下载日志已记录，仅供内部参考

中信建投证券 CHINA SECURITIES

证券研究报告·行业深度

10 月 AI 月报：模型迭代与投入不减，Agent 商业化持续兑现

计算机

维持 强于大市

应瑛
yingying@csc.com.cn
SAC 编号:S1440521100010
SFC 编号:BWB917

发布日期： 2026 年 10 月 07 日

市场表现



相关研究报告

26.09.27	【中信建投计算机】:周报 26 年第 35 期：阿里展示一站式 AI 方案，算力+应用双提升
26.09.21	【中信建投计算机】:周报 26 年第 34 期：国内超节点方案持续迭代
26.09.13	【中信建投计算机】:周报 26 年第 33 期：光博会圆满收官，华为全联接大会召开在即
26.09.06	【中信建投计算机】:周报 26 年第 32 期：行业中报收入稳健增长，利润改善显著
26.08.30	【中信建投计算机】:周报 26 年第 31 期：IGV 创年内新高，AI 商业化推动海外软件重估

核心观点

海外安全讨论尚未改变模型高频迭代节奏，RSI 逐步成为 OpenAI、Anthropic 等头部模型企业共识；国内阿里、智谱分别完成约 800 亿港元、50 亿美元融资，规模化投入支撑模型能力追赶，腾讯亦加大海外算力采购。商业化方面，海外企业合作深化，ChatGPT 广告年化收入运行率达到 10 亿美元，Muse 推动消费端变现路径拓宽；国内 Agent 仍有望通过提升单用户任务密度，带动 Token 调用及新一轮 ARR 增长。国内 AI 板块整体趋势上行，建议关注算力服务、国产芯片与超节点，以及具备场景、数据和企业交付能力的 B 端 AI 应用厂商。

标题

海外：模型能力继续提升，企业部署和消费端变现落地

安全讨论尚未改变模型高频发布节奏，OpenAI 与 Anthropic 接连更新 GPT-6 Sol/Luna、Opus 5.5 和 GPT-6.1 Sol，后者能力接近 GPT-6 Astra 的同时成本仅 1/5。研发自动化开始反哺模型迭代，截至 8 月，Claude 在人类监督下主导的研发工作占比升至 26%。商业化方面，Anthropic 与 Salesforce、贝恩深化合作，推动 Agent 进入企业流程；ChatGPT 广告年化收入运行率达到 10 亿美元，Muse 拓展个人任务入口，消费端商业价值开始释放。

国内：融资支撑模型扩展，Agent 任务加速 ARR 上行

阿里完成约 800 亿港元配售，智谱完成约 50 亿美元股债融资，腾讯亦加大海外算力采购。商业化方面，MiniMax 7 月 Token 消耗量达到 1 月的 20 倍，智谱将年末 ARR 目标由 24 亿美元上调至 30 亿美元。办公 Agent 竞争及 Harness 生态仍在演进，下一阶段增长更取决于有效任务量、使用留存和付费转化，模型能力提升有望带动新一轮收入扩张。

行情回顾：硬件与海外软件修复领先

以 7 月 31 日收盘为基期，截至 9 月 30 日，IGV 和费城半导体指数分别上涨 12.6%、11.6%；A 股通信、电子上涨 4.7%、3.2%，计算机、传媒下跌 5.2%、3.2%。估值方面，IGV 综合 PS-TTM 为 8.59 倍，申万计算机、传媒分别为 2.80 倍、2.68 倍，较 2025 年末分别下降 10.1%、21.6%、18.6%。海外软件的 AI 付费与数据调用增长为估值修复提供支撑，国内应用端表现仍较分化。后续定价将更加重视客户采购、持续使用及利润贡献，具备明确 AI 收入和交付能力的企业更有望获得估值支撑。

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

下载日志已记录，仅供内部参考

目录

一、月度观点 1
二、全球 AI 商业化进展 3
2.1 海外企业进展 3
2.2 国内企业进展 23
2.3 AI for Science 44
三、行情回顾 52
3.1 海外软件与 A 股 AI 应用板块表现复盘 52
3.2 海外软件与 A 股 AI 应用板块估值复盘 56
3.3 八月至今 AI 板块表现分析 62
四、投资建议 65
五、风险提示 66

下载日志已记录，仅供内部参考

一、月度观点

海外安全讨论不改模型高频迭代节奏，国内能力追赶与研发投入仍具较强确定性。海外关于放缓前沿研发的呼吁，尚未转化为全行业产品发布降速。9 月 22 日，OpenAI 与 Anthropic 分别发布 GPT-6 Sol、Luna 及 Claude Opus 5.5；9 月 29 日，OpenAI 进一步推出 GPT-6.1 Sol，继续提升编程、计算机操作和专业任务表现。国内小米、DeepSeek、智谱及 MiniMax 等亦围绕多模态、长程任务和推理效率推进升级。研发自动化开始成为能力提升的新驱动力，千问披露通过一个多月的自动化运行完成 33 轮迭代，Artificial Analysis 评分由 40 提升至 45。模型扩展、后训练与 RSI 探索同步推进，进一步验证前期关于能力仍将向上突破的判断。安全要求可能影响部分前沿模型的开放方式和发布时间，但竞争与商业需求仍在推动厂商投入。

蒸馏是行业通用的知识迁移手段，国产模型趋势仍向上。基于蒸馏产生的模型间的知识迁移本身是行业共识，海外厂商同样在利用国产模型构建训练数据，例如英伟达官方文档披露，Nemotron 3 Super 使用 DeepSeek、Kimi 作为教师模型，并通过 Qwen3-Coder-480B 生成软件工程蒸馏数据。访问限制可能提高特定训练数据的获取成本，但难以据此推导国产模型的迭代路径被切断。国内厂商正在同步投入自主预训练、强化学习、架构优化及合成数据，后续仍应以真实任务表现、推理效率及付费调用验证产业趋势。只要模型持续迭代、能力差距收窄并转化为客户使用，就不宜因单一技术路径的限制否定行业上行方向。

轻量模型与结构化决策模型丰富 Agent 分工，同等任务质量下的使用成本仍有下降空间。GPT-6.1 Sol 在部分编程与专业任务中接近 Astra，短上下文标准 API 每百万 Token 输入、输出价格为 2 美元和 10 美元，均为 Astra 的五分之一；Luna 则以 0.10 美元和 0.50 美元覆盖低成本任务，Gemini Flash 系列继续以较低价格承接规模化调用。Jev 和 OpenAI Decisions API 进一步将分类、路由及动作选择封装为结构化判断，使 Agent 能够把高频、答案范围明确的任务交由低成本模型，将复杂规划和异常处理交由高能力模型。未来全球同等能力下的标准 Token 价格仍有下行压力，旗舰能力、高速服务及特定供给时段保持溢价。商业化评价应更多关注有效任务的完成成本、耗时和成功率，模型单价下降与使用深度增加可以同时发生。

