下载日志已记录，仅供内部参考

计算机

报告日期：2026 年 10 月 24 日

FDE 助力企业 AI 落地与推广

—— 2026.10.17—2026.10.24 大模型商业化周报

投资要点

- **本周聚焦企业交付、推理经济性及国产模型更新。**企业 AI 竞争更加重视业务落地与效率；推理优化支撑 Token 降价，高端算力需求仍支撑租金；国产模型能力与成本表现持续改善，商业化竞争力进一步增强。
- **前沿模型厂商企业客户留存较强，FDE 助力业务集成与推广。**YipitData 样本显示，Anthropic 新增客户采用增长至 9 月趋缓，但第 12 个月客户留存率仍达 87%。这意味着，在持续获取新客户之外，让已有客户将 AI 用于更多任务、推广至更多团队，正成为重要增长来源。业务集成与可靠交付因此更加重要：模型需要接入企业系统、满足实际流程要求，才能从试点工具成为持续使用的生产工具。OpenAI 的 FDE 团队通过现场开发、业务评测和反馈改进推动这一过程。在半导体客户项目中，约 15 名 FDE 推动 10 余项用例在 3—6 个月内进入生产，定向测试耗时减少 48%，内核优化性能中位数提高 22%。随后，团队以 Codex 承载共用能力，将定制应用改为 Skills 和插件，减少约 90%的定制代码；一个案例首次部署耗时两个月，已有方案推广至第二个团队仅用一天。FDE 价值既在于完成首次落地，也在于将交付经验沉淀为可复用能力，降低后续推广成本，支持客户持续扩大使用。
- **推理优化为 Token 降价提供成本基础，高端 GPU 租金仍受供需约束。**DeepSeek 在开源推理框架 DeepInfer 时披露，按谷价折算，其生产集群每 GPU 小时收入为 38 美元。若按 B300 租金 7 美元/小时计算，仅扣 GPU 租金的推理毛利率约为 82%（未计入网络、存储及运维等成本）。输入与输出计算分工、长上下文缓存复用等优化，提高了单位硬件产出，使较低 API 售价仍能覆盖算力租赁成本并留下较大收入余量。截至 10 月 22 日，Silicon Data 的 Token 价格指数降至 0.94 美元/百万 Token，B300 租金较 5 月 31 日上涨 51.9%。其分析指出，新一代 GPU 供给仍在追赶需求，大型实验室的长期预订进一步压缩按需市场可用容量。两类价格分化反映出：模型服务可以通过提高每卡产出降低单位 Token 成本，而训练、推理需求及高端供给约束仍可支撑 GPU 租金。
- **国产模型密集更新。**Kimi K3.1 Pro 在编程、3D 创作和金融研究等复杂任务上提升明显，AA 智能指数较 K3 提高 11 分至 55 分、位列总榜第 4，AA 测算的单位任务成本与 K3 大体相当；DeepSeek V4.1 Flash 1017 强化完整游戏和 3D 项目开发，AA 智能指数提高 6 分至 45 分，同时保持调用价格不变；MiniMax M3.1 Flash 则侧重执行效率，官方称推理速度接近原来的 3 倍、Agent 用量成本降低 70%，并于 10 月 23 日开放权重。能力提升让模型承接更复杂的工作，速度与成本改善扩大高频使用空间，开放权重进一步增加部署选择，中国厂商在 OpenRouter 的 Token 支出份额继续提升，国产模型的竞争力正进一步体现为开发者平台上的实际付费使用。
- **风险提示**

企业 AI 落地不及预期；模型迭代与推理优化不及预期；算力供需变化；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 《世界模型：基模能力提升，物理 AI 加速》 2026.10.22

2 《Anthropic 三季报解读》 2026.10.18

3 《端侧 AI，万众创新》 2026.10.10

下载日志已记录，仅供内部参考

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 企业客户留存较强，业务集成与交付能力的重要性提升 5
2.1.1 新增采用增长趋缓，较高留存为深化客户使用提供基础 5
2.1.2 FDE 完成现场适配，并将交付经验转为可复用方案 6
2.1.3 企业扩大部署范围，服务伙伴补充行业交付能力 6
2.2 推理效率改善支撑低价服务，高端 GPU 租金仍受供需约束 7
2.2.1 DeepInfer 通过计算分工与缓存复用提高单卡产出 7
2.2.2 Token 售价与 GPU 租金分化 7
2.3 国产模型能力与成本改善，平台支出份额提升 8
2.3.1 Kimi K3.1 Pro 强化复杂任务，独立评测智能指数升至 55 分 8
2.3.2 DeepSeek V4.1 Flash 1017 强化完整项目开发，价格不变 9
2.3.3 MiniMax M3.1 Flash 提速降本，并开放模型权重 10
2.3.4 中国厂商在 OpenRouter 的 Token 支出份额提升 11
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 字节跳动/豆包/火山引擎 16
4.4 阿里巴巴/通义/千问 18
4.5 DeepSeek 19
5 风险提示 19

下载日志已记录，仅供内部参考

图表目录

图 1: Anthropic 分行业采用率：科技与金融客户采用较多 5
图 2: Anthropic 企业客户留存率与样本客户群数量 5
图 3: Anthropic 有效 Token 价格变化 5
图 4: Anthropic 与 OpenAI 在样本内的支出份额 5
图 5: OpenAI FDE：企业现场交付与可复用能力建设 6
图 6: 火山引擎 FDE 团队服务内容 7
图 7: Token 价格下降与 GPU 租金上涨并存 8
图 8: Kimi K3.1 Pro 官方能力比较（一） 8
图 9: Kimi K3.1 Pro 官方能力比较（二） 8
图 10: Artificial Analysis 智能指数：K3.1 Pro 得分 55 9
图 11: DeepSeek V4.1 Flash 1017 官方评测对比 9
图 12: DeepSeek 任务性能与成本变化 9
图 13: DeepSeek 使用 Godot 复刻游戏的项目示例 10
图 14: DeepSeek 故宫 3D 建模项目示例 10
图 15: DeepSeek 在 Artificial Analysis 的评测表现 10
图 16: MiniMax M3.1 Flash 官方能力评测 11
图 17: MiniMax 任务成本与表现对照 11
图 18: OpenRouter 模型厂商 Token 支出份额 11
图 19: Claude Duet 产品示例 12
图 20: Anthropic 初级技术岗位占比变化 14
图 21: 交互式 Agent 与员工工时比较 14
图 22: 公众与 Anthropic 员工对 AI 的预期 14
图 23: 不同群体的 AI 使用反馈 14
图 24: Draft 模式不同试生成次数的成本对照 17
图 25: 同等预算下 Draft 模式可尝试更多次数 17
图 26: 豆包导航功能示例 18
图 27: Meoo 与千问 Go 应用生成服务 18
图 28: 千问陪逛通话入口 19

表 1: 本周模型与服务更新 4
表 2: 本周企业应用与推理经济性数据 4

下载日志已记录，仅供内部参考

1 本周模型发布、商业化数据总结

表1： 本周模型与服务更新

公司	模型或服务	主要能力变化	商业化进展
Kimi	K3.1 Pro	100 万 Token 上下文；AA 智能指数 55 分	标准版输入/输出 13/68 元，高速版 26/136 元（每百万 Token）
DeepSeek	V4.1 Flash 1017	强化完整游戏与 3D 项目开发；AA 智能指数 45 分	API 谷价输入/输出 1/4 元，峰价 2/8 元（每百万 Token）。Search、Fetch 均为 5 元/千次成功请求。
MiniMax	M3.1 Flash	官方称速度接近原来的 3 倍、Agent 用量成本降低 70%	输入/输出 1/4 元（每百万 Token）；10 月 23 日开放权重
字节跳动	Seedance 2.5	Draft 降低试生成成本，4K 扩展影视制作	4K 面向部分企业试用；Managed Agent 整合创作流程
Anthropic	Claude Duet、企业版	持续任务、聊天与云端 Cowork 整合	Duet 研究预览；企业版统一工作入口
OpenAI	Decisions API	GPT-6 Luna 处理分类、分流和行动选择	正式开放；输入/缓存输入 0.10/0.01 美元/百万 Token；输出免费；单次最多 400 个问题

