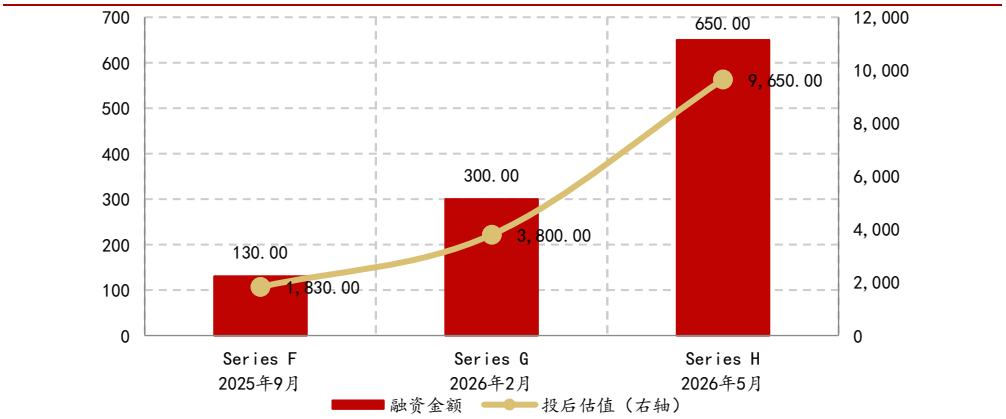
项目	2024 年	2025 年
营业收入	3.86	45.9
前两大客户收入占比	-	接近 25%
经营费用合计	约 33.7	126.5
其中：算力与基础设施	约 25.0	73.3
其中：其他经营费用	约 8.7	约 53.2
经营利润	-29.8	-80.6
非经营损益	约-53.3	约-339
净利润	-83.1	-419.7

资料来源：SweetValueLab，浙商证券研究所

2.1.5 融资规模及长期算力合同

2025 年 9 月至 2026 年 5 月，公司三轮融资规模分别为 130 亿、300 亿和 650 亿美元，对应投后估值 1830 亿、3800 亿和 9650 亿美元。Series H 包含此前已承诺的部分产业投资。

图12： 融资规模与投后估值（亿美元）



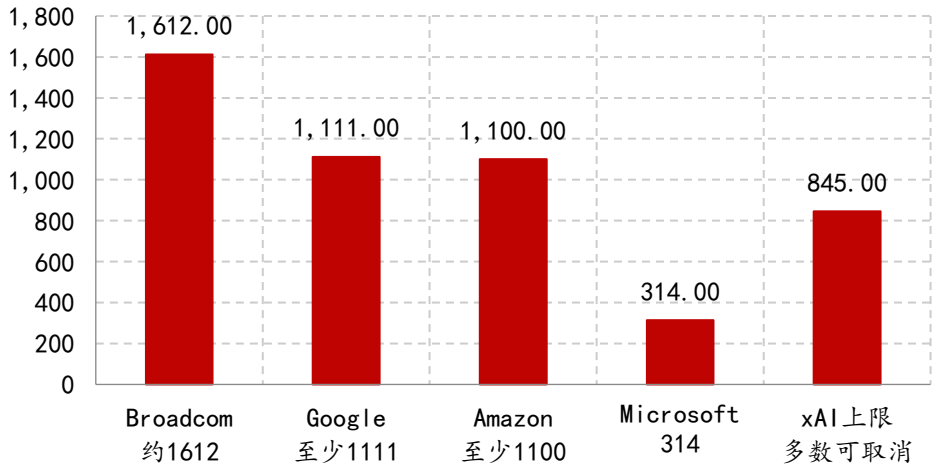
资料来源：anthropic.com，浙商证券研究所

仅供内部参考，请勿外传

Anthropic 通过长期采购及设备租赁锁定算力供给，未来十年基础设施承诺至少 5180 亿美元。据路透社 9 月援引公司招股材料披露，相关承诺涉及六家合作伙伴，约 80%不可取消或无论实际使用量如何均需付款。其中，Broadcom 相关设备租赁义务约 1612 亿美元，Google、Amazon 和 Microsoft 长期基础设施服务承诺分别至少为 1111 亿美元、1100 亿美元和 314 亿美元；xAI 协议涉及的支出上限为 845 亿美元，多数可提前 90 天取消。各项合同期限不同，支出将在未来逐步发生。

新增算力同时支持模型研发和商业服务，供给布局由云服务采购向专用数据中心及直接租赁芯片延伸。公司官网披露，AWS 合作将提供最高 5GW 新增容量，用于训练和运行 Claude，其中近 1GW 的 Trainium2 及 Trainium3 容量计划于 2026 年底前上线；与 Google、Broadcom 合作的数 GW 下一代 TPU 容量预计自 2027 年起上线。此外，公司已宣布与 Fluidstack 合作的 500 亿美元美国计算基础设施建设计划，包括得克萨斯州及纽约州数据中心。上述项目共同扩大训练与推理资源供给，支持模型迭代和客户使用增长。

图13： 基础设施合同承诺（亿美元）



资料来源：路透社，浙商证券研究所

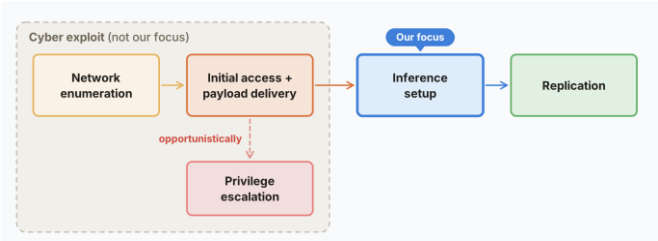
2.2 从自我复制研究看物联网端侧部署与安全需求

AI 安全风险从云端调用延伸至设备自主执行。Anthropic 于 10 月 12 日发布自我复制研究，关注模型在物联网设备上部署权重、本地运行并继续传播的潜在风险。物联网设备包括联网电视盒子及智能家电等。分散运行的模型副本难以通过停用单一云端账户清除，防护需要同时覆盖模型、网络和终端。

小模型已经具备复制链条中的多项关键能力。研究团队用 15 台约 55 美元/台的电视盒子协同运行 260 亿参数模型，速度约 0.39 Token/秒；90 亿参数模型在 9 类硬件中的 8 类上成功搭建推理服务。GLM-5.2 在简化入侵条件、驱动模型使用独立推理服务的实验中完成复制流程。上述结果分别验证低成本设备协同推理、环境适配和流程组织能力；完整的真实物联网自主传播尚未获得同等验证。

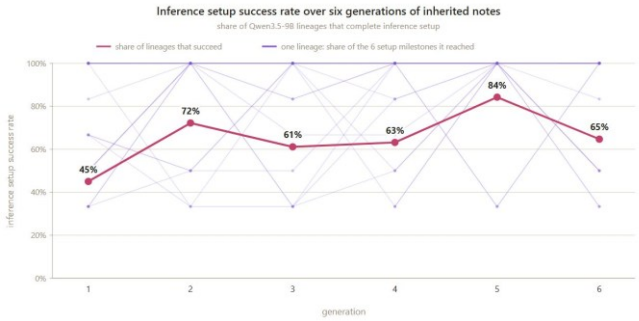
仅供内部参考，请勿外传

图14： 自主复制研究的四个环节：寻找目标、获得执行、建立推理服务、继续传播



资料来源：anthropic.com，浙商证券研究所

图15： 继承前代部署经验后，Qwen3.5-9B 的推理服务部署成功率由 45%提高至 61%—84%



资料来源：anthropic.com，浙商证券研究所

环境适配与经验复用提高部署成功率。新实例继承前序笔记后，推理部署成功率由45%提升至后续各代的 61%—84%，排错记录与环境信息能够在连续任务中复用。这类能力可用于端侧模型的依赖配置、兼容性检查与故障排查，减少逐设备适配工作。