国产模型能力提升下，下一轮 ARR 增量仍可能主要来自 Agent 带来的单用户任务密度提升，产品格局尚未定型。MiniMax 披露，7 月平台 Token 消耗量达到 1 月的 20 倍；据上海证券报，智谱 9 月将年末 ARR 目标由 24 亿美元上调至 30 亿美元，上调幅度 25%。调用扩张与收入预期上修，与前期关于 Agent 提高使用深度的判断一致。当同一用户可以将更多编程、调研和办公任务交由模型执行，每项任务又包含多轮规划、工具调用与结果验证，付费需求因而能够快于用户数量增长。与此同时，千问办公、豆包工作、WorkBuddy 等仍在争夺产品入口，Harness 及其插件生态也处于迭代中。插件依赖、运行环境和接口版本不一致，可能引发集成冲突，插件化组合未必是最终产品形态。竞争重点仍是稳定完成任务、提高留存与付费转化。

融资与长期算力采购同步扩张，高端资源获取及交付能力成为国产模型发展的重要支撑。阿里完成约 800 亿港元配售，智谱完成约 50 亿美元股债融资，为模型研发和基础设施建设补充资金。9 月 30 日，据《金融时报》报道、路透社转述，腾讯与 Oracle 达成为期五年的海外租赁安排，通过东南亚数据中心获取约 10 万颗先进 AI 芯片的使用能力，交易金额估计约 70 亿美元。海外采购体现厂商对高端资源性能、可获得性和交付周期的重视，也提示模型扩张过程中仍需关注结构性供给约束。预计境外云资源采购、境内基础设施建设与国产算力适配将并行推进，能够提供稳定交付、集群优化和有效算力的服务商，更有条件承接训练与推理需求。

海外 B 端合作深化、C 端变现路径拓宽，国内商业化有望沿相近方向推进，安全治理的重要性同步上升。Anthropic 与 Salesforce 将模型接入企业数据和销售流程，贝恩则通过全公司部署及客户项目扩大 Claude 的企业

下载日志已记录，仅供内部参考

覆盖，验证了模型厂商借助软件平台和咨询伙伴拓展交付的路径。消费端方面，ChatGPT 广告业务在上线不足 200 天后达到 10 亿美元年化收入运行率，Muse 则将个人 Agent 拓展至日常任务执行，推动 AI 应用在订阅之外探索广告与服务变现。国内模型能力差距缩小后，产品与商业化仍需要完成场景适配和付费习惯培育，收入兑现节奏可能滞后，但企业提效与消费端任务服务的需求方向具有共通性。随着 Agent 进入企业系统、个人账户及终端设备，权限控制、数据保护、操作审计与人工接管将逐步成为采购和使用的基础要求，具备场景、交付及安全治理能力的厂商更有条件形成持续收入。

AI4S 由专业预测向科研 Agent 与实验协同扩展，有望成为 AI 应用的重要新增方向。Claude 科研实验、微软 Quine 及罗氏自主实验室建设，体现模型正在连接文献分析、专业计算、候选设计与实验反馈。晶泰上半年 AI4S 智慧解决方案收入同比增长 136.4%，英矽智能授权合作与临床推进，则为科研平台的产业价值提供了验证。随着计算端候选生成能力提升，高质量数据、高通量验证及自动化实验设施的重要性将进一步提高。未来几个季度，应用预计率先在工具基础较好、评价标准相对明确的研发环节落地，具备专业模型、科研 Agent 与实验交付能力的平台更有条件积累持续竞争优势。

未来一至两个季度，国产模型突破、旗舰模型提价和 Agent 落地有望同步推进。展望下半年，国内厂商仍有更大版本模型储备，国产开源模型在部分代码和 Agent 任务上超过海外头部模型具备可实现性；算力供给偏紧将推动顶级模型价格及毛利率逐步回升；DeepSeek Harness 等开源执行框架则有望降低 Agent 开发门槛，加快模型能力向企业办公和生产流程渗透。

下载日志已记录，仅供内部参考

二、全球 AI 商业化进展

2.1 海外企业进展

海外模型迭代依旧未降速，智能上界提升与降本同步推进，RSI 得到一致关注。GPT-6 Astra、Fable 5.1、Grok 4.7 和 Gemini 4 Argon 继续提升复杂编程、计算机操作及专业工作的任务完成能力，竞争重点由单次回答质量延伸至长程规划、工具调用和结果验证。与此同时，GPT-6.1 Sol、GPT-6 Sol/Luna 将旗舰模型的部分能力下放至更低价格区间，Opus 5.5 也通过降低 Token 价格、减少任务消耗和优化缓存，将典型工作负载成本较 Opus 5 降低约 40%。高能力模型仍可获得价格溢价，但同等任务质量下的使用成本正在下降，模型商业化的评价标准应由每百万 Token 价格，进一步转向完成一项有效任务所需的总成本和时间。

此外，RSI 成为大家的一致目标，OpenAI 表示已实现自动化研究实习生阶段目标，AI 能够在人类指导下完成定义明确、通常需要熟练研究人员数天的任务；首席科学家 Jakub Pachocki 发文将 RSI 列为重要研究方向。Anthropic 则在 RSI 专题文章中展示了 Claude 自主提出假设、设计实验与迭代验证的研究流程；截至 8 月，Claude 在人类监督下主导的研发工作占比由 2 月不足 1%升至 26%。长程任务能力与研发自动化相互促进，有望缩短模型迭代周期，并带动多 Agent 并行、持续推理及实验训练需求，为递归自我改进积累基础。

结构化决策模型丰富 Agent 的模型选择，高频判断与复杂推理分工有助于提高执行效率。TypeSafe 推出的 Jev 面向分类、评分和预设选项判断，发布时输入价格为每百万 Token 0.042 美元、输出免费；OpenAI 的 Decisions API 则将 Luna 用于内容分类、请求路由及下一步动作选择。Agent 可以将答案范围明确的高频判断交由低成本决策模型处理，将开放式规划、复杂生成和异常处理交由高能力模型承担。不同模型按任务分工，有助于减少不必要的长文本生成、输出解析和高成本调用，产品竞争因此进一步延伸至模型选择、任务路由与执行框架的协同优化。

企业 Agent 开始进入既有业务系统，模型厂商与软件平台、咨询伙伴的合作加深。Anthropic 与 Salesforce 打通销售数据、权限及预置技能，并通过贝恩等咨询伙伴扩大企业部署；xAI 将 Grok 接入主流云平台，Grok Bot 企业版则补齐访问控制、审计与独立运行环境。企业采购的对象正在由模型调用能力，扩展至能够连接业务数据、执行具体任务并接受组织管理的完整产品。模型能力提升解决任务能否完成的问题，软件平台和实施伙伴则提供客户入口、业务规则及交付能力，双方合作有助于缩短 AI 由试点进入生产环境的周期，也使企业软件的竞争优势更多取决于数据、流程和执行权限的积累。