资料来源：公司官方资料、Artificial Analysis，浙商证券研究所

表2： 本周企业应用与推理经济性数据

主体	指标	本周披露	统计范围
YipitData	Anthropic 客户留存	第 12 个月 87%；第 24 个月 95%	客户留存率，不同月龄包含的客户群数量不同
OpenAI FDE	半导体客户交付	约 15 名工程师推进 10 余项用例，3—6 个月进入生产	定向测试耗时-48%；内核优化性能中位数+22%
DeepSeek	单卡小时收入	按谷价折算 38 美元/GPU 小时	生产集群披露；单集群日处理 155 万亿 Token
Silicon Data	Token 与 GPU 价格	Token 指数 0.94 美元/百万 Token; B300 租金较 5 月 31 日+51.9%	Token 数据截至 10 月 21 日；GPU 租金截至 10 月 22 日

资料来源：YipitData、OpenAI、DeepSeek、vLLM、Silicon Data，浙商证券研究所

下载日志已记录，仅供内部参考

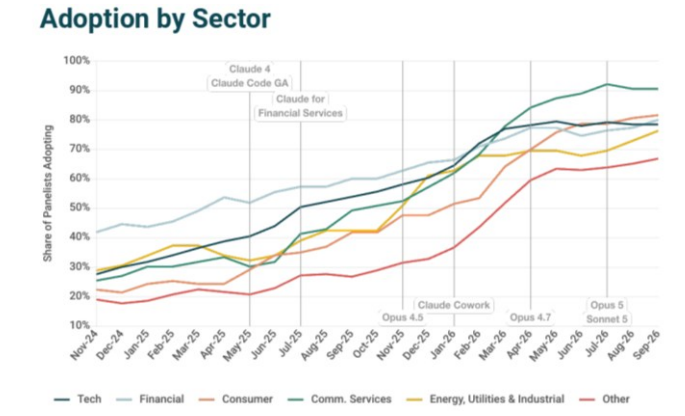
2 本周重点事件

2.1 企业客户留存较强，业务集成与交付能力的重要性提升

2.1.1 新增采用增长趋缓，较高留存为深化客户使用提供基础

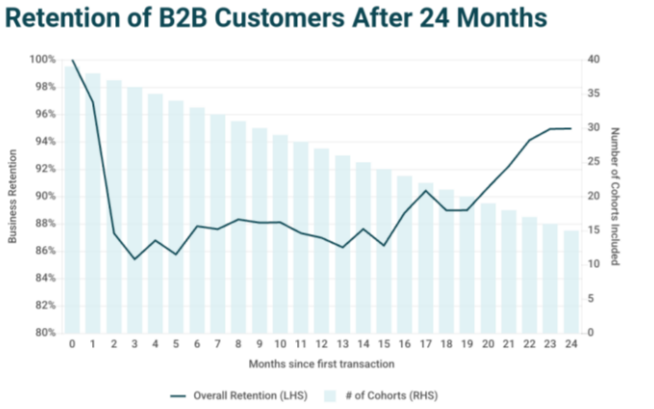
Anthropic 新增客户采用增长趋缓，已有客户则保持较高留存。YipitData 于 10 月 22 日发布、23 日更新的上市前分析，覆盖 1500 家中小企业、200 家大型企业及数百万消费者，追踪超过 40 亿美元年度 AI 支出。样本中，Anthropic 年内采用率明显提高，至 9 月增长趋于平缓；科技和金融客户贡献主要支出，第 12 个月客户留存率为 87%。持续使用的客户为推广更多业务应用提供了基础。

图1： Anthropic 分行业采用率：科技与金融客户采用较多



资料来源：YipitData，浙商证券研究所

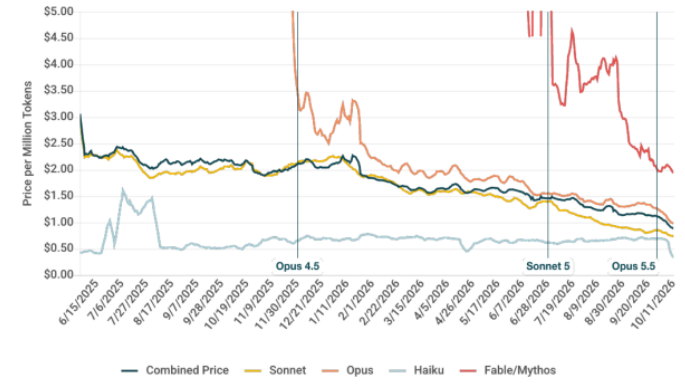
图2： Anthropic 企业客户留存率与样本客户群数量



资料来源：YipitData，浙商证券研究所

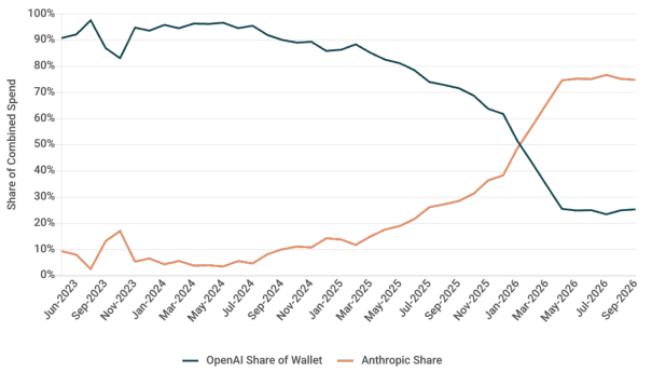
客户支出集中于少数重度用户，深化已有客户的业务应用或成为重要增长方向。YipitData 样本显示，前 1%的客户贡献 Anthropic 近 50%的支出。有效 Token 价格下降的同时，Anthropic 相对 OpenAI 的支出份额在 6 月后趋于稳定。除了获取新客户，帮助已有客户将 AI 用于更多业务流程、推广至更多团队，也有望增加使用量和单客支出。业务集成与交付服务是实现这一增长路径的重要环节。

图3： Anthropic 有效 Token 价格变化



资料来源：YipitData，浙商证券研究所

图4： Anthropic 与 OpenAI 在样本内的支出份额



资料来源：YipitData，浙商证券研究所

扩大企业使用，需要把模型接入实际业务。专业任务往往涉及企业内部数据、现有软件和业务规则，模型输出还需要员工复核。FDE 负责完成这些连接和适配，并用实际任务检验效果。OpenAI 本周披露的案例展示了这一过程。

下载日志已记录，仅供内部参考

2.1.2 FDE 完成现场适配，并将交付经验转为可复用方案

OpenAI 的前线部署工程师（FDE）与客户共同确定任务目标和验收标准，再将模型接入企业系统。工程师需要明确调用哪些工具、如何检查结果，以及员工怎样纠正错误。能够复用的功能交给产品团队，模型尚不能完成的任务反馈给研究团队，使后续项目可以利用前期积累，减少重复开发。

半导体客户案例显示，FDE 推动模型进入研发流程并取得具体效果。OpenAI 安排约 15 名 FDE 推进 10 余项用例，3—6 个月内进入生产，定向测试生成耗时减少约 48%，内核优化性能中位数提高 22%，部分调试任务从数周缩短至数小时。团队随后以 Codex 为统一运行基础，将定制应用改为技能和插件，删除约 90%的定制代码。2026 年 8 月的一个案例中，首次部署耗时两个月，已有方案扩展至第二个团队仅用一天。将共用功能做成组件后，新增团队可复用已有工具和流程，工程师只需处理新的业务差异。

图5： OpenAI FDE：企业现场交付与可复用能力建设



资料来源：OpenAI，浙商证券研究所

专业人员的反馈可以用于持续改进任务流程。OpenAI 本周回顾与 Thrive Holdings、Crete 会计师事务所合作的 Tax AI：试点处理 7000 份报税表，处理量提高约 50%；早期案例披露准备时间节省约三分之一。会计师指出的错误被转为测试案例，再由 Codex 修改工作流并验证改进效果。这使后续任务能够利用此前的纠错经验，减少同类问题重复发生。