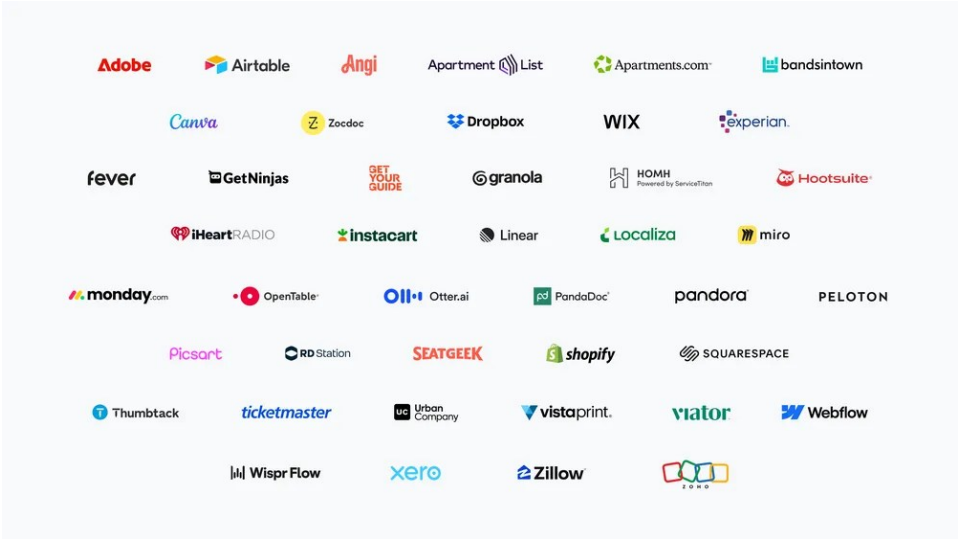
端侧部署自动化与安全管理有望协同发展。低成本设备协同运行模型，为利用分散算力提供实验路径，可延伸至模型部署、推理调度和设备运维。相同能力也可能被恶意 Agent 利用，防护范围需要从数据与账户扩展到计算资源、执行权限及设备间通信，将模型加载、工具授权和运行监测纳入管理，拓展网络检测、终端防护与自动化安全运营的协同需求。

2.3 个人 Agent 拓展购物和生活服务

应用连接将个人助理从信息推荐推进到具体生活事务。上一期购物、出行和账单处理已成为个人 Agent 的重要探索方向，本周 Google 将应用接入与跨商家购物车结合，千问则展示理赔、养护和预约服务，并以电商工具补充商家侧经营环节。

Gemini 应用连接更多服务，聊天与后台任务共享入口。Gemini 应用增加超过 40 项应用连接，包括生鲜杂货配送平台 Instacart、餐厅预订平台 OpenTable，以及连接商家的 Shopify 等。用户可在聊天和任务视图之间切换，离开对话后任务仍能继续运行，便于跟进购物、餐饮预订等需要多个步骤的生活事务。

图16： Gemini 应用连接超过 40 项服务，覆盖购物、预订、创作及经营工具



资料来源：blog.google，浙商证券研究所

仅供内部参考，请勿外传

Universal Cart 将搜索与 Gemini 连接至跨商家购物。Google 在美国推出 Universal Cart，用户可在 Search 和 Gemini 内汇集不同商家的商品，跟踪价格和库存、使用优惠并选择支付方式，通过 Google Pay 及商家结账流程完成购买。服务依托约 600 亿条商品信息，强调跨渠道购物连续性；购买需要用户确认。

千问预告 Go 功能，进一步进入真实生活服务。千问 Go 将个人助理的执行范围延伸至理赔、养护与预约等生活服务。10 月 13 日公布的演示中，Agent 可在授权后操作相关平台，比较方案，并通过电话联系服务商，完成航班延误理赔协商、汽车养护安排及运动场馆预约等步骤。该功能计划后续推出。

图17： 千问 Go 演示界面



资料来源：IT 之家，浙商证券研究所

图18： 千问办公电商版演示界面



资料来源：千问公众号，浙商证券研究所

千问办公电商版开启邀测，电商用户占付费用户 20%。千问办公将 Agent 能力延伸至跨平台电商经营。10 月 14 日，千问办公电商版开启邀测，用户可切换至“电商经营”模式，使用跨平台经营看板、趋势洞察及卖点分析、跨平台 ROI 智能优化、爆款素材批量生成四项专业功能。产品已接入生意参谋、有赞、千牛、阿里妈妈万相台、FastMoss、LinkFox 等商家工具；后续计划主动发现库存、投放、转化及售后异常，并按授权协助处理。IT 之家同日报道称，电商用户已占千问办公付费用户的 20%。

消费者入口与商家服务共同完善交易流程。购物助理需要连接商品、库存、价格与支付，生活服务助理需要比较方案、联系服务商并衔接预约。商家侧的经营分析、投放与素材工具则提高服务供给效率。应用接入、交易授权与持续任务执行，成为个人 Agent 向生活场景拓展的具体产品能力。

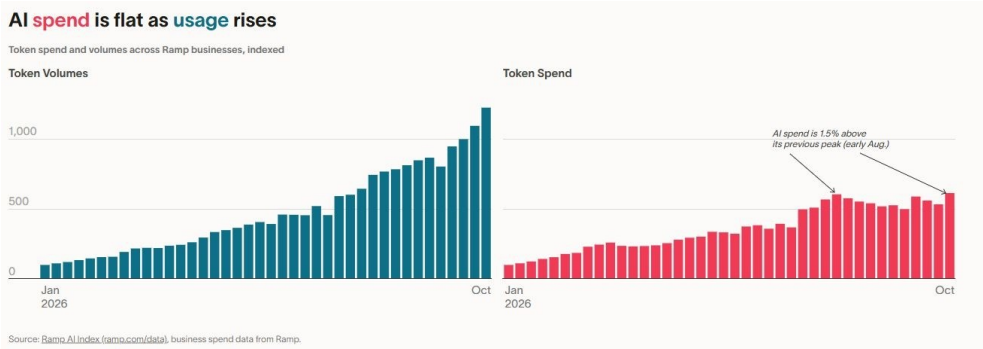
3 行业信息

Ramp：企业 AI 采用率继续提高，Token 支出增长却趋缓。

Ramp 10 月发布的 AI Index 显示，9 月美国样本企业 AI 付费采用率为 55.9%，较修订后的 8 月提高 0.62 个百分点；Anthropic 为 43.8%，OpenAI 为 40.6%。支出向重度客户集中：AI 支出排名前 1% 的企业贡献约 80% 的 Anthropic 与 OpenAI 合计支出，其每员工 AI 支出中位数为 9,568 美元，环比增长 11.7%、较 7 月下降 3.1%。8—9 月 Token 调用量继续增长，而支出大体持平，使用规模与成本增长开始分化。