消费端 AI 由用户增长向变现延伸，入口的商业价值开始显现。ChatGPT 广告业务取得进展，并通过 Sponsored Agents 及 CRM、电商平台集成，将广告触达与后续交互连接；Meta 推出 Muse，借助独立 App 和 WhatsApp 扩大个人 Agent 覆盖。相比主要依赖问答和内容生成的产品，个人 Agent 能够围绕购物、预订和日常事务连续执行任务，更接近用户的实际消费决策。消费端竞争因此不仅取决于访问量和订阅转化，也取决于能否获得持续任务委托、连接商家服务并完成交易流程，广告、订阅与企业服务有望形成互补。

未来半年海外模型迭代及 AI 商业化发展预计呈现以下四点趋势：

1）模型扩展尚未见顶，研发自动化为能力提升增加新的驱动力。延续上期对预训练、强化学习及推理时计算继续扩展的判断，本月 OpenAI、Anthropic 披露的内部研发进展，进一步体现模型能力提升对研发效率的反馈。OpenAI 已能够在人类指导下完成通常需要研究人员数天的明确研究任务，Anthropic 则将更多实验与工程

下载日志已记录，仅供内部参考

工作交由 Claude 主导。未来模型扩展不仅体现为训练规模增加，也体现为 Agent 承担更长程任务、开展更多并行实验及缩短验证周期。RSI 相关探索将强化模型能力与研发效率之间的相互促进，并增加研究型 Agent 的推理及实验算力需求。

2）头部模型的领先位置将继续交替，决策模型进一步验证 Agent 效率的重要性。本月 GPT-6/6.1 系列、Claude Fable 5.1 与 Opus 5.5、Gemini 及 Grok 的更新，延续了各家围绕复杂编程、专业研究、多模态及实时交互分别突破的态势。单次发布或单项榜单领先仍难以形成长期稳定的竞争格局，后训练、工具体系和推理效率改善均可能改变模型的相对表现。未来企业及开发者将更多根据任务难度、响应速度和交付成本选择模型，多模型调用与动态路由的重要性提高，厂商需要将阶段性的能力优势转化为产品留存和工作流黏性。此外，Jev 与 Decisions API 的出现进一步表明，模型竞争不仅体现为通用能力排名，也体现为对不同工作环节的适配。未来 Agent 有望组合高能力模型、轻量模型与结构化决策模型，并根据任务难度、响应要求及判断置信度动态分配工作。

3）模型定价分野将延续，开源模型的冲击进一步加深。上期关于高能力模型与轻量模型定价分化的判断仍然成立，本月变化则表明，前沿能力溢价并不意味着同一档位模型持续涨价。GPT-6 Sol、Luna 及 Opus 5.5 通过推理效率改善降低调用成本，旗舰能力、高速服务、缓存调用及批量任务之间仍保持价格区分。未来半年，轻量模型、缓存复用和任务路由有望继续压低常规工作负载成本，国产开源模型也将增加全球开发者的替代选择，但高能力模型和低延迟服务仍具备溢价空间。对 Agent 而言，更重要的是固定任务质量下的完成成本；单价下降与任务数量、执行深度增加可以并行，模型收入增长将更多取决于实际使用扩张。

4）B 端仍是商业化兑现较快的方向，C 端开始出现新的增长来源，可靠性与安全治理成为规模化部署的关键。Anthropic 与 Salesforce、贝恩等伙伴深化合作，继续验证上期关于企业工作流、渠道和交付能力推动商业化的判断；ChatGPT 广告业务扩张及 Muse 用户增长，则显示消费端的流量变现与任务型应用开始形成新的进展。预计企业付费仍将率先贡献稳定收入，消费端则围绕广告、订阅及生活服务探索增量。随着 Agent 获得更多数据访问和操作权限，企业采购将更加重视身份权限、过程审计、提示注入防护及人工接管机制。具备模型能力、部署渠道和治理体系的厂商，更有条件将试点项目转化为持续使用。

2.1.1 OpenAI

GPT-6 Astra 正式发布，OpenAI 模型竞争进一步向端到端任务执行延伸。 9 月 3 日，OpenAI 正式发布 GPT-6 Astra，定位于复杂端到端工作任务，在计算机操作、软件工程、网络安全、科学研究及专业工作等方向实现全面升级。相比此前 GPT-5.6 通过 Sol、Terra 和 Luna 覆盖不同能力与成本区间，GPT-6 Astra 进一步强化模型自主调用工具和执行多步骤任务的能力，可直接完成网页表单填写、CRM 信息更新、日程管理、在线研究以及文档和表格处理等工作。官方测试显示，Astra 在 FrontierMath Tier 4、ARC-AGI-3 和 ExploitBench 中分别达到 98%、99.9%和 100%，同时在计算机操作、浏览和专业工作等场景取得进一步提升。

GPT-6 Astra 计算机操作能力进一步提升，复杂任务执行效率明显改善。GPT-6 Astra 在 Computer Use 相关评测中全面超过 GPT-5.6 Sol，其中 Agents' Last Exam、OSWorld 2.0 和 ScreenSpot-Pro 得分分别达到 59.3%、72.6%和 92.7%，较 GPT-5.6 Sol 分别提升 5.7pct、6.9pct 和 15.8pct。除任务完成能力外，Astra 在执行效率方面同样实现改善，OpenAI 在 OSWorld 2.0 离线子集的延迟模拟测试中显示，Astra 完成单项任务平均约需 40 分钟，而 GPT-5.6 Sol 约需 75 分钟，任务耗时减少约 47%。此外，在更新后的 Codex harness 下，Astra 在 Mind2Web 上的任务完成速度约为 GPT-5.6 Sol 的 1.9 倍。模型在 Computer Use 上的能力提升意味着 Agent 可以进一步承担网页操作、信息整理、软件使用等此前需要人工参与的环节，有助于推动 AI 从回答问题向直接完成任务转变。

下载日志已记录，仅供内部参考

Astra 专业任务能力同步提升，Agent 应用进一步向高价值知识工作渗透。除 Computer Use 外，GPT-6 Astra 针对专业工作场景进行了进一步训练，在 AutomationBench、BenchCAD 和 BrowseComp 等评测中均较 GPT-5.6 Sol 实现提升。其中 AutomationBench 得分达到 41.4%，较 GPT-5.6 Sol 的 18.1%提升 23.3pct；BenchCAD 和 BrowseComp 得分分别达到 95.9%和 91.5%，较 GPT-5.6 Sol 分别提升 12.6pct 和 1.1pct。相比通用问答和单轮 Coding，专业工作任务通常具有更长的执行链条和更高的交付要求，Astra 在自动化、工程设计和复杂信息检索等任务中的能力提升，有望进一步扩大 Agent 的可应用范围。OpenAI 同时表示，Astra 能够根据既有模板生成结构化的文档、表格和演示文稿，并结合 Computer Use 直接处理专业软件中的数据和工作流程。随着模型在复杂任务中的稳定性和执行效率持续提升，Agent 有望逐步从通用办公助手向研发、设计、金融等高价值知识工作场景渗透。

图表 1:GPT-6 Astra 计算机操作能力及任务效率显著提升

Computer Use

Computer Use	GPT-6 Astra	GPT-5.6 Sol[2]
Agents' Last Exam	59.3%	53.6%
OSWorld 2.0 (v2026.08.08, offline set, partial score)	72.6%	65.7%
ScreenSpot-Pro (no tools)	92.7%	76.9%