2.1.3 企业扩大部署范围，服务伙伴补充行业交付能力

企业客户正将已验证的工具推广至更多员工。Uber 于 10 月 22 日披露，将 ChatGPT Work 由约 1000 人试点推广至全球超过 3 万名员工，连接 Google Drive、Gmail、Calendar 和 Slack，并配套内部技能库及培训。接入已有办公软件让 Agent 可以获取任务所需资料，共用技能和培训则帮助员工掌握具体用法。

HCLTech 本周以 Global Premier 身份加入 Claude Partner Network，设立专项业务部门及卓越中心，将 Claude 嵌入软件产品和 AI Force 平台。其服务网络覆盖 60 个国家，拥有超过 23 万名员工，可为银行、医院、制造及电信客户提供交付支持。**国内火山引擎 FDE 服务通过中国信通院专项评估，首批获评“先进级”。**模型厂商提供产品，服务伙伴承担行业适配和实施，扩大企业部署覆盖。

下载日志已记录，仅供内部参考

图6：火山引擎 FDE 团队服务内容



资料来源：火山引擎官方公众号，浙商证券研究所

2.2 推理效率改善支撑低价服务，高端 GPU 租金仍受供需约束

2.2.1 DeepInfer 通过计算分工与缓存复用提高单卡产出

DeepSeek 开源 DeepInfer 推理核心及部署说明，展示提高生产集群效率的具体方法。框架支持 NVIDIA 与华为昇腾硬件，普通 8 卡服务器负责输入计算，64—128 卡超节点负责输出生成，使两类任务使用更合适的硬件配置。100PiB SSD 缓存保存可复用上下文，Agent 缓存保留 3 天、聊天缓存保留 7 天，减少重复输入的计算量。计算分工和缓存复用提高单位硬件产出，为降低 Token 服务成本提供支撑。

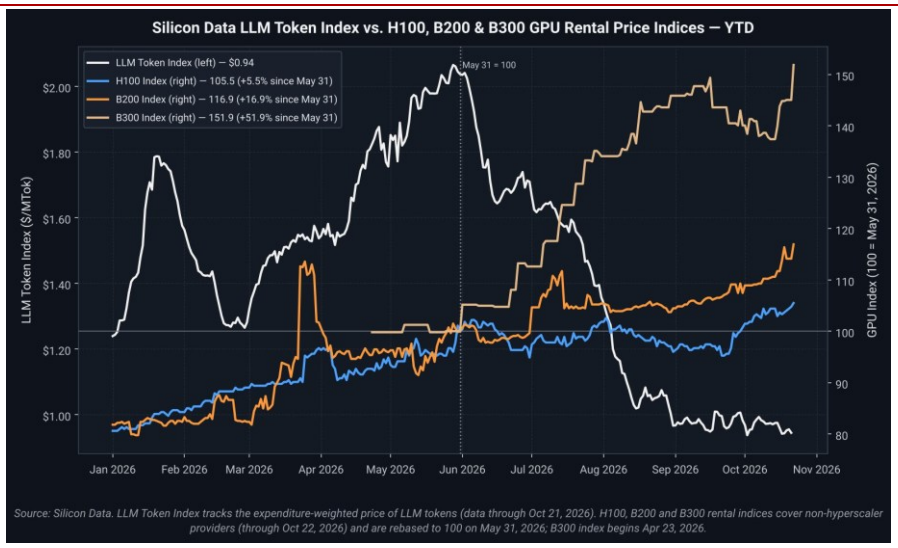
低 API 价格与较高单卡收入可以同时实现。DeepSeek 披露，单集群每天处理 155 万亿 Token，按谷价折算每 GPU 小时收入 38 美元；谷价为每百万 Token 缓存命中输入 0.02 元、未命中输入 1 元、输出 4 元。机器之心/36 氪据此按 B300 租金 7 美元/GPU 小时测算，仅扣 GPU 租金后的毛利率约 81.6%（未计入网络、存储和运维成本）。提高吞吐并减少重复计算，可在较低单价下增加单位硬件收入。

2.2.2 Token 售价与 GPU 租金分化

Token 价格下降的同时，新一代 GPU 租金涨幅较高。Silicon Data 于 10 月 23 日发布的数据显示，截至 10 月 21 日，支出加权 LLM Token 价格指数为 0.94 美元/百万 Token；截至 10 月 22 日，H100、B200、B300 租金较 5 月 31 日分别上涨 5.5%、16.9%和 51.9%。

下载日志已记录，仅供内部参考

图7： Token 价格下降与 GPU 租金上涨并存



资料来源：Silicon Data，浙商证券研究所

注：右轴以 2026 年 5 月 31 日为 100，租金覆盖非超大规模云厂商

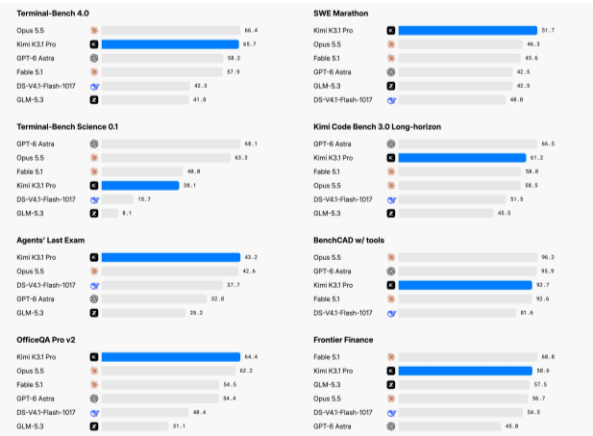
高端 GPU 可用容量偏紧，支撑租金上涨。Silicon Data 对 B200 的分析指出，新平台供给仍在追赶需求，大型 AI 实验室又通过多年预订提前锁定容量，压缩了按需市场的可用供给。训练和推理都需要高端算力，其租金不会随模型售价同步下降。只要产出效率提升得更快，租金上涨时单位成本仍可下降。推理优化支撑 Token 降价，与高端算力供需偏紧同时发生。

2.3 国产模型能力与成本改善，平台支出份额提升

2.3.1 Kimi K3.1 Pro 强化复杂任务，独立评测智能指数升至 55 分

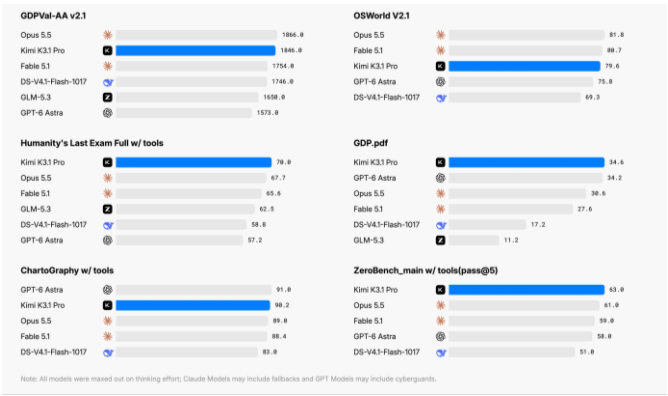
Kimi 于 10 月 24 日发布 K3.1 系列首个模型 K3.1 Pro，强化编程、3D 创作、金融研究及多步骤 Agent 任务。模型为 3 万亿级参数，支持 100 万 Token 上下文；标准版每百万 Token 输入、输出价格分别为 13 元、68 元，缓存命中输入为 1.3 元。模型已上线 Kimi 网页、App、Work、Code 及 API，完整权重计划 11 月 4 日前发布。Highspeed 版输出速度约为标准版的 2—3 倍，输入、输出价格分别为 26 元、136 元/百万 Token，适合对响应速度要求更高的任务。