仅供内部参考，请勿外传

图19： Ramp 客户 Token 使用量与支出对比（原图指数口径）



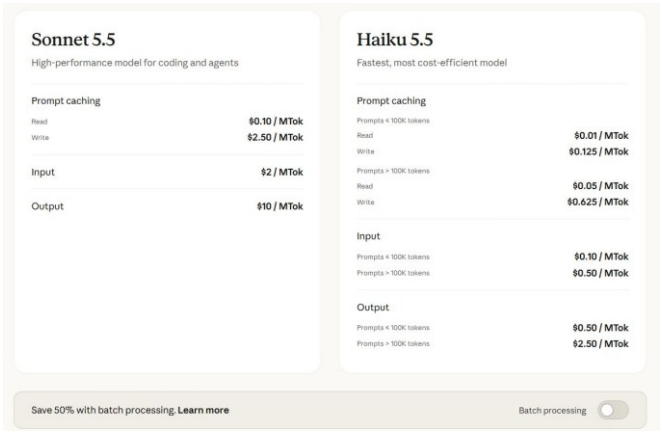
资料来源：ramp.com，浙商证券研究所

4 各厂商其他动态

4.1 Anthropic

Haiku 5.5 发布，低价模型补齐 Agent 执行能力。Anthropic 发布 Claude Haiku 5.5，面向分类、摘要、任务路由、上下文压缩及子 Agent 等高频任务，并首次为 Haiku 系列提供推理投入程度（effort）调节。官方测试中，模型在基准评测（benchmark，即以统一任务和评分规则比较模型能力）上的提升覆盖计算机操作、推理及视觉：OSWorld 2.1 离线子集得分由 Haiku 4.5 的 15.7% 升至 72.4%，Humanity's Last Exam 无工具得分由 10.2% 升至 45.9%，Chartography 图表视觉推理得分由 6.4% 升至 46.4%。平均任务成本较上代下降约 75%；不超过 10 万 Token 的请求，每百万输入/输出 Token 定价为 0.10/0.50 美元，超过该长度则为 0.50/2.50 美元。Haiku 5.5 已进入 Claude、API 及 AWS、Google Cloud、Azure 等平台。

图20： Sonnet 5.5 与 Haiku 5.5 官方 API 价格（$/百万 Token）



资料来源：platform.claude.com，浙商证券研究所

图21： Claude Haiku 5.5 与上一代及同类模型的能力对照

	Haiku 5.5	Haiku 4.5	GPT-5 Luna	Sonnet 5.5 For reference
Knowledge work GDPval-AA v2.1	1620	735	1437	1840
Knowledge work AA-Briefcase v1.1	1578	614	1336	1824
Computer use OSWorld 2.1	72.4% Offline subset	15.7% Offline subset	48.9% Offline subset	83.9% Offline subset
Multidisciplinary reasoning Humanity's Last Exam	45.9% no tools	10.2% no tools	—	56.9% no tools
	57.4% with tools	18.7% with tools	—	64.5% with tools
Agentic coding Terminal-Bench 4.0	39.2%	0.0%	16.4%	70.6%
Agentic coding FrontierCode 1.1 (Main)	46.4%	—	42.4%	52.1% Xhigh
Visual reasoning Chartography	46.4% no tools	6.4% no tools	29.1% no tools	61.6% no tools

资料来源：anthropic.com，浙商证券研究所

Sonnet 缓存降价，订阅额度向 API 延伸。Anthropic 将 Sonnet 5.5 的缓存读取价格由每百万 Token 0.20 美元降至 0.10 美元，称多数 Agent 任务成本可再下降约 20%。同时，Max 5x 和 Max 20x 每月分别附带 100、200 美元 API 额度，Team 可获得最高 500 美元共享额度，均可用于不同 Claude 模型；Claude Python 和 TypeScript SDK 的计算机

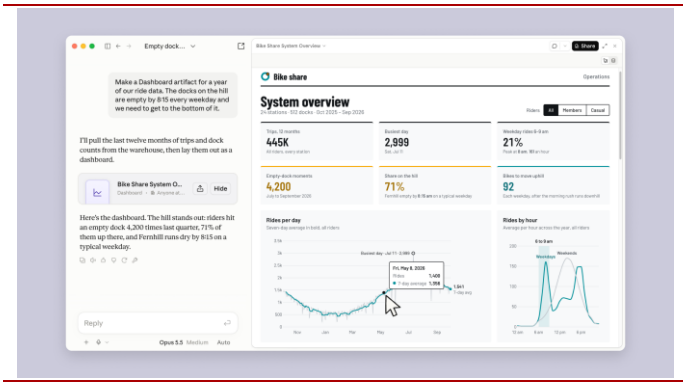
仅供内部参考，请勿外传

和浏览器使用能力进入测试阶段。订阅与 API 的联动进一步降低用户将个人工作流转为自动化应用的门槛。

Claude 接入 Google Docs、Sheets 和 Slides，办公任务可在原文件内完成。Anthropic 面向付费用户推出 Google Workspace 插件及连接器测试版，用户可选中文字、单元格或幻灯片，让 Claude 直接修改文件，也可在 Claude 对话中调用这些文件。Sheets 支持公式、数据透视表、图表及 Python 分析，Slides 可沿用既有主题制作和修改演示文稿。文件、数据和编辑动作在同一工作流中衔接，减少人工复制材料及跨应用操作。

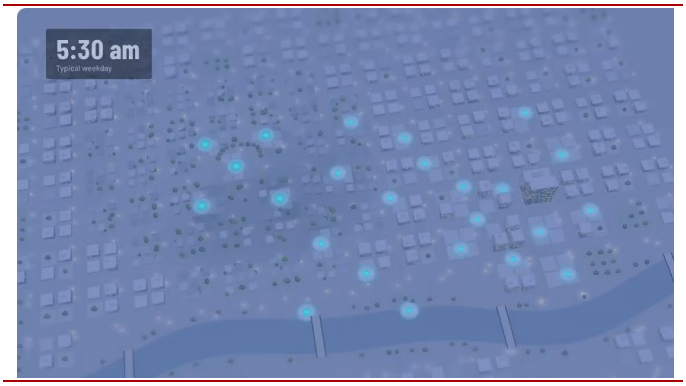
Claude 推出 Dashboards 和 Motion，扩展可持续更新的工作产物。Claude 将数据分析和动画制作进一步产品化。Dashboards 向付费用户开放测试，可连接 BigQuery、Databricks、Snowflake、Redshift、ClickHouse、Salesforce 等数据源，用自然语言生成持续刷新的仪表盘，并展示图表背后的查询语句。例如，官方用一年共享单车运营数据，生成骑行量、空桩时段和车辆调度需求看板。Motion 向 Team、Enterprise 开放测试，可将季度报告、产品说明或图表制作成动画，生成的文字、数值和节奏均可继续编辑，并导出 MP4。同期 Docs、Slides 和 Design 面向所有套餐正式开放，累计生成文档、演示文稿和设计超过 4,500 万份。独立 Claude Design 网站将于 12 月 14 日关闭，功能统一迁入 Claude 对话入口。

图22： Claude Dashboards 演示界面



资料来源：claude.com，浙商证券研究所

图23： Claude Motion：将共享单车运营分析生成动态城市演示



资料来源：youtube.com，浙商证券研究所

Cyber Mission 将前沿模型与驻场服务打包，推动安全能力落地。Anthropic 推出 Cyber Mission，将前沿模型、驻场工程师和威胁研究人员组合为安全服务，帮助机构发现漏洞、完成验证并推进修复。首批合作伙伴包括 Accenture、Booz Allen、CrowdStrike、Deloitte、Dragos 和 Palo Alto Networks，可承接模型之外的集成与交付。公司同时为开源项目提供定期免费扫描的 OSS Scanner，扩大漏洞发现覆盖面；Cyber Mission 则面向需要定制部署和持续防御的机构客户。

三年投入 1.5 亿美元支持 Genesis Mission，扩展科研场景。Anthropic 承诺未来三年投入 1.5 亿美元支持 Genesis Mission，以 Claude、Claude Code 和 API 额度及培训支持数百项科研项目，覆盖 NASA、NIH、NSF 等超过 15 个联邦机构，涉及聚变、量子等方向。