资料来源：OpenAI，中信建投

图表 2: GPT-6 Astra 专业任务能力显著提升

Professional

Professional	GPT-6 Astra	GPT-5.6 Sol
AutomationBench	41.4%	18.1%
BenchCAD	95.9%	83.3%
BrowseComp	91.5%	90.4%
OpenScore String Quartets (1 - OMR-NED)	0.84	0.19
Internal Design Tasks	50.0%	47.4%
Internal Data Science Tasks	40.9%	30.5%
Artificial Analysis Intelligence Index v4.1.1	61.2	60.9

资料来源：OpenAI，中信建投

Astra 定位最高端复杂任务，API 定价明显高于 GPT-5.6 系列其他模型。GPT-6 Astra 在模型定位上进一步向高价值、复杂端到端任务倾斜，OpenAI API 标准模式下输入和输出价格分别为 10 美元和 50 美元/百万 tokens，均为 GPT-5.6 Sol 当期限时优惠价 4/20 美元的 2.5 倍；GPT-5.6 Terra 和 Luna 的标准输入/输出价格分别为 2/12 美元和 0.2/1.2 美元。相比 7 月发布的 GPT-5.6 通过 Sol、Terra 和 Luna 形成从高性能到低成本的模型梯度，9 月 Astra 进一步抬高了旗舰模型的能力和价格上限。与此同时，Astra 支持约 105 万 tokens 上下文窗口和 12.8 万 tokens 最大输出，并支持计算机操作、web search、file search、code interpreter 等工具，更适合长周期、多步骤的复杂任务。旗舰模型较高的 API 价格一方面反映其在复杂推理、Agent 执行及专业工作场景中的能力溢价，另一方面也意味着随着模型能力持续提升，AI 模型厂商有望通过高端模型收取能力溢价+轻量模型覆盖规模化调用的方式进一步实现商业化分层。

下载日志已记录，仅供内部参考

图表 3: GPT-6 Astra API 定价显著高于 GPT-5.6 系列其他模型

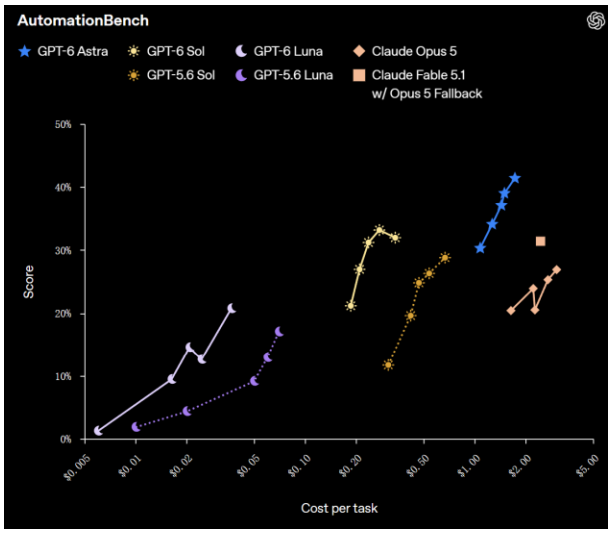
Prices per 1M tokens.

	Short context				Long context			
Model	Input	Cached input	Cache writes	Output	Input	Cached input	Cache writes	Output
gpt-6-astra	$10.00	$1.00	$12.50	$50.00	$20.00	$2.00	$25.00	$75.00
gpt-5.6-sol	$4.00	$0.40	$5.00	$20.00	$8.00	$0.80	$10.00	$30.00
gpt-5.6-terra	$2.00	$0.20	$2.50	$12.00	$4.00	$0.40	$5.00	$18.00
gpt-5.6-luna	$0.20	$0.02	$0.25	$1.20	$0.40	$0.04	$0.50	$1.80

资料来源：OpenAI，中信建投

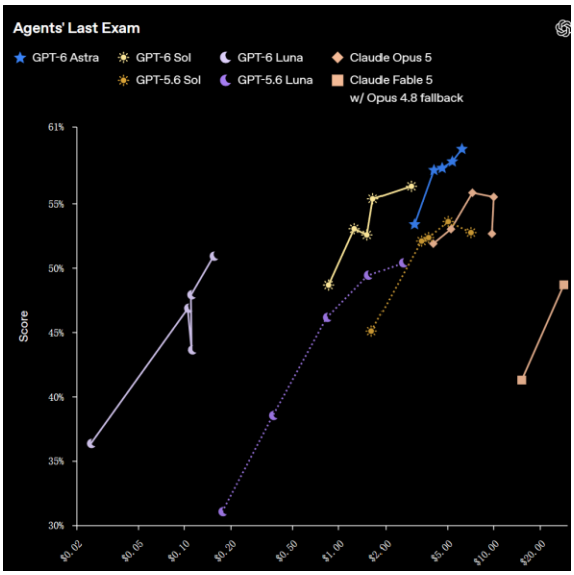
GPT-6 产品线向高性价比模型扩展，降低 Agent 规模化调用成本。9 月 22 日，OpenAI 发布 GPT-6 Sol 和 Luna，将 Astra 的训练方法向更低成本模型延伸，重点提升编程、计算机操作及专业工作能力。Sol 每百万 Token 输入、输出价格分别为 2 美元、10 美元，较 GPT-5.6 Sol 促销价格均下降 50%；Luna 分别为 0.10 美元、0.50 美元。同日，公司升级 GPT-6 提示词缓存，提高重复上下文的复用效率，缓存输入享受 90%的价格折扣，进一步降低长对话与多步骤 Agent 的调用成本。

图表 4:GPT-6 Sol/Luna Automation Benchmark



资料来源：OpenAI，中信建投

图表 5: GPT-6 Sol/Luna Agent'Last Exam Benchmark

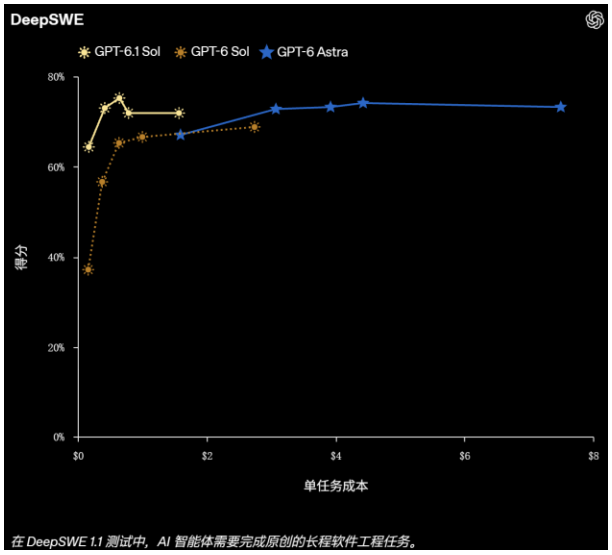


资料来源：OpenAI，中信建投

GPT-6.1 Sol 以较低价格承接旗舰能力，进一步改善复杂任务的调用经济性。9 月 29 日，OpenAI 发布 GPT-6.1 Sol，在编程、电脑操作及专业工作中接近 Astra 的能力水平。标准 API 在输入不超过 27.2 万 Token 时，每百万 Token 输入、输出分别为 2 美元和 10 美元，均为 Astra 标准价格的五分之一；缓存输入降至 0.10 美元，较 GPT-6 Sol 再降 50%。模型支持 105 万 Token 上下文，优先进入 ChatGPT Work、Codex 及 API。官方 DeepSWE v1.1 测试中，其最佳成绩较 GPT-6 Sol 提高 6.4 个百分点。高能力模型与低成本模型之间的性能差距缩小，使企业能够在固定预算下扩大复杂任务调用，并按任务难度配置模型。