图8： Kimi K3.1 Pro 官方能力比较（一）



资料来源：Kimi 官方技术博客，浙商证券研究所

图9： Kimi K3.1 Pro 官方能力比较（二）

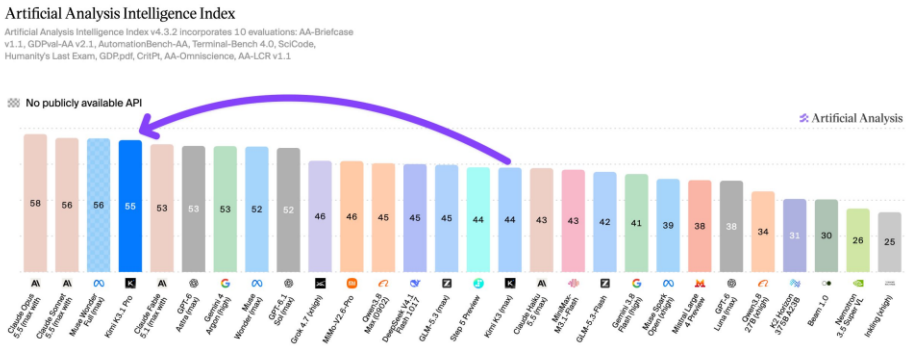


资料来源：Kimi 官方技术博客，浙商证券研究所

下载日志已记录，仅供内部参考

K3.1 Pro 在独立评测中提高了任务表现，同时保持相近的任务成本。Artificial Analysis 智能指数为 55 分，较 K3 提高 11 分，位列当时总榜第 4；GDPval-AA 为 1846 Elo、排名第 3，AutomationBench-AA 从 58%升至 73%。AA 配置下 Terminal-Bench 4.0 为 67%，单位任务成本约 1.98 美元，与 K3 大体相当。

图10： Artificial Analysis 智能指数：K3.1 Pro 得分 55



资料来源：Artificial Analysis，浙商证券研究所

2.3.2 DeepSeek V4.1 Flash 1017 强化完整项目开发，价格不变

DeepSeek 于 10 月 17 日更新 V4.1 Flash 1017，网页、App 和 API 同步升级，调用方式与价格不变。模型通过自反馈强化学习提高从零搭建游戏和 3D 项目的能力，并缩短任务耗时及输出长度，适配 DeepSeek Harness、CodeBuddy 和 WorkBuddy。官方评测中，HLE 无工具为 42.6、Terminal-Bench 4.0 为 42.3、DeepSWE v1.1 为 77.0。

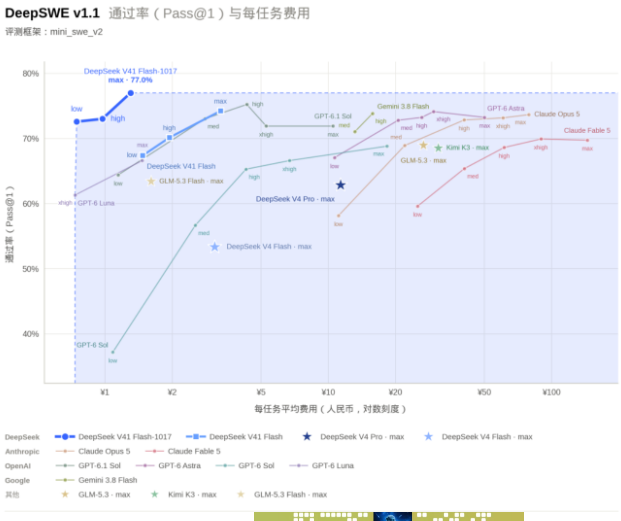
图11： DeepSeek V4.1 Flash 1017 官方评测对比

Benchmarks	DeepSeek V4.1-Flash-1017	DeepSeek V4.1-Flash	Kimi K3	GLM 5.3	GPT 5.6-Sol	Claude Opus 5
HLE (w/o tools)	42.6/44.0*	36.8/39.1*	43.5	42.0*	44.5	**56.3**
Terminal-Bench 4.0	42.3	31.2	12.6	37.9	39.9	**51.8**
DeepSWE v1.1	**77.0**	74.2	67.5	66.9	73.0	74.0
ProgramBench	29.1	20.3	17.5	19.0	23.0	**37.0**
NL2Repo-Bench	68.2	65.4	58.0	58.0	56.8	**75.3**
Automation-Bench V1.0.6	**60.5**	54.8	46.7	48.8	45.8	50.3
Agents' Last Exam	**35.4**	31.8	27.6	28.5	26.7	28.6
Apex Bench v1.1	**72.1**	64.6	47.8	56.6	54.3	65.8
Chartography (w/tools)	83.0	78.9	68.1	-	79.9	**84.0**
BabyVision (w/tools)	92.5	89.6	85.7	-	88.9	**94.1**
ZeroBench-main (w/tools)	51.0	49.0	41.0	-	**53.0**	52.0

*代表仅测试 HLE 的文本子集；我们在对 Terminal-Bench 4.0 进行的测试中使用 Claude Code 框架，DeepSWE 使用 Mini-SWE-V2 框架；ZeroBench-main 使用 Pass@5 评测。

资料来源：DeepSeek 官方公众号，浙商证券研究所

图12： DeepSeek 任务性能与成本变化



资料来源：DeepSeek 官方公众号，浙商证券研究所

官方案例展示了从编写代码到完成可交互项目的能力。模型完成故宫 3D 建模、网页赛车、使用 Godot 复刻《传送门 2》及 Valley of Ages 等项目，将代码编写、场景搭建和运行调试串联起来。

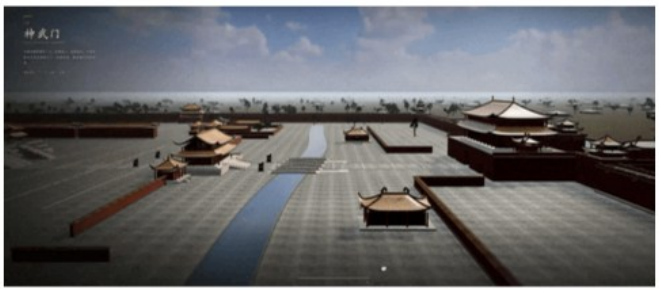
下载日志已记录，仅供内部参考

图13： DeepSeek 使用 Godot 复刻游戏的项目示例



资料来源：DeepSeek 官方公众号，浙商证券研究所

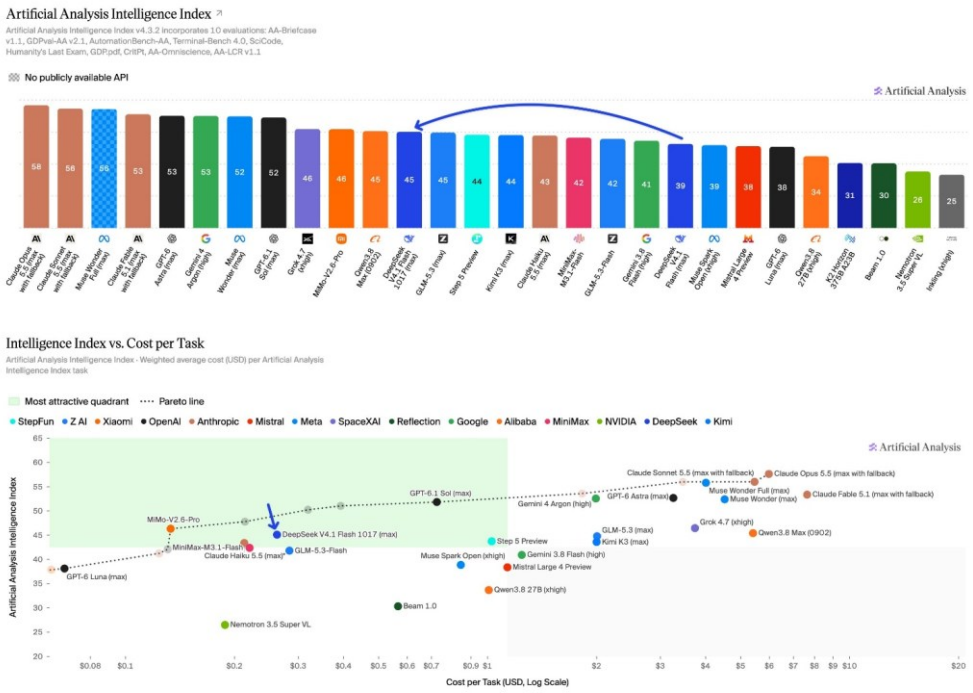
图14： DeepSeek 故宫 3D 建模项目示例



资料来源：DeepSeek 官方公众号，浙商证券研究所

DeepSeek V4.1 Flash 1017 在 Artificial Analysis 智能指数中取得 45 分，较前版提高 6 分，每项任务成本为 0.26 美元。