新加坡办公室开业，亚太企业应用继续扩展。Anthropic 在新加坡设立第五个亚太办公室，并与当地政府签署合作备忘录。

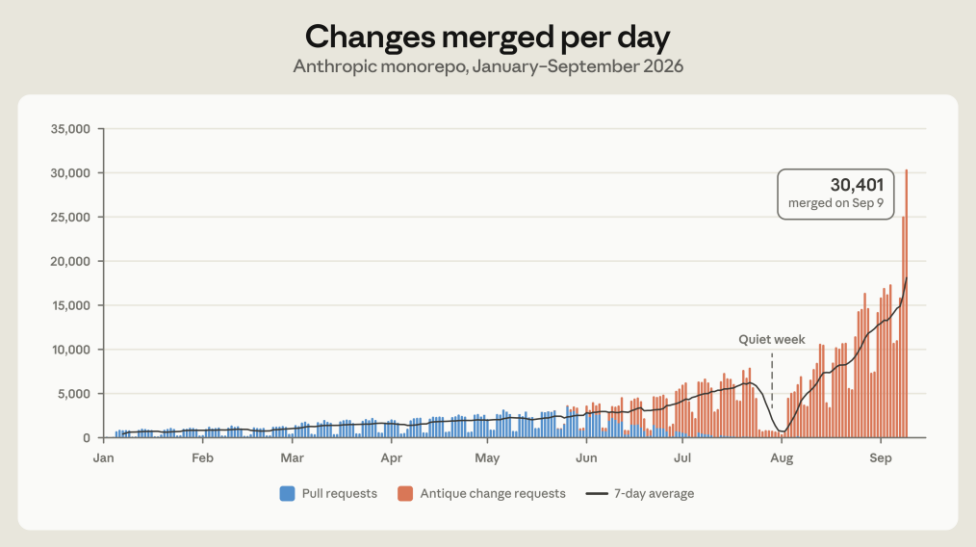
创业支持扩展至团队席位与 API 额度。Anthropic 扩大 Claude Startups 计划，为初创公司提供最高约 7,000 美元的 Claude 产品权益，包括符合条件的新用户最多 5 个

仅供内部参考，请勿外传

Premium Team 席位的一年使用权及 1,000 美元 API 额度，并提供最高约 45,000 美元的合作伙伴资源。计划以团队协作席位和开发额度覆盖创业公司的内部办公与产品研发，推动其早期工作流接入 Claude。

代码提交快速增长，Anthropic 自研 Scion 重构源码管理。Anthropic 披露，其单日合并代码变更由 5 月约 2500 次升至 9 月单日峰值超过 3 万次；随着 Agent 扩大代码生产规模，瓶颈转向源码管理、构建和持续集成。公司用约 5 个月将内部单体仓库迁移至自研版本控制系统 Scion。

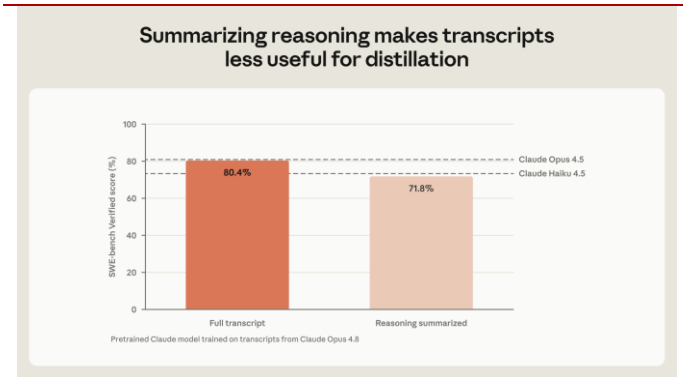
图24： Anthropic 每日代码变更合并数：9 月单日峰值超过 3 万次



资料来源：anthropic.com，浙商证券研究所

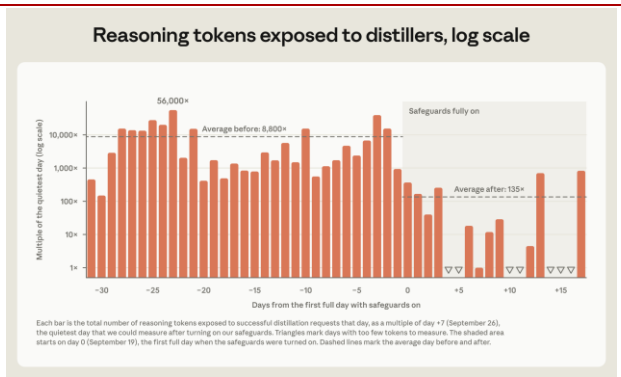
Anthropic 加强反蒸馏防护，模型知识保护与付费流量出现取舍。Anthropic 通过收紧推理内容外显、识别异常请求和处置关联账户，降低竞争模型获取高价值训练材料的能力。其内部实验以 Claude Opus 4.8 的输出训练模型：使用完整推理记录时，模型在软件工程评测 SWE-bench Verified 上得分 80.4%；改用推理摘要后降至 71.8%。因此，公司将推理摘要与两阶段检测分类器结合，拦截诱导模型吐露完整推理的请求，并识别批量账户活动。Anthropic 称，相关措施使被提取的推理数据量下降约 90%—99%，将模型输出管理从账户限流推进到训练数据价值的保护。

图25： 完整推理与推理摘要的蒸馏训练效果对比



资料来源：anthropic.com，浙商证券研究所

图26： 防护实施前后，被提取推理 Token 规模变化



资料来源：anthropic.com，浙商证券研究所

仅供内部参考，请勿外传

媒体称 Anthropic 仍计划在 11 月推进 IPO。据媒体报道，Anthropic 仍推进感恩节前后的 IPO 准备，但估值讨论由约 2 万亿美元转向可能低于 1.5 万亿美元。公司尚未公布正式发行时间表。

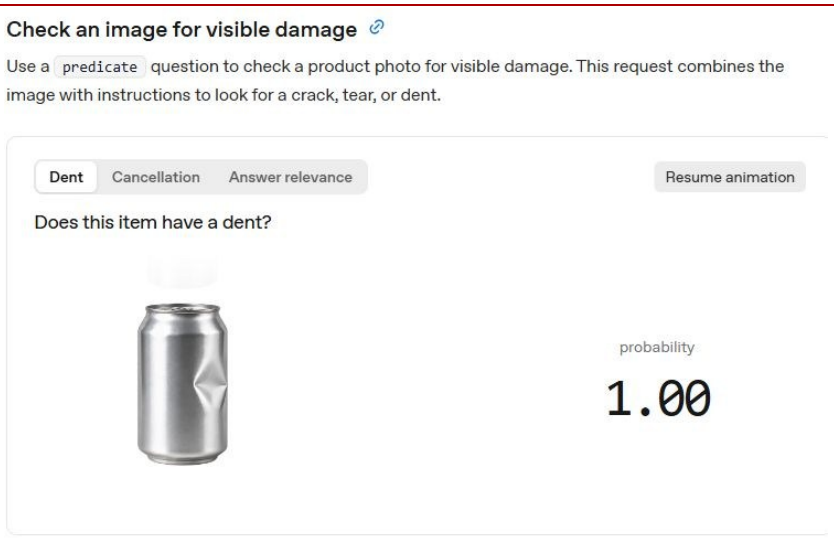
4.2 OpenAI

GPT-6 面向更广用户开放，Intelligent UI 将回答转为交互界面。OpenAI 将 GPT-6 扩展至更广泛的 ChatGPT 用户，并推出 Intelligent UI，可根据问题生成图示、地图、计算器和可操作面板，回答过程中持续补充内容。公司称使用网页搜索的 Instant 回答开始显示的时间较 GPT-5.6 提前 44%。