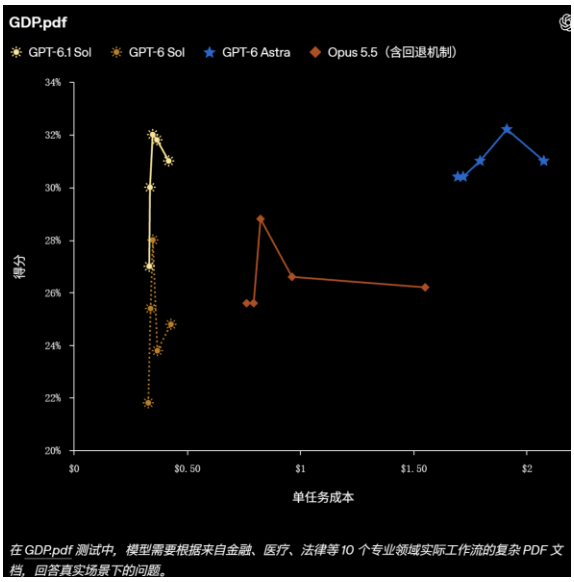
下载日志已记录，仅供内部参考

图表 6: GPT-6.1 Sol 在 DeepSWEv1.1 的表现



资料来源：OpenAI，中信建投

图表 7: GPT-6.1 Sol 在 GDP.pdf 的表现



资料来源：OpenAI，中信建投

ChatGPT Images 2.5 发布，图像生成与编辑能力进一步提升。9 月 8 日，OpenAI 发布 ChatGPT Images 2.5，新模型在图像生成质量、编辑精度及多轮修改一致性方面进一步提升，相比 Images 2.0 生成延迟最高降低 50%。同时，OpenAI 新增 Sketch 功能，用户可以直接在 ChatGPT 中绘制草图并生成完整图片，并通过 Templates 降低海报、商品图等常见内容的创作门槛。OpenAI 同步推出 GPT-Image-2.5 Flare 和 Sunburst 两款模型，分别侧重高速度生成和高精度创意工作流。

图表 8: Images 2.5 手绘成图效果

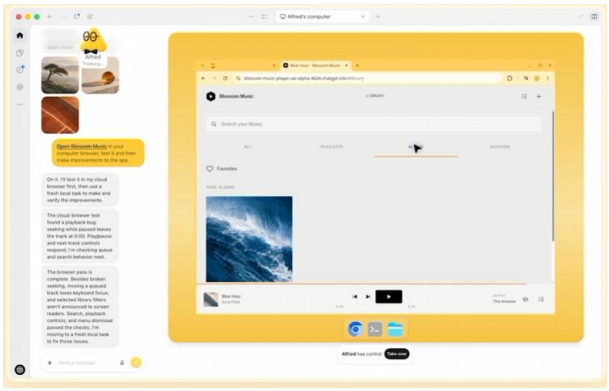


资料来源：APPSO，中信建投

Dots 与 ChatGPT Space 拓展持续任务和团队协同，产品入口由单次对话延伸至长期项目。9 月 29 日，OpenAI 推出常驻 Agent Dots 和团队工作空间 ChatGPT Space。Dots 能够围绕用户目标持续处理任务，在需要授权或关键判断时请求确认；Space 将团队成员、知识资料、页面和 Agent 连接在同一项目环境中。Dots 首先面向符合条件市场的 Pro 和 Business Premium 用户，企业及教育等用户可由管理员开启测试。OpenAI 同时披露周活跃用户达到 12 亿。个人助手、团队工作空间与开发者插件共同扩大既有用户的任务覆盖，为订阅升级、企业采购和第三方应用分发增加产品基础。

下载日志已记录，仅供内部参考

图表 9: OpenAI 发布 24x7 工作助手 dots



资料来源：OpenAI，BubbleBrain，中信建投

图表 10: OpenAI 发布团队办公空间 Space



资料来源：OpenAI，BubbleBrain，中信建投

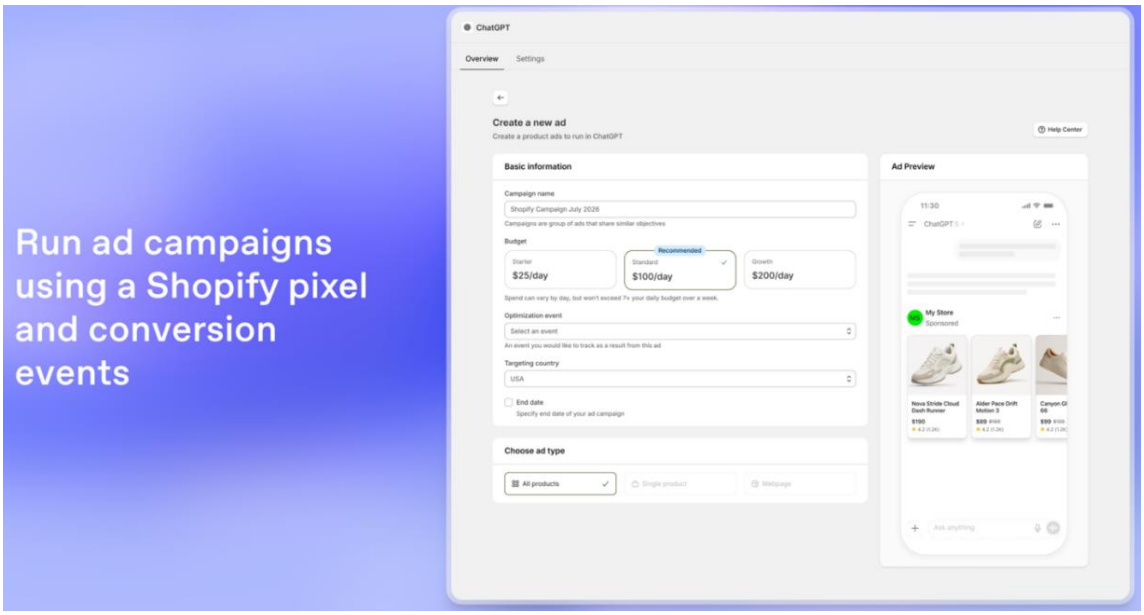
Decisions API 将轻量模型用于结构化判断，Agent 开发工具向执行基础设施延伸。在 DevDay 上，OpenAI 还推出 Decisions API 有限预览，基于 Luna 对开发者预先定义的问题和有限答案集合进行判断，可用于内容分类、请求分流与 Agent 下一步动作选择。同时，Agents API 加入托管浏览器电脑操作，并整合多 Agent、工具搜索及上下文压缩能力；Codex 云端任务支持共享开发环境和权限配置。结构化决策接口与托管执行环境结合，有助于将高频判断和复杂规划分开处理，降低开发者自行建设运行基础设施的负担，扩展模型服务的企业使用深度。