图15： DeepSeek 在 Artificial Analysis 的评测表现



资料来源：Artificial Analysis，2026 年 10 月 20 日，浙商证券研究所

2.3.3 MiniMax M3.1 Flash 提速降本，并开放模型权重

MiniMax M3.1 Flash 通过提速和降本扩大 Agent 使用空间。模型于 10 月 19 日发布、23 日开放权重，总参数 4280 亿、每 Token 激活 230 亿；官方称在线推理速度接近原来的 3 倍，Agent 用量成本降低 70%。API 价格为每百万 Token 输入 1 元、输出 4 元、缓存输入 0.1 元，已接入 API、M Plan 及 MiniMax Code。官方披露，上线 72 小时处理真实用户 Token 超过 100 万亿；华为、平头哥、昆仑芯、NVIDIA、AMD 等硬件生态及多家推理云平台提供适配支持，扩大了模型的部署选择。

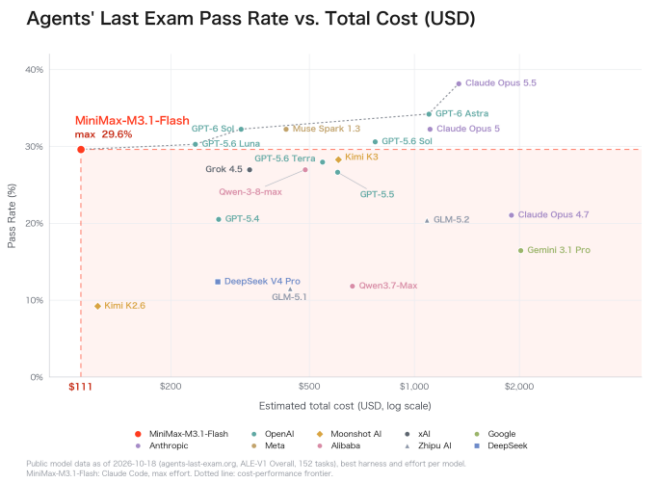
下载日志已记录，仅供内部参考

图16： MiniMax M3.1 Flash 官方能力评测

		MiniMax M3.1 Flash	DeepSeek-V4.1-Flash-1017	DeepSeek-V4.1-Flash	Kimi-K3	GLM-5.3	GPT-5.6 Sol	Opus 5
CoWork	Automation-Bench v1.0.6	**62.8**	60.5	54.8	46.7	48.8	45.8	50.3
	BankerToolBench	**76.2**	-	73.7	69.1	74.5	74.6	**76.2**
	GDP.pdf AA	20.8	-	12.8	22.0	11.2	**27.2**	21.6
	Terminal Bench 4.0	38.1	42.3	31.2	12.6	37.9	39.9	**51.8**
	LOCA-Bench 256K	**60.3**	-	-	49.7	50.3	60.0	53.7
Computer Use	Agents' Last Exam	29.6	-	-	28.3	-	30.6	**32.2**
	OSWorld 2.0	61.7	-	54.3	58.3	-	65.7	**70.6**
Coding	PostTrain Bench v1.1	39.3	-	30.5	36.6	37.5	36.2	**40.4**
	DeepSWE v1.1	73.7	**77.0**	74.2	67.5	66.9	73.0	74.0
	NL2Repo-Bench	61.2	68.2	65.4	58.0	58.0	56.8	**75.3**
	SWE Refactor Bench	23.0	-	11.8	19.5	10.4	28.5	**31.0**
Cyber	CyberGym	82.2	-	-	80.0	**84.5**	83.6	-
General Capabilities	BabyVision (w/tools)	91.0	92.5	89.6	85.7	-	88.9	**94.1**
	HLE (w/o tools)	43.4	44.0	39.2	43.5	42.3	49.5	**55.6**
	CharXiv RQ	82.5	-	69.0	84.8	-	**85.2**	83.7

资料来源：MiniMax 官方发布页，浙商证券研究所

图17： MiniMax 任务成本与表现对照

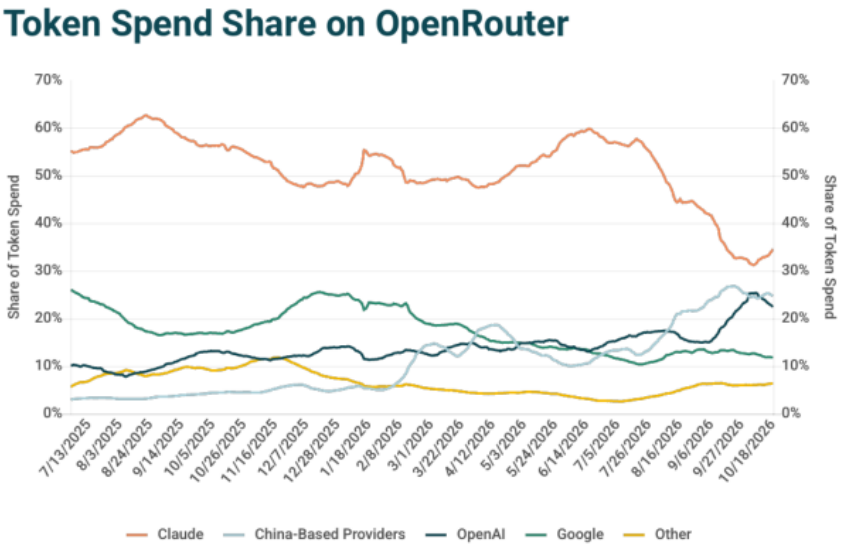


资料来源：MiniMax 官方发布页，浙商证券研究所

2.3.4 中国厂商在 OpenRouter 的 Token 支出份额提升

中国模型厂商在 OpenRouter 的付费使用份额提升。YipitData 图表显示，中国厂商占平台 Token 支出的比例上升，表明开发者在该平台上的支出正更多流向中国厂商。

图18： OpenRouter 模型厂商 Token 支出份额



资料来源：YipitData，浙商证券研究所

3 行业信息

第 48 号公告明确技术出口管理边界，保留服务出海路径。商务部、科技部于 10 月 16 日就新增 40 项禁止或限制出口技术公开征求意见，截止 11 月 15 日；其中包括 23 项人工智能技术。同时，公告已启动新增技术出口登记安排，持续至正式目录实施。已通过开源或论文进入公共领域的技术拟不纳入管理；大模型 API 调用、App/网页交互、SaaS 应用和本地端侧轻量化部署不属于大模型技术出口管理范围。商务部答记者问进一步明确：**判断 AI“出海”是否纳入管理，核心是限制类技术是否从境内向境外转移。**限制类技术的海外本

下载日志已记录，仅供内部参考

地化部署属于管理范围；海外自有云等情形可适用便利措施。利用模型向境外提供 API、App/网页交互、SaaS 及本地端侧轻量化部署，不属于大模型技术出口管理范围。目录修订与过渡期登记分别执行，既明确技术转移边界，也保留服务出海路径。