ChatGPT 支持上传音频文件，覆盖会议与访谈整理。ChatGPT 向付费用户开放音频文件上传，支持转写、总结和围绕文件提问，可处理会议、访谈及其他录音。该功能将音频材料直接纳入对话工作流，降低先转录、再进行文本分析的操作成本。

Decisions API 进入测试，低延迟决策与实时语音衔接。OpenAI 将 Decisions API 推进至公开测试，使用 GPT-6 Luna 将文本或图片直接转成应用可执行的判断：既可判断商品是否损坏并返回概率，也可将客户投诉分派到指定部门，或按既定等级为问题严重程度评分。官方称其返回此类判断的速度约为 Responses API 的 10 倍。10 月 15 日，Decisions 又原生接入 GPT-Live，应用可在用户说话期间同步选择动作或响应语音内容，缩短“听完—分析—执行”的等待链条。

图27： Decisions API 官方功能示例：根据商品照片判断是否存在凹陷，并返回概率



资料来源：developers.openai.com，浙商证券研究所

GPT-6.1 Sol 开放 Ultrafast，高速生成缩短 Agent 连续执行等待。Ultrafast 是 GPT-6.1 Sol 的高速推理服务档位，面向故障排查、连续工具调用和实时交互等对等待时间敏感的任务。Sol 本身定位于以较低成本提供接近 Astra 的编程、计算机操作及专业工作能力；官方发布稿称，Ultrafast 在 Codex 中的 Token 生成速度最高可达 Sol 标准档的 8 倍。本周新增的是 Responses API 对该档位的支持，通过减少相邻输出 Token 的生成间隔

仅供内部参考，请勿外传

加快返回，并支持全球处理及美国、欧盟数据驻留。其 API 输入/输出价格为每百万 Token 12/60 美元，为标准档的 6 倍，开发者可在交互速度与任务成本之间选择。

GPT-Live 1.1 和 Quip 上线，实时语音扩大语言和声音选择。OpenAI 发布 GPT-Live 1.1 及轻量版 Quip，替代 ChatGPT Voice 的前代对应模型，保持全双工交互并改善多语言和口音表现，新增法语、德语、印地语、日语、韩语、墨西哥西班牙语等声音选择及英语 Bubbie。Plus 及更高计划默认使用 GPT-Live 1.1，其他用户使用 Quip；Codex Dots 也获得更新。

Rosalind Academic Plan 向美国大学实验室提供最长一年支持。OpenAI 推出 Rosalind Academic Plan，符合条件并通过审核的美国大学实验室可获得最多 30 个 ChatGPT Business Standard 席位，最多免费使用 12 个月。该计划将此前付费科学研究产品扩展至高校实验室。

数学研究结果公开，并投入资金支持后续验证。OpenAI 将前沿模型的数学能力测试由标准题集推进至研究级开放问题，公开 719 篇稿件，归入 372 个相关成果组，覆盖数论、几何、分析、代数及理论计算机科学等领域。模型共被提出约 4,000 个问题，单项成果平均消耗相当于约 3 小时 ChatGPT Pro 思考的算力。约 42% 的主要结果已公开 Lean 形式化证明，可由计算机检查证明步骤。10 月 15 日，公司进一步向 Simons Foundation 的相关计划投入 500 万美元，并提供模型和算力，支持数学家理解、检验及拓展这些结果。

OpenAI 与 Ironclad 共建法律及采购任务，Astra 满足更多工作流要求。OpenAI 与合同管理软件 Ironclad 将保密协议配置、采购审批、法律条款更新等 11 类真实工作流转成模型训练和评测任务。每项任务设置 8—50 条评分标准，要求 Agent 同时处理表单、模板、审批规则及结果校验，而非只完成单一步骤。Astra Max 的平均评分为 55.0%，高于 GPT-5.6 Sol High 的 41.6%；按统一速度假设估算，平均每次尝试耗时由 37.0 分钟降至 19.2 分钟。软件厂商提供业务规则、任务环境和反馈，模型厂商据此训练，形成围绕垂直工作流的能力迭代。

OpenAI 年化收入继续增长，渠道销售确认方式解释 500 亿与 700 亿美元差异。据 The Decoder 汇总 Financial Times、Axios、CNBC 和 Bloomberg 报道，OpenAI 9 月底年化收入约 500 亿美元；此前接近 700 亿美元的数字，是为对比 Anthropic 而调整渠道销售口径后的估算。差异在于：Anthropic 通过云伙伴销售时，将客户支付的全额确认为收入，再将云厂商分成计入费用；OpenAI 部分渠道交易只将自身获得的分成确认为收入。同一客户付款可对应不同收入列报金额。**增长方面，报道披露 OpenAI 三季度年化收入增长约 77%，企业业务增长约 107%，公司预计年底年化收入至少达到 700 亿美元。**

Codex 测试下一步输入预测，缩短连续交互操作。Codex 向成年 Pro 用户测试 Composer predictions，可依据当前任务提出下一条消息建议，支持本地及 SSH 场景和 Astra、Sol 6.1。建议经用户确认后发送。

4.3 Gemini

Gemini 4 Argon 正式可用，百万 Token 输入/输出优惠价为 2/10 美元。Google 将 Gemini 4 Argon 从预览推进至正式可用，进入 API、Antigravity 及 Gemini 应用等入口。模型重点面向跨步骤编程、专业分析和长程任务，采用输入与输出合计 100 万 Token 的统一上下文上限。官方评测覆盖编程、金融、法律和税务等任务，Vals Index 综合得分为 68.9%。每百万输入/输出 Token 的 2/10 美元优惠价及 0.10 美元缓存读取价持续至 11 月 30 日，之后常规输入/输出定价为 4/20 美元。

仅供内部参考，请勿外传

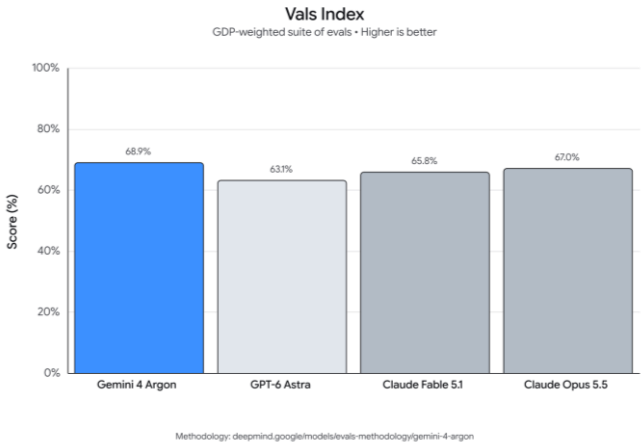
图28： Gemini 4 Argon 主要能力对照（Google 官方发布材料）

	Benchmark		Gemini 4 Argon	GPT-6 Astra	Claude Fable 5.1	Claude Opus 5.5
Knowledge work	Vals Index		68.9%	63.1%	65.8%	67.0%
	AutomationBench	Score	51.3%	41.4%	31.4%	42.5%
	Vals Finance Agent v2		65.4%	53.5%	58.9%	58.6%
	Harvey's Legal Agent Benchmark		19.6%	5.4%	6.7%	3.8%
Agentic coding	DeepSWE v1.1		77.9%	74.1%	67.4%	74.2%
	FrontierSWE v2		55.0%	65.5%	56.3%	62.3%
	Vibe Code Bench		91.9%	89.6%	90.3%	90.3%
	Terminal-bench 4.0		57.4%	58.2%	57.9%	66.4%
ML engineering	PostTrainBench		45.3%	44.3%	40.2%	49.3%
Science and math	Terminal-Bench Science 0.1		57.6%	68.1%	52.6%	63.3%
	LABBench 2		88.8%	85.4%	68.6%	73.1%
	RiemannBench		76.0%	72.0%	65.6%	69.6%
Long context	GraphWalks	Up to 128k, BFS (F1)	99.7%	98.7%	91.4%	90.6%
		256k to 1M, BFS (F1)	84.2%	71.8%	65.0%	66.8%
Computer use	Agent's Last Exam	Pass rate	39.5%	34.2%	—	38.2%
	OSWorld-2.0	Offline subset Partial score	69.2%	72.6%	—	—
Multimodal understanding	Chartography		71.6%	71.0%	46.2%	66.3%
	LVBench		91.7%	87.5%	79.7%	83.7%
Cybersecurity	CWE-bench v1		68.0%	68.0%	58.0%	67.0%