ChatGPT Ads 年化收入运行率突破 10 亿美元，消费端流量开始贡献新增商业化收入。除模型和 Agent 产品持续迭代外，OpenAI 在消费端商业化方面也取得进一步进展。8 月 31 日，OpenAI 披露 ChatGPT Ads 推出不足 200 天，广告业务年化收入运行率已达到 10 亿美元，目前覆盖数万家广告主及 40 多个国家，同时 ChatGPT 周活跃用户超过 10 亿。9 月 16 日，OpenAI 进一步宣布测试 Sponsored Agents，用户点击广告后可与企业赞助的 Agent 交互；同时将广告创建接入 ChatGPT Work，升级 Ads Manager 创意工具，并与首个 CRM 伙伴 HubSpot 及首个电商伙伴 Shopify 集成。

随着 ChatGPT 用户规模持续扩大，广告业务有望成为订阅、API 和企业服务之外的新增收入来源。OpenAI 加快广告商业化，一方面反映 ChatGPT 已经形成较大的消费端流量入口，另一方面也说明 AI 产品正在从单纯依靠订阅付费向用户流量+广告变现+企业服务的多元商业模式延伸。相比传统互联网平台，AI 产品还可以结合用户交互和任务场景进一步提升广告匹配效率，消费端流量的商业价值仍有进一步释放空间。

下载日志已记录，仅供内部参考

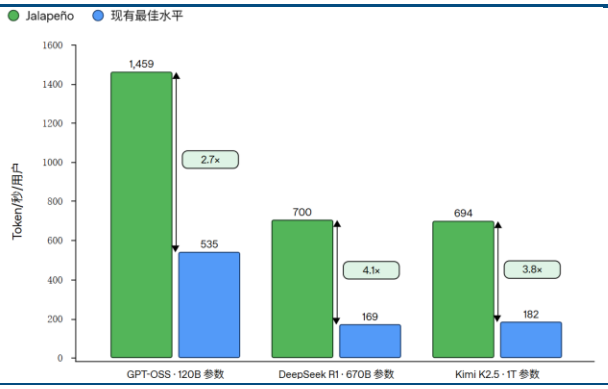
图表 11：ChatGPT Ads 接入 Shopify，支持广告投放与效果追踪



资料来源：OpenAI，中信建投

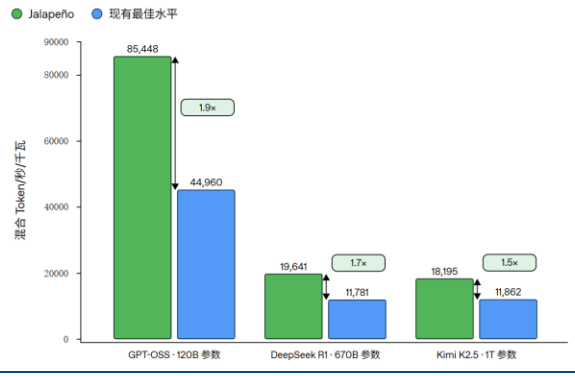
OpenAI 自研 Jalapeño 推理芯片取得阶段性进展，模型与芯片协同有望改善长期推理成本。8 月 25 日，OpenAI 公布自研推理芯片 Jalapeño 首批测试结果，在 GPT-OSS 120B、DeepSeek R1 和 Kimi K2.5 等模型测试中，Jalapeño 峰值单位功耗吞吐量达到对比系统的 1.5 至 1.9 倍，同时具备更低的端到端延迟。OpenAI 表示，Jalapeño 目前仍处于生产验证和软件完善阶段，计划于 2026 年底前开始部署。相比单纯扩大外部算力采购，自研芯片能够使 OpenAI 进一步参与模型、芯片、内存、网络及推理系统的协同优化，通过针对自身模型和 Agent 工作负载进行软硬件适配，提高单位功耗下的有效算力产出。随着模型推理需求持续增长，推理成本和算力利用率的重要性不断提升，Jalapeño 有望成为 OpenAI 优化 AI 服务成本、提升算力供给效率的重要基础设施。

图表 12：Jalapeño 为每个用户提供更高的 Token 速率



资料来源：OpenAI，中信建投

图表 13：Jalapeño 提供更高的每千瓦吞吐量



资料来源：OpenAI，中信建投

2.1.2 Anthropic

Fable 5.1 进一步提升复杂任务能力，Anthropic 持续强化企业级 Agent 布局。 9 月 1 日，Anthropic 发布 Fable 5.1 和 Mythos 5.1，其中 Fable 5.1 面向更广泛用户开放，Mythos 5.1 主要面向经过审核的机构客户。相比上一代模型，Fable 5.1 重点提升复杂编程和知识工作场景表现，在 Terminal-Bench 4.0 和 Terminal-Bench-Science

下载日志已记录，仅供内部参考

0.1 中的得分分别由 Fable 5 的 42.0%和 24.7%提升至 55.8%和 52.6%，长周期、多步骤任务执行能力进一步增强。与此同时，Anthropic 持续完善企业 Agent 产品和部署体系，通过 Claudeforce、Computer Use、Skills API 及 Files API 等产品加强模型与企业业务流程的连接。Anthropic 当前的竞争重点已由单纯提升模型能力逐步向复杂任务执行和企业工作流渗透延伸，模型能力、工具生态和企业部署能力有望共同推动 Claude 商业化进一步落地。

图表 14:Fable 5.1 复杂任务能力全面提升，Agent 应用场景进一步拓展

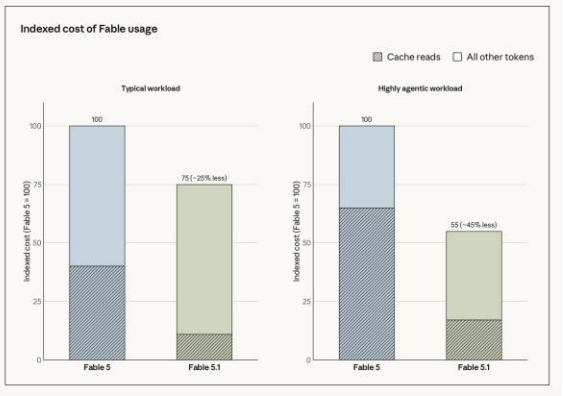
	Fable 5.1	Fable 5	Opus 5	GPT-5.6 Sol
Agentic scientific research Terminal-Bench-Science 0.1 [1]	**52.6%**	24.7%	29.0%	22.4%
Agentic coding Terminal-Bench 4.0	**55.8%** 60.9% (Mythos 5.1)	42.0%	52.3%	37.3%
Knowledge work GDPval-AA v2	**1853**	1723	1824	1711
Computer use OSWorld 2.0 [2]	**77.9%** partial	72.9% partial	75.4% partial	— partial
Computer use OSWorld 2.0	**41.7%** strict	36.1% strict	39.6% strict	— strict
Multidisciplinary reasoning Humanity's Last Exam	**60.9%** no tools	57.8% no tools	56.6% no tools	— no tools
	65.0% with tools	63.8% with tools	63.6% with tools	— with tools
Business workflows AutomationBench	**31.4%**	17.1%	26.9%	19.6%
Agentic coding CursorBench 3.2.0	**73.4%**	70.5%	70.0%	67.2%