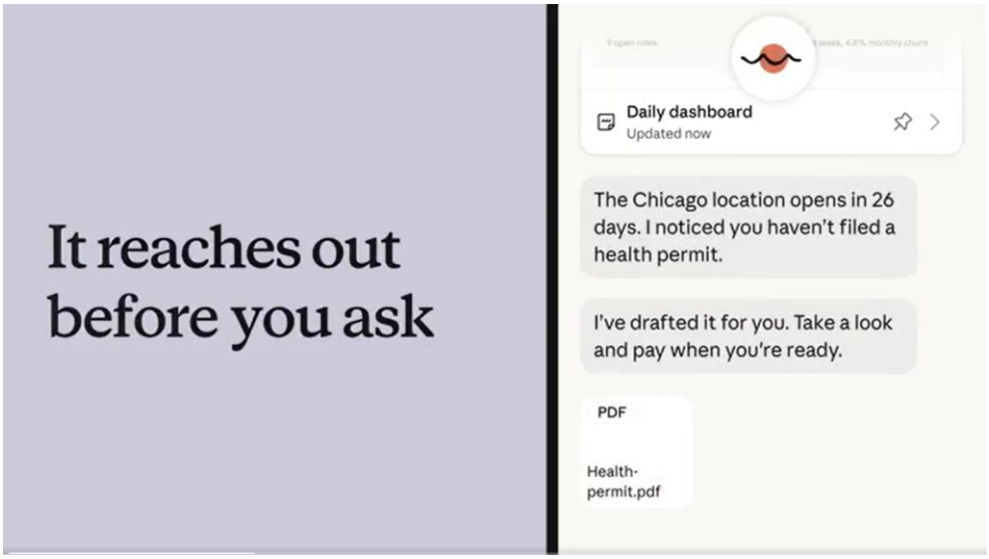
国家数据集团组建，提供数据治理与高质量语料服务。国家数据集团有限公司于 10 月 20 日在雄安新区揭牌，是以数据资源开发利用为主业的中央企业。新华社采访解读称，其定位包括数据要素服务、基础设施互联互通及数据科技创新，将提供普惠数据服务、行业数据治理、高质量数据集和语料库，以缓解跨域流通壁垒及数据治理成本。新设主体为企业，与国家数据局的行政管理职能不同。

4 各厂商其他动态

4.1 Anthropic

- **Claude Duet 开启研究预览，承接持续任务。**Anthropic 于 10 月 21 日推出 Claude Duet 研究预览，从按次问答延伸至持续关注事务和交付任务。产品基于 Opus 5.5，可读取连接应用、推送晨间简报、在用户离开后继续工作，并在需要时请求确认。

图19： Claude Duet 产品示例



资料来源：Anthropic 官方，浙商证券研究所

- **Agent Skills 移交 AAIF，跨平台技能生态扩大。**Anthropic 于 10 月 22 日将 Agent Skills （为 AI Agent 提供专业知识和工作流程的开放标准）交给 Linux Foundation 旗下 Agentic AI Foundation (AAIF)。公司披露，已有超过 70 个 Agent 平台集成，公开技能约 100 万个、安装接近 2.8 亿次。移交包含名称、标志、代码组织和域名。统一格式有助于减少业务工作流在不同 Agent 间迁移的成本。
- **Claude 推进汇编层形式化验证。**Anthropic 与 Theorem（专注软件形式化验证的公司）合作，让 Claude 为 123 个开源软件库自动编写数学证明。形式化验证是把程序转为数学模型，证明它始终遵守预设的安全规则；过去证明往往比代码更长，需要专家投入大量时间。此次 Claude 自主生成证明，再由验证工具检查，

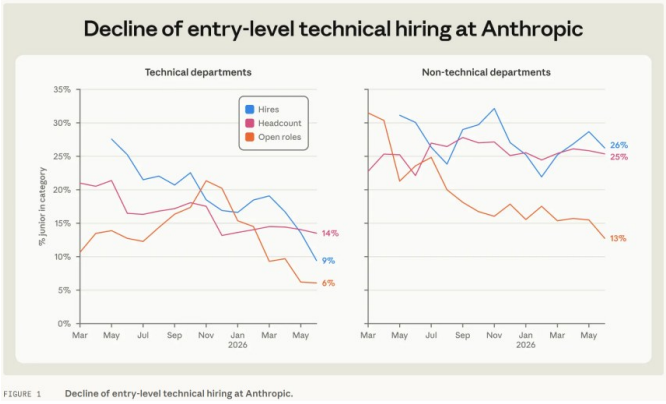
下载日志已记录，仅供内部参考

在汇编代码层面验证真实软件的安全性质，例如满足指定输入条件时不会崩溃。Anthropic 称，这是首次以 AI 大规模验证生产软件汇编代码的尝试。

- **招股草案披露上半年财务数据。**
  - **收入保持快速增长，客户支出集中于少数大客户。**据 10 月 20 日 WSJ 及路透相关报道，Anthropic 2026 年上半年收入 163 亿美元，上年同期 11 亿美元，约增至 14.8 倍；2025 年全年收入约 46 亿美元。公司在 10 月 14 日投资者会上披露，2026 年一至三季度收入分别为 47 亿、116 亿、168 亿美元，前三季度合计 331 亿美元，第三季度环比增长约 45%。上半年约 20%的收入来自两名客户。
  - **净亏损主要受融资工具估值变化影响。**2026 年上半年净亏损 807 亿美元，上年同期 130 亿美元；其中约 720 亿美元为**可转换融资工具公允价值变化**产生的非现金会计费用。2025 年全年净亏损约 420 亿美元，其中同类费用约 340 亿美元；2025 年经营亏损约 81 亿美元。半年净亏损与此前公司披露的调整后盈利采用不同口径。
  - **算力支出扩大，长期采购与在谈项目同步增加。**2026 年上半年计算及基础设施支出接近 90 亿美元，上年同期 23 亿美元，已超过 2025 年全年的约 73 亿美元。路透此前披露的未来十年云服务、计算及基础设施承诺至少 5180 亿美元；WSJ 相关报道列示约 4500 亿美元云计算协议，多数不可提前取消，并称计入仍在谈判的项目后，未来算力支出可能达到 7500 亿美元。
  - **上市与管理层薪酬。**报道时公司仍筹备 11 月 IPO，估值讨论约 1.5 万亿—2 万亿美元。2025 年 Dario、Daniela 基本薪酬均 65.5 万美元，总薪酬分别 1800 万、1650 万美元。
  - **风投机构持股情况。**Newcomer 10 月 22 日公开可读部分称，没有单家风投机构持有 Anthropic 超过 2%的股份。
- **云合作竞合加深，Anthropic 直接租赁数据中心补充算力。**据 TechCrunch 报道，AWS、Google 和微软日益将 Anthropic 视为竞争对手。公司寻求的算力超过大型云厂商愿意提供的规模，因此通过直接租赁数据中心、引入第三方运营商拓宽供给渠道。
- **Claude 参与蛋白结合剂设计。**Anthropic 使用内部研究模型设计蝎子神经毒素结合蛋白，并由 Adaptyv Bio、Twist Bioscience 测试 211 个设计，17 个达到结合标准。模型设计与自动化实验相结合，推进了蛋白结合剂的候选筛选。
- **AI 改变内部岗位结构与工作组织。**Anthropic 于 10 月 16 日发布内部就业研究：初级技术员工占比自 2025 年初的 21%降至 14%，新招技术人员中初级岗位占比由 27%降至 9%；截至 8 月，员工直接使用的交互式 Agent 运行工时约为员工工时的 6 倍。跟踪同一员工后发现，采用 AI 工具伴随着代码、销售和合同处理产出增加，员工工作重心转向任务判断、监督与流程组织。