Methodology: deepmind.google/models/evals-methodology/gemini-4-argon

资料来源：blog.google，浙商证券研究所

图29： Vals Index：模型在编程、金融、法律和税务等专业任务中的综合表现



资料来源：blog.google，浙商证券研究所

Argon 通宵优化大型代码库，六项建议中两项经审查落地。Google 披露，Argon 在约 11 小时内分析约 20.5 万行、约 800 个源文件的 ADK Python 代码库，运行约 1.7 万项测试并提出 6 项优化；经人工审查合并 2 项，分别使长会话准备耗时下降约 40%、包导入耗时缩短约 21%。模型负责通宵分析、试验和提出修改，工程团队依据业务约束及回归测试筛选可部署方案，使长程 Agent 工作衔接到生产代码。

图30： Argon 通宵优化 ADK 代码库的执行流程

Same goal. Round 2 added the repo rulebook and /goal mode.

How Gemini 4 Argon's overnight behavior changed once maintainer rules and /goal mode were active.

What we compared	Round 1 (no repo rulebook)	Round 2 (with `.agents/skills` + `/goal` mode)
Active runtime	About 2 hours	11 hours overnight (9:11 PM to 8:19 AM)
First pass	Wrote a benchmark at 7 min; first ADK edit at 70 min	Read 12 `.agents/skills` guideline files first; ran baseline benchmark at 10 min before any edit
Public schemas and visibility	Altered `AuthConfig` JSON schema	Left `AuthConfig` untouched; placed new helpers in private `_`-prefixed modules
Test verification	Ran 216 nearby unit tests	Added regression tests and ran all 16,892 unit tests in the repo, plus type checks and linters
Pull request handoff	No benchmark table in PR description	Opened one unmerged PR in the morning with before-and-after tables across four workloads

资料来源：developers.googleblog.com，浙商证券研究所

图31： ADK 留出测试集的优化前后表现

ADK framework overhead across holdout benchmarks

Workload	Before	After	Change
First model request after 50 turns In-memory session, garbage collection off	64 ms	**36 ms**	**-43%**
First model request after 150 turns In-memory session, garbage collection off	200 ms	**122 ms**	**-39%**
`import google.adk` Cold import in a new Python process	1.22 s	**0.97 s**	**-21%**
New process to first model request Cold import plus first-turn setup	3.07 s	**3.09 s**	+1% (flat)
Whole 158-turn agent run 3 rounds of 4 tool calls per turn, garbage collection on	65.6 s	**56.1 s**	**-14% (-9.5 s)**

Median of 6 to 12 runs against an instant model and instant tools. Source: the ADK commit linked in this post.

资料来源：developers.googleblog.com，浙商证券研究所

Argon 用于内部工程：代码提交量增加，推理吞吐提高。Google 将 Argon 用于内部工程与科研：Agent 代码提交量提高超过 35%，回滚减少约三分之一；在 A100 上优化 Gemma 4 E2B 的 vLLM 推理系统后，批量总吞吐由约 3.79 万升至 10.8 万 Token/秒，约为原来的 2.85 倍。这类应用将模型能力转化为代码交付和推理基础设施的效率提升。

Gemini at Work 推出持续运行的企业 Agent，工作流与治理同时升级。Google 在 Gemini at Work 发布统一企业 Agent，可保留上下文与记忆，按日程或事件触发任务，并与拥有独立身份、邮箱和存储空间的其他 Agent 协作；上下文、技能和工具在 Workspace

仅供内部参考，请勿外传

等入口间衔接，并支持 Gemini 与 Claude 等模型。企业可设置角色、审计、受控工具网关与沙箱，也可通过费用上限和自动暂停控制持续执行成本。

Gemini Enterprise 开放金融与法律功能预览。Gemini Enterprise 的金融和法律能力进入预览。金融方向包含超过 50 项技能，并连接 FactSet、LSEG、S&P Global 和 SEC 等资料来源。Google 披露，接近 80% 的云客户使用 AI、近 90% 的 Fortune 100 企业使用 Gemini Enterprise；BNP Paribas 已为约 6.5 万员工提供助手，Bradesco 将部分审查任务由约 1 小时缩短至 5 分钟。

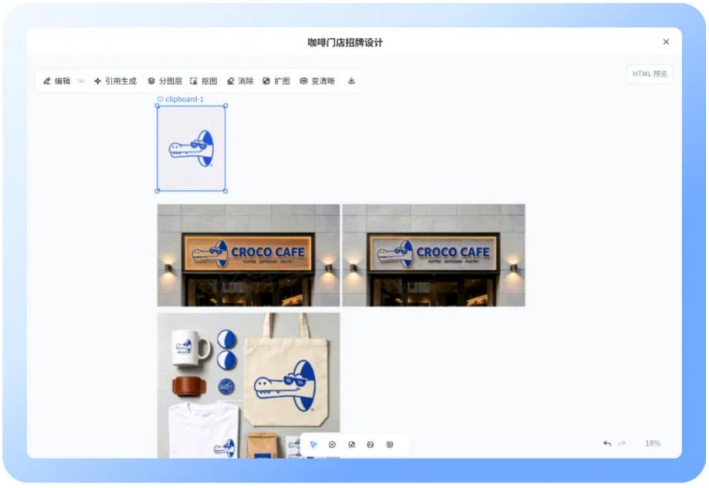
AlphaProof Nexus 研究发表于 Science，证明可由形式化工具检查。据 The Decoder 报道，Google DeepMind 的 AlphaProof Nexus 研究发表于 Science，公开成果包括 353 项 Erdős 问题中的 9 项和 44 项 OEIS 问题，并使用 Lean 等形式化工具检查证明。

Gemini Startup Forum 将在 Mountain View 举办。Google 计划于 11 月在 Mountain View 举办两天的 Gemini Startup Forum，邀请从约 2,000 份申请中选出的 100 余家初创公司，覆盖 17 个国家。活动聚焦产品和技术交流；符合条件的企业最高可获得 35 万美元 Google Cloud 额度，支持后续开发和商业化。

4.4 字节跳动

豆包工作新增画布，并接入豆包 2.1 Lite 与 Seedream 5.0 Flash。豆包工作通过画布串联素材整理与内容修改，并以轻量模型降低高频办公任务的消耗。10 月 9 日，任务模式新增无限画布，用户可将素材、方案和创作成果集中展示、比较与整理，并继续调整文字、配色和布局；产品同时接入豆包 2.1 Lite 和 Seedream 5.0 Flash，分别服务文档、表格、PPT 等办公任务及图像生成。

图32：豆包工作画布：将参考素材与多个设计方案集中展示并继续调整



资料来源：mp.weixin.qq.com，浙商证券研究所

TRAE 整合 TraeWork 与 TraeCode，统一开发与办公工作流。TRAE 将开发与办公任务整合到同一平台。10 月 9 日，TRAE 宣布合并 TraeWork 与 TraeCode，开始灰度开放 Agent 与 IDE（集成开发环境）模式切换，并覆盖桌面、网页和移动端；用户可在编程和通用办公任务之间延续既有账号、历史会话、本地数据与付费权益。