资料来源：Anthropic，中信建投

Fable 5.1 能力提升与推理成本下降并行，Agent 任务经济性进一步改善。Fable 5.1 在 Terminal-Bench Science、Terminal-Bench 4.0、AutomationBench 等复杂任务中均较 Fable 5 实现明显提升，同时 Anthropic 将缓存读取价格下调 75%至 0.25 美元/百万 tokens。官方测算显示，在典型工作负载下，Fable 5.1 的任务成本较 Fable 5 下降约 25%；对于高度 Agentic 的工作负载，成本最高可下降约 45%。随着 Agent 任务执行链条不断拉长，缓存机制通过复用已计算的输入状态减少重复计算，并以更低的缓存读取价格计费，成本下降有望进一步提升长周期 Agent 任务的实际使用经济性。

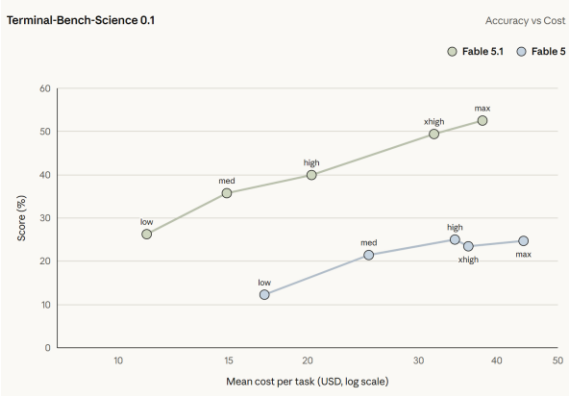
下载日志已记录，仅供内部参考

图表 15: Fable 5.1 缓存读取价格下降 75%，Agent 任务成本显著降低



资料来源：Anthropic，中信建投

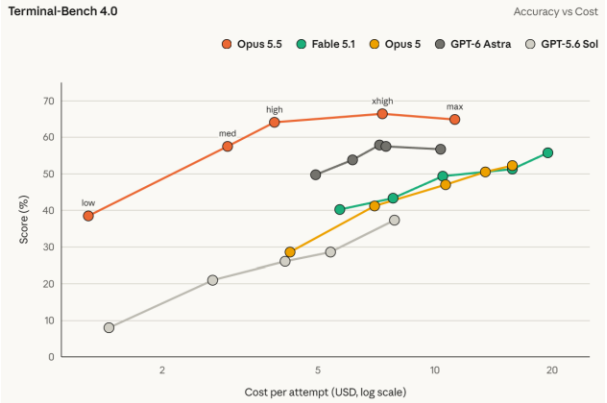
图表 16: Fable 5.1 在科学研究任务中实现更高得分与更低成本



资料来源：Anthropic，中信建投

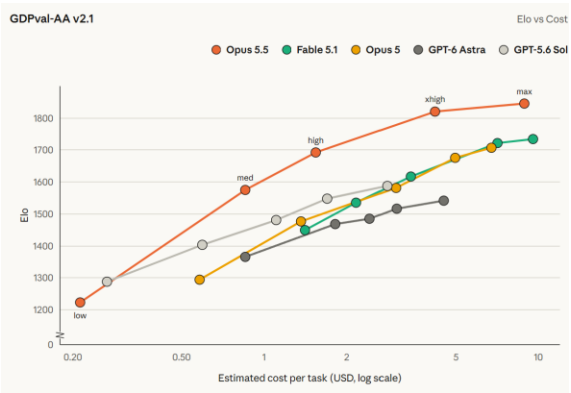
Opus 5.5 强化复杂任务能力，定价下调与模型提效共同降低任务成本。9 月 22 日，Anthropic 发布 Claude Opus 5.5，同步覆盖 Claude 平台及 AWS、Google Cloud、Microsoft Azure。新模型每百万 Token 输入、输出价格分别为 4 美元、20 美元，较 Opus 5 均下降 20%；缓存读取价格降至 0.20 美元，降幅达 60%。结合单任务 Token 消耗减少，公司测算默认设置下典型工作负载成本较 Opus 5 下降约 40%。能力方面，Opus 5.5 在多数工作任务中接近 Fable 5.1 水平，在官方 CursorBench 4.0 测试中以中等推理强度取得 52.5%的成绩，高于 Opus 5 最高推理强度下的 46.6%，推动复杂编程与企业 Agent 应用的性价比提升。

图表 17: Opus 5.5 Terminal-Bench 4.0 benchmark



资料来源：Anthropic，中信建投

图表 18: Opus 5.5 GDPval-AA v2.1 benchmark



资料来源：Anthropic，中信建投

Sonnet 5.5 维持 Token 单价并提高任务效率，补齐 Claude 中端模型的规模化调用能力。9 月 28 日，Anthropic 发布 Sonnet 5.5，标准 API 每百万 Token 输入、输出及缓存读取价格分别为 2 美元、10 美元和 0.20 美元，与 Sonnet 5 一致。公司测试显示，新模型输出速度提高 30%以上，多数工作任务的单次成本最高下降 30%；CursorBench 4.0 成绩由 34.1%升至 55.5%，OSWorld 2.1 部分得分由 57.0%升至 80.1%。Sonnet 承担范围明确的编程、文件制作及常规执行任务，Opus 侧重需要持续判断的复杂工作，形成按任务难度分配预算的产品组合。