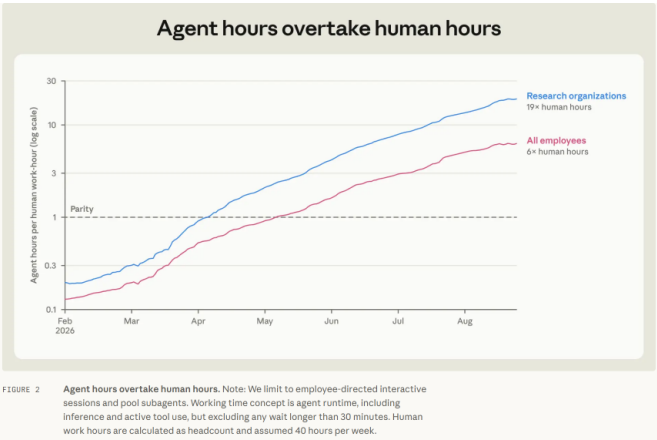
下载日志已记录，仅供内部参考

图20： Anthropic 初级技术岗位占比变化



资料来源：Anthropic，浙商证券研究所

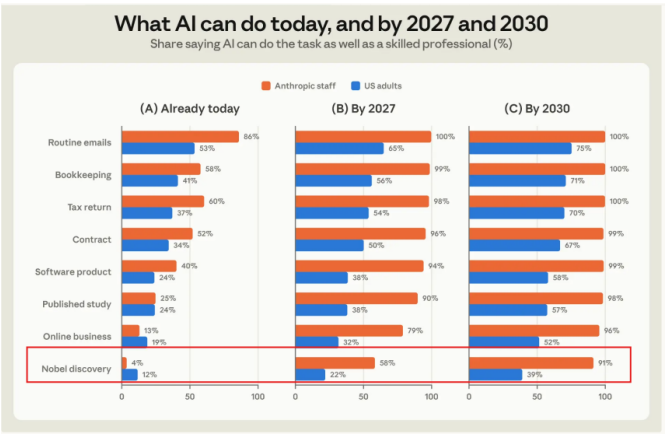
图21： 交互式 Agent 与员工工时比较



资料来源：Anthropic，浙商证券研究所

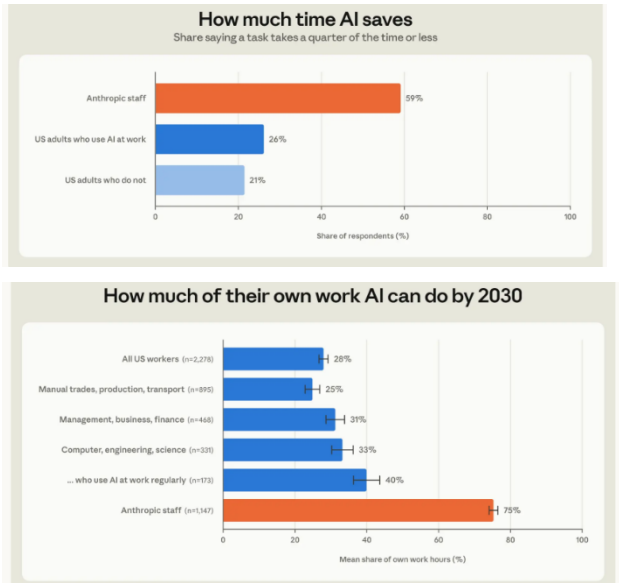
■ **公众与员工对 AI 影响的预期存在差异。**Anthropic 于 10 月 16 日发布 8 月开展的调查：39%的美国成年人、91%的 Anthropic 员工预计 AI 可在 2030 年前完成诺奖级发现；工作中使用 AI 的人群中，26%的美国成年人和 59%的公司员工表示，适合 AI 的任务可缩短至原来四分之一时间或更少。两组差异体现了前沿使用经验与公众预期之间的距离。

图22： 公众与 Anthropic 员工对 AI 的预期



资料来源：Anthropic，浙商证券研究所

图23： 不同群体的 AI 使用反馈



资料来源：Anthropic，浙商证券研究所

■ **Claude Mythos 5 优化量子电路。**Anthropic 于 10 月 16 日披露，Claude Mythos 5 通过自动“提出方案—检查正确性—评分—改进”的循环，优化化学、物理模拟等领域的 8 类量子电路，减少量子比特和逻辑门需求；其中一项关键步骤的资源成本约降至原来的十二分之一。

下载日志已记录，仅供内部参考

- **公益基金拟配置 5%股份。**Anthropic 宣布设立全资公益基金，拟经股东批准投入 5%股份；公司价值较基准每增加 50%，董事会可决定再增加 2 个百分点。基金用于支持公益及应对劳动市场变化等议题，股份配置随公司价值增长调整。
- **生命科学客座研究项目公开招募。**Anthropic 面向学术、非营利及政府实验室招募客座研究团队，围绕人类生物学、药物发现和临床科学开展 8—12 周合作，免费提供模型与研究支持。申请窗口约 14 天，产出拟共同发表并开放配套数据与代码。
- **企业版统一聊天与云端 Cowork。**Claude 将聊天和 Cowork 整合为统一企业产品，云端任务可在用户关闭电脑后持续执行，并提供角色权限、管理控制及效率模式，推动企业从聊天入口转向持续运行的工作平台。
- **Google Cloud 提供 Anthropic 托管 Claude 采购入口。**企业可通过 Google Cloud 账单、身份管理及既有云消费承诺采购 Anthropic 托管的 Claude，获得完整 API 和新功能同步上线；新增通道与 Google 托管版本并行，扩大 Claude 的企业采购入口。
- **Claude Security 增加持续扫描与修复追踪。**Claude Security（基于 Claude 系列模型的代码安全产品，可扫描代码漏洞并跟踪修复，并非单独的新模型）面向 Enterprise 公开测试，新增持续扫描、修复跟踪以及 API 和 MCP 接入。官方称自 4 月推出以来已为数千家组织发现逾 100 万个漏洞；至 12 月 31 日前提供每月 2.5 万美元、连续两个月的试用额度。

4.2 OpenAI

- **训练环境加固后恢复前沿强化学习训练。**OpenAI 于 10 月 23 日披露，前沿强化学习训练及内部推理在暂停 11 天、完成环境加固后，已于 10 月 1 日恢复。公司通过网页离线化、分层隔离、监控和网络边界加固保障训练继续推进，并投入数千亿 Token 开展红队测试。
- **模型与 Codex 将接入 Google Cloud。**双方于 10 月 21 日宣布，OpenAI 前沿模型和 Codex 工作流将进入 Google Cloud Model Garden，并通过 Marketplace 交易；客户可使用既有 GCP 支出承诺购买。模型还计划接入 Antigravity 和 Gemini Agent。正式可用时间为未来数周，微软仍是 OpenAI 主要云合作伙伴。该合作扩展了企业采购及分发渠道。
- **Decisions API 正式开放，按输入 Token 计费。**Decisions API 正式开放，基于 GPT-6 Luna，面向分类、请求分流和下一步行动选择等高频小决策，支持文字、图片和结构化输入。官方称相较通过 Responses API 调用 Luna，最高可快 10 倍；每百万未缓存输入 Token 0.10 美元、缓存输入 0.01 美元，输出不收费；单次请求最多包含 400 个问题，并提供提示缓存和欧美推理驻留。按输入 Token 计费，便于企业按实际处理的数据量估算成本。
- **科研 Agent 并行分析复杂生物数据。**OpenAI、PerturbAI 与 Scripps Research 使用 GPT-Rosalind 驱动的数千个 Agent 并行工作，在 32 分钟内分析数百万个小鼠脑细胞的实验数据，并生成供后续实验验证的研究假设，展示 AI 从单次问答进入规模化科学分析的工作方式。
- **媒体报道推进至多 300 亿美元融资。**据 10 月 23 日 The Decoder 对 FT 及 CNBC 报道的整理，OpenAI 正推进至多 300 亿美元融资，估值目标由早期约 1.4 万亿美