豆包工作支持导入外部对话与记忆，降低迁移成本。豆包工作通过导入对话、记忆及工作资料，减少用户切换 AI 工具时重新交代任务背景的成本。10 月 14 日，产品上线数据

仅供内部参考，请勿外传

导入与同步，支持从 ChatGPT、Claude Code、TraeWork 等产品迁入对话及记忆，并承接指令、技能、连接器和项目资料；新增“在新对话中继续”、批注与摘要侧栏，便于拆分长任务和复用已有成果。

豆包鸿蒙电脑端上线，补齐桌面端覆盖。豆包将桌面端覆盖扩展至鸿蒙电脑。10 月 7 日，豆包鸿蒙电脑客户端上线应用市场，首版本搭载豆包 2.1 Turbo 及 Pro，提供聊天与工作模式，支持深度研究、PPT 制作、写作、生图和云盘等功能。

豆包手机助手逐步开放第三方屏幕自动化。豆包手机助手开始按应用声明开放屏幕自动化，头部应用的拒绝名单限制了部分高频任务的覆盖范围。10 月 14 日，SAEP（屏幕自动化操作声明协议）的 30 天公示期结束，未明确拒绝的第三方应用将逐步开放操作权限；已声明拒绝的应用不会被自动操作。微信、QQ、淘宝、天猫、支付宝、高德地图、拼多多、美团，以及 Kimi、DeepSeek 等均列在拒绝名单中。开发者还可按应用、页面或操作类型设置权限，并在公示期结束后申请退出。

Seed 据报设立安全与对齐部门，强化 Agent 风险治理。字节 Seed 将模型安全与对齐设为独立部门，覆盖智能体执行及前沿模型风险。新浪科技 10 月 13 日报道，Seed 通过内部邮件宣布成立 Seed Model-Alignment Preparedness 部门，由严林负责并向吴永辉汇报，原 Responsible AI 团队与部分 Chat 团队成员转入，研究范围包括智能体模型安全及化学、生物、放射、核等领域。

OmniHuman v2 上线 fal，人物视频生成按秒计费。OmniHuman v2 是字节跳动的人物视频生成模型，输入一张人物图片和一段音频，即可生成与音频对应的表情、动作和人物视频。10 月 15 日（北京时间 10 月 16 日），fal 上线该模型的商业调用服务，标注为合作伙伴模型，按生成视频时长收费，每秒 0.105 美元。人物形象与音频可分别准备，便于将口播、配音等素材制作成数字人视频。

4.5 阿里巴巴

Qwen-Image-2.1-Turbo 开放权重，采用 8 步采样。Qwen-Image-2.1-Turbo 以 8 次去噪迭代完成图像生成：模型从噪声状态逐步还原出画面，“8 步采样”指这一生成过程迭代 8 次。10 月 9 日，千问开放该加速版的模型权重和推理示例，视觉生成模块为 70 亿参数，统一支持文生图和图像编辑，并提供透明背景、2K 分辨率与多图参考等能力。权重可免费用于研究和评估，商业用途需另行向千问申请授权；开发者也可调用已上线的 Pro 与 Turbo 托管 API。

Qoder 上线 Fast 模式，缩短响应时间并降低 Credits 消耗。Qoder 通过 Fast 模式同时缩短等待时间与降低额度消耗。10 月 10 日发布的 0.4.4 及以上桌面版本开始支持该模式，覆盖个人版、企业版和国内外版本。官方对比中，国际版原 Performance 模式升级后首次响应由 8 秒缩短至 3 秒，Credits 消耗系数由 1.1 倍降至 0.8 倍，降低约 27%；后续计划接入其他 Qoder 产品。

苹果折叠屏国行备案增加通义模型编号。通义模型进入苹果国行终端的部署准备环节。10 月 14 日披露的备案信息显示，国行 iPhone Duo 于 9 月 29 日换证时增加 Apple 智能、通义万相和通义千问备案编号；证券时报援引阿里方面信息称，千问将集成至 Apple 智能，为中国用户提供文本、图像理解与内容生成能力。

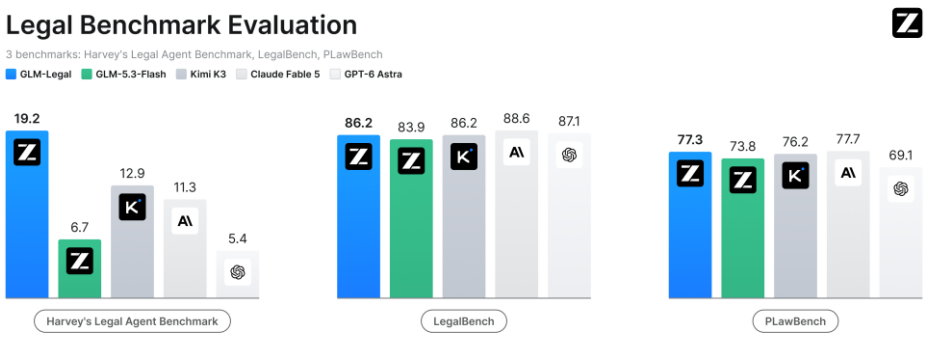
4.6 智谱

智谱推出 GLM-Legal，法律专业任务与轻量化模型结合。智谱以轻量基座结合专业训练，推进法律 Agent 从问答走向完整任务交付。10 月 12 日发布的 GLM-Legal 基于

仅供内部参考，请勿外传

GLM-5.3-Flash，先以法律语料补充专业知识，再通过数万个法律环境的强化学习与 Agent 运行框架协同优化，覆盖合同审阅、证据分析和跨境合规。其在 Harvey LAB（覆盖 24 个法律实务领域、要求任务全部标准均通过的评测）中取得 19.2%的通过率，较基座提高 12.5 个百分点；在 LegalBench（法律推理测试集）和 PLawBench（中文法律咨询、案例分析及文书撰写评测）中也达到较强表现。Harvey LAB 与 LegalBench 结果由 Vals 独立评测。API 已开放测试申请，后续计划接入 AutoClaw。

图33： GLM-Legal 在法律任务评测中的表现（Harvey LAB、LegalBench 由 Vals 独立评测；PLawBench 为中文法律实务评测）



资料来源：z.ai，浙商证券研究所

4.7 MiniMax

MiniMax Design Agent V3 上线，增加跨会话素材和上下文复用。MiniMax Design Agent V3 以素材和上下文复用支持连续创作。10 月 14 日，官方宣布新版可在同一创作页面的不同对话间共享素材与上下文，并支持自定义 Agent 指令，保留用户的风格和工作习惯。公司披露创作成本降低 30%以上，用户可沿用已有素材继续修改，减少重复上传、描述与生成。

FastH3-Omni 开放预览，计划 10 月 20 日开放权重与训练代码。MiniMax H3 的第三方加速生态扩展至图像与音频参考。10 月 14 日，Nuva Lab 与 Reactor 开放 FastH3-Omni 预览，该版本由 FastVideo 联合两家合作方进行后训练，支持参考图、参考音频、首尾帧约束和片段续接，可输出约 5—15 秒、24 帧/秒的音画同步视频。实时模式最多输入 3 张参考图，非实时模式最多 9 张，音频参考最多 3 段；Reactor 提供托管 API，Nuva Lab 提供专属部署。团队计划 10 月 20 日开放权重和训练代码，并支持 DGX Spark 与 RTX GPU。

fal 基于 H3 Max 推出视频扩画，保留原视频音轨与时间结构。fal 基于 MiniMax H3 Max 推出 Reframe，将既有视频扩展为不同画幅，同时保留原素材、音轨和时间结构。10 月 14 日上线的服务支持通过文字指引，将竖屏扩展至宽屏或方形画面，适配不同平台的内容发布。fal 提供 480p 与 768p 两档输出，每生成秒分别收费 0.05 美元和 0.08 美元；长视频分窗口处理，窗口重叠部分也计入生成时长。