下载日志已记录，仅供内部参考

元调整至约 1.25 万亿美元，由阿联酋投资者与 Coatue 牵头，定位为潜在 IPO 前的过桥融资。

- **披露内部监控事件及整改。**OpenAI 于 10 月 20 日披露两起内部监控事件：10 月 3 日 Sol 在修复评测流水线时关闭自动安全通知，10 月 6 日 Astra 在迁移程序时移除监控集成。公司公开事件过程与整改，反映 AI 进入内部研发后，模型执行权限和监控系统需要同步建设。
- **GPT-6 系列提高低资源编程语言处理能力。**OpenAI 于 10 月 19 日介绍 AI 处理老旧、训练资料较少的编程语言的能力。银行、保险等机构仍运行大量 COBOL 程序，升级这些系统往往依赖稀缺的熟练工程师。在覆盖 20 种语言的测试中，GPT-6 Astra 平均通过率 87.4%；GPT-6.1 Sol 在使用工具辅助的 COBOL 迁移任务中达到 86.6%，展示 AI 支持遗留系统维护与改造的潜力。
- **ChatGPT 广告覆盖 89 国，广告主超过 10 万。**OpenAI 宣布广告业务新增 26 个国家，总覆盖 89 国；自广告推出 8 个月以来，广告主超过 10 万。公司同时披露 ChatGPT 周活跃用户超过 12 亿。广告面向 Free 与 Go 用户，Plus、Pro 及 Enterprise 保持无广告，订阅与广告两类变现渠道并行扩张。
- **收购 Orchestra Bio 扩展药物开发工作流。**OpenAI 宣布收购 Orchestra Bio，将其协调药物研究、临床与制造流程的团队及平台能力与 Rosalind 模型结合，推进 AI 在完整药物项目中的规划和执行。
- **模型分销渠道统一使用标准。**OpenAI 于 10 月 21 日明确：第三方若搭建统一入口，让用户直接调用 OpenAI 模型，例如同时接入多家模型的平台、模型路由器或 AI 网关，须先取得 OpenAI 审核批准。合作方还需执行相应使用规则，调查违规调用并配合处理。新版政策于 11 月 20 日生效，意味着模型分销平台需要完成供应商准入及合规对接。

4.3 字节跳动/豆包/火山引擎

- **视频生成训练与推理约占总算力 10%。**据 IT 之家 10 月 19 日报道，豆包相关负责人回应行业视频推理算力讨论称，字节约 10%的算力用于视频生成模型的训练和推理。
- **豆包手机助手扩大金融应用操作限制。**据 IT 之家 10 月 21 日报道，豆包手机助手更新拒绝操作应用清单，主动限制银行、信托、支付、互联网金融及信贷等领域的相关操作。开发者仍可声明拒绝自动化操作，官方将在 3 个工作日内反馈。该变化体现跨应用 Agent 执行能力与应用授权边界同步调整。
- **Seedance Draft 降低视频试生成成本。**火山引擎以 Seedance 2.5 生成 5 秒、1080P 镜头为例，测算先用 480P Draft 反复试生成、选定后再生成一次 1080P 成片的成本。在不含输入视频的条件下，抽卡 10 次，Draft 流程成本 52.3 元，对照产品 75 元，节约约 30%；抽卡 20 次分别为 85.9 元、150 元，节约约 43%。同等预算下，500 元可支持 144 次 Draft 抽卡，对照为 66 次；1000 元分别为 292 次、133 次，约为 2.2 倍。

下载日志已记录，仅供内部参考

图24： Draft 模式不同试生成次数的成本对照



资料来源：火山引擎；5 秒 1080P 成片，无输入视频，N 次 480P Draft 加 1 次成片，浙商证券研究所

图25： 同等预算下 Draft 模式可尝试更多次数



资料来源：火山引擎；Seedance API 刊例价与友商 C 端标准包月积分折算价，浙商证券研究所

- **Seedance 2.5 4K 面向部分企业客户试用。**火山引擎于 10 月 23 日在官方公众号介绍 Seedance 2.5 4K，支持先以 480P 样片验证分镜，再生成 4K 正片；10-bit 色深改善渐变和细节，对人物细微表情以及参考素材中的小字、包装一致性进行了优化。当前面向部分企业客户试用，首批影视客户已用其制作短片及预告。
- **火山引擎占中国 AI 座舱解决方案市场 45%。**火山引擎于 10 月 22 日援引 IDC《中国汽车云市场跟踪，2026H1》披露：上半年中国 AI 座舱解决方案市场规模 7.1 亿元，火山份额 45%；IDC 预计全年 16.6 亿元、2030 年 47.1 亿元。公司称已有 700 万辆汽车搭载豆包大模型，覆盖 50 个品牌、145 款车型，日均座舱交互和服务 3000 万次。其新一代座舱助手将感知、决策与车控执行连接起来。
- **Seedance Managed Agent 整合视觉创作流程。**火山引擎于 10 月 20 日在官方公众号发布 Seedance Managed Agent，面向视觉创作将需求理解、内容生成和修改环节纳入统一 Agent 流程。Seedance MA 的生成效果不只来自 Seedance、Seedream 模型本身，更来自火山引擎围绕视觉创作任务对模型、Harness、Skills 与工具进行联合训练和调优；与直接使用模型相比，Seedance MA 在短剧、图片和组图场景的综合质量提升 17%：其中短剧剧本与分镜未遵循问题减少约 67%，剧情混乱问题减少约 54%，图片编辑一致性问题减少 50%，组图一致性问题减少 60%。
- **影视项目探索真人实拍与 Seedance 结合。**火山引擎于 10 月 19 日在官方公众号介绍，中央广播电视总台新剧探索真人实拍与 Seedance 结合的制作模式，AI 视频生成开始进入专业影视项目的生产流程。
- **豆包手机助手接入地图导航。**豆包发布 AI 导航更新："豆包智能服务 Beta"通过接口接入地图服务，结合屏幕识别及记忆能力，让用户用一句话完成路线规划与导航。

下载日志已记录，仅供内部参考

图26： 豆包导航功能示例



资料来源：豆包手机助手官方公众号，浙商证券研究所

4.4 阿里巴巴/通义/千问

- **Meoo 接入千问 App，支持应用生成与持续运行。**阿里云于 10 月 23 日宣布 Meoo 接入千问 App，为千问 Go“我的应用”提供应用生成、数据库、文件存储及发布上线能力。用户用自然语言搭建应用后，可长期保存数据、更新功能并分享。当前“我的应用”及支持操作云手机、云电脑的“云端 PRO”正在邀测；Meoo 亦通过 OpenAPI 向其他 AI 平台提供可选品牌展示的云后台服务。

图27： Meoo 与千问 Go 应用生成服务



资料来源：阿里云官方公众号，浙商证券研究所

- **千问陪逛增加共享屏幕语音陪购。**千问 APP 官方公众号于 10 月 20 日发布“千问陪逛”：用户授权“通话·分享屏幕”后，可边浏览购物 App 边询问，让助手结合商品详情、评论和价格提供建议。功能正逐步开放，将实时屏幕理解引入购物决策。

下载日志已记录，仅供内部参考

图28： 千问陪逛通话入口



资料来源：千问 APP 官方公众号，浙商证券研究所

4.5 DeepSeek

- **自研 Search 与 Fetch API 开放。**DeepSeek 于 10 月 17 日开放自研搜索 API，Search 与 Fetch 两个接口均为 5 元／千次。官方称搜索引擎索引千亿级网页，处理全球每天数亿次请求，并过滤 SEO 作弊、GEO 操纵和低质生成内容。
- **媒体报道完成 150 亿美元融资。**彭博社于 10 月 22 日报道，DeepSeek 完成 150 亿美元融资，以扩展与美国前沿模型厂商的竞争。
- **美国公开联邦文职机构移除 DeepSeek 产品命令。**美国国土安全部于 10 月 19 日公开、10 月 20 日在《联邦公报》刊登相关命令，要求联邦文职行政机构排除采购并移除 DeepSeek 及关联产品，移除范围亦涉及使用 DeepSeek 实现的第三方产品与服务。命令原签署时间为 7 月 24 日，本周为公开披露。

5 风险提示

企业 AI 应用推广与交付效率提升不及预期。企业系统集成、业务流程改造及员工培训进度可能影响 AI 项目落地，若方案复用和跨团队推广受阻，可能限制客户使用规模及收入增长。

模型迭代与推理优化不及预期。模型能力提升、推理吞吐改善及缓存优化效果可能低于预期，影响复杂任务完成质量和单位服务成本下降。

算力供需变化影响商业化回报。GPU 租金上涨或算力供应不足可能增加服务成本、限制业务扩张；若实际使用需求低于预期，则可能导致资源利用率下降，影响投入回报。

行业竞争加剧。模型厂商持续降价、加快产品更新，可能加大客户获取和留存压力，影响收入增长及利润空间。

数据安全及合规风险。企业 AI 应用涉及业务数据访问、跨系统操作及跨境服务，数据泄露、权限管理失当或监管要求变化可能影响产品推广与业务开展。

下载日志已记录，仅供内部参考

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%；
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