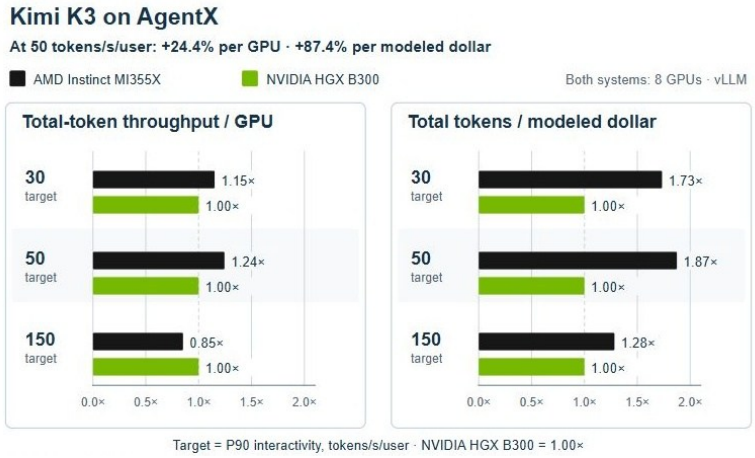
MiniMax 开设天猫官方旗舰店，订阅销售拓展至电商渠道。MiniMax 将模型订阅销售扩展至天猫。第一财经 10 月 13 日报道，MiniMax 官方旗舰店已上架 M Plan 的 Go、Explore 和 Build 三档订阅，支持月付及年付。此前智谱、阿里云和千问办公等已在天猫开店，模型服务的购买渠道从开发者官网扩展至消费电商平台。

仅供内部参考，请勿外传

4.8 Kimi

Kimi K3 的 AMD 推理栈优化提供更多硬件部署选择。AMD 与 Embedded LLM 优化 Kimi K3 的推理部署，提高 AMD 硬件上单位资源可承载的 Agent 工作量。10 月 16 日发布的八卡 MI355X 测试显示，在 AgentX（模拟 Agent 调用负载的推理系统评测）中，将 90%请求的交互速度门槛设为每用户 50 Token/秒时，每 GPU 总吞吐为 7567 Token/秒，高于 HGX B300 的 6083 Token/秒，增幅 24.4%；按测试采用的总持有成本模型，每美元处理 Token 数提高 87.4%。总吞吐包含复用输入和生成输出，费用比较采用硬件持有成本。

图34： Kimi K3 在 MI355X 上的推理性能（AMD 测试，原图保留快照日期与配置）



资料来源：vllm.ai，浙商证券研究所

4.9 DeepSeek

DeepSeek 新一轮融资规模有望接近 1000 亿元。DeepSeek 新一轮融资规模向千亿元扩大。彭博 10 月 6 日报道，该轮已接近筹得至少 800 亿元，按已签署条款计算最终可能接近 1000 亿元，超过最初 500 亿元目标，腾讯和宁德时代是主要出资方；《财经》10 月 13 日进一步报道，该轮拟融资约 1000 亿元、已临近完成。

DeepSeek API 中断后宣布返还 7 日消耗赠金。DeepSeek 对 API 故障提供 7 日使用额的赠金补偿。10 月 15 日，公司开放平台通知，北京时间 08:11—13:04 发生系统故障，API 服务中断；将按各账号 10 月 8—14 日的消耗金额全额返还赠金，自动到账并优先抵扣后续 API 费用，同时排查并剥离单点依赖。

仅供内部参考，请勿外传

5 风险提示

商业化增长及盈利改善不及预期。客户预算、模型定价和算力投入共同影响收入与现金流，增长放缓或成本上升可能影响盈利兑现。

端侧部署及安全产品落地不及预期。异构设备的性能、兼容性和权限管理可能限制部署效率，实验能力向产品转化存在不确定性。

个人 Agent 服务拓展不及预期。购物、预约及理赔依赖外部平台接入、用户授权与执行可靠性，合作和产品推广进度可能影响使用规模。

行业竞争加剧及数据安全风险。模型、应用与渠道竞争可能影响商业回报，跨设备运行和外部服务连接增加数据保护与合规要求。

仅供内部参考，请勿外传

股票投资评级说明

以报告日后的 6 个月内，证券相对于沪深 300 指数的涨跌幅为标准，定义如下：
1.买 入 ： 相对于沪深 300 指数表现 + 20％以上；
2.增 持 ： 相对于沪深 300 指数表现 + 10％ ~ + 20％；
3.中 性 ： 相对于沪深 300 指数表现 - 10％ ~ + 10％之间波动；
4.减 持 ： 相对于沪深 300 指数表现 - 10％以下。

行业的投资评级：

以报告日后的 6 个月内，行业指数相对于沪深 300 指数的涨跌幅为标准，定义如下：
1.看 好 ： 行业指数相对于沪深 300 指数表现 + 10%以上；
2.中 性 ： 行业指数相对于沪深 300 指数表现 - 10% ~ + 10%以上；
3.看 淡 ： 行业指数相对于沪深 300 指数表现 - 10%以下。
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。投资者不应仅仅依靠投资评级来推断结论。

法律声明及风险提示

本报告由浙商证券股份有限公司（已具备中国证监会批复的证券投资咨询业务资格，经营许可证编号为：Z39833000）制作。本报告中的信息均来源于我们认为可靠的已公开资料，但浙商证券股份有限公司及其关联机构（以下统称“本公司”）对这些信息的真实性、准确性及完整性不作任何保证，也不保证所包含的信息和建议不发生任何变更。本公司没有将变更的信息和建议向报告所有接收者进行更新的义务。
本报告仅供本公司的客户作参考之用。本公司不会因接收人收到本报告而视其为本公司的当然客户。
本报告仅反映报告作者的出具日的观点和判断，在任何情况下，本报告中的信息或所表述的意见均不构成对任何人的投资建议，投资者应当对本报告中的信息和意见进行独立评估，并应同时考量各自的投资目的、财务状况和特定需求。对依据或者使用本报告所造成的一切后果，本公司及/或其关联人员均不承担任何法律责任。
本公司的交易人员以及其他专业人士可能会依据不同假设和标准、采用不同的分析方法而口头或书面发表与本报告意见及建议不一致的市场评论和/或交易观点。本公司没有将此意见及建议向报告所有接收者进行更新的义务。本公司的资产管理公司、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的意见或建议不一致的投资决策。
本报告版权均归本公司所有，未经本公司事先书面授权，任何机构或个人不得以任何形式复制、发布、传播本报告的全部或部分内容。经授权刊载、转发本报告或者摘要的，应当注明本报告发布人和发布日期，并提示使用本报告的风险。未经授权或未按要求刊载、转发本报告的，应当承担相应的法律责任。本公司将保留向其追究法律责任的权利。

浙商证券研究所

上海总部地址：杨高南路 729 号陆家嘴世纪金融广场 1 号楼 25 层
北京地址：北京市东城区朝阳门北大街 8 号富华大厦 E 座 4 层
深圳地址：广东省深圳市福田区广电金融中心 33 层
上海总部邮政编码：200127
上海总部电话：(8621) 80108518
上海总部传真：(8621) 80106010
浙商证券研究所：https://www.stocke.com.cn