下载日志已记录，仅供内部参考

图表 19: Sonnet 5.5benchmark 得分

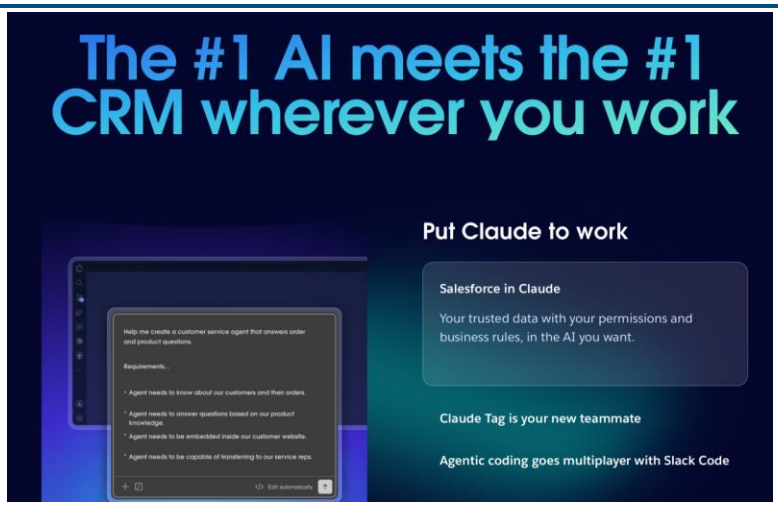
	Sonnet 5.5	Sonnet 5	Opus 5.5	GPT-6 Sol
Agentic coding Terminal-Bench 4.0	**70.6%**	10.3%	66.4%[1]	—
Agentic coding FrontierCode 1.1 (Main)	46.2% Max[2] 52.1% Xhigh	42.4%	**54.4%**	49.3%
Agentic coding CursorBench 4.0	55.5%	34.1%	**57.8%**	—
Knowledge work GDPval-AA v2.1[3]	1844	1449	**1846**	1487[4]
Knowledge work AA-Briefcase v1.1[3]	1811	1359	**1822**	1483[4]
Multidisciplinary reasoning Humanity's Last Exam	64.5% with tools	54.9% with tools	**67.7%** with tools	—
Computer use OSWorld 2.1	80.1% partial	57.0% partial	**81.8%** partial	—
Visual chart recognition Chartography	61.6% no tools	15.6% no tools	**64.4%** no tools	53.6%[4] no tools

资料来源：Anthropic，机器之心，中信建投

Claudeforce 进一步打通企业业务流程，模型能力开始向企业软件执行层延伸。 8 月 26 日，Anthropic 与 Salesforce 扩大合作，推出 Salesforce in Claude，公告时向部分试点客户提供，计划 9 月开放公测，包含 37 项预置销售技能，可支持会议准备、商机检查和销售管线更新等任务，并沿用 Salesforce 现有的权限和业务规则。同时，Claude 成为 Slackbot 等产品的默认模型，双方进一步扩大内部相互采用。相比单纯通过 API 调用模型，Claudeforce 进一步将模型能力嵌入企业已有的数据、权限和业务流程，使 Claude 能够直接参与销售等核心业务环节。模型厂商与企业软件平台的结合，有助于降低 Agent 进入企业核心工作流的门槛，未来企业 AI 的竞争不仅取决于模型能力，也取决于能否连接已有数据和业务系统并完成实际任务。

下载日志已记录，仅供内部参考

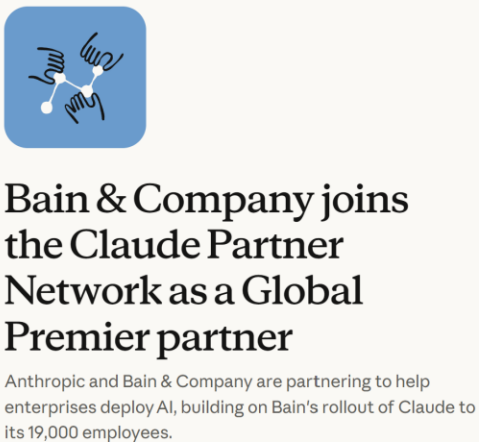
图表 20: Claudeforce 打通 Salesforce 企业数据与业务流程



资料来源：Anthropic，中信建投

Anthropic 拓展咨询伙伴合作，完善企业部署与模型验证能力。8 月 25 日，贝恩成为 Claude Partner Network 的 Global Premier 合作伙伴，并已向约 1.9 万名员工部署 Claude；试点阶段数周内活跃使用人数超过 7000 人，超过三分之二的试点参与者采用 Claude for Excel，双方将联合推进企业 AI 部署。9 月 18 日，Anthropic 与埃森哲宣布开展嵌入式模型评估合作，由埃森哲旗下 Faculty 牵头，覆盖模型评测、红队测试、对齐评估及安全机制验证，双方分别预计未来五年投入至少 10 亿美元用于相关能力建设。两项合作分别强化企业交付渠道与独立评估能力，有助于降低客户部署和验证门槛，推动 Claude 在企业场景中的应用。

图表 21: Anthropic 与贝恩达成合作



资料来源：Anthropic，中信建投

图表 22: Anthropic 与埃森哲达成合作



资料来源：Anthropic，中信建投

巴克莱扩大 Claude 生产部署，银行客户开始形成可跟踪的高频工作负载。10 月 1 日，Anthropic 披露与巴克莱深化合作，覆盖软件开发、遗留系统现代化及业务运营。巴克莱的知识助手已被超过 1.6 万名员工采用，累计处理超过 100 万次搜索；全球市场业务使用 Claude 对每日约 12 万封邮件进行分类、信息补全和处理路径分配。银行预计 2026 年底 Claude Code 覆盖 50%的开发人员，并于 2027 年扩展至多数软件工程师。企业知识检索与日常运营的持续使用，为模型调用提供稳定需求，也使权限管理、业务规则和人工监督成为部署体系的重要组成部分。

下载日志已记录，仅供内部参考

Frontier Academy 投入 1 亿美元建设交付人才体系，Anthropic 强化企业 AI 实施渠道。10 月 2 日，Anthropic 推出 Claude Frontier Academy，承诺投入 1 亿美元，目标于 2027 年底前培训 1 万名前沿部署工程师。首批参与机构包括埃森哲、贝恩、凯捷、德勤、麦肯锡、摩根士丹利及诺和诺德。项目包含现场训练和 12 周企业实践，由工程师带回真实业务场景实施并接受考核。模型能力提高后，企业落地仍需要系统集成、安全审查与流程改造，培训计划有助于扩大合作伙伴的交付能力，将模型厂商的销售触达转化为客户生产环境中的实际应用。

2.1.3 xAI

Grok 4.7 强化长程任务能力，维持定价推动编程与知识工作性价比提升。9 月 21 日，xAI 发布 Grok 4.7，采用更大规模的基础模型，并针对长时间复杂任务加强强化学习训练，改善结果自检、长上下文管理及 Grok Bot 适配能力。新模型标准 API 输入、输出价格分别为每百万 Token 2 美元、6 美元起，与 Grok 4.6 保持一致；另提供输出速度提升一倍的快速版本，价格相应翻倍。在官方 CursorBench 4.0 评测中，Grok 4.7 以 xHigh 推理强度取得 46.3%的成绩，高于 Grok 4.6 在 High 强度下的 40.4%。模型已上线 Cursor、Grok Build 及 API，在维持基础调用价格的同时提升复杂任务表现，拓展编程与专业知识工作的应用空间。

Grok 4.6 接入三大云平台，xAI 加快进入企业既有采购和治理体系。8 月 19 日、21 日及 26 日，Grok 4.6 先后接入 Amazon Bedrock、Google Gemini Enterprise Agent Platform 及 Microsoft Foundry，扩大企业客户的模型调用渠道。模型支持 50 万 Token 上下文及多档推理强度，适用于编程、研究助手和多步骤 Agent 任务。AWS 及 Google 平台公布的输入、缓存输入、输出价格分别为每百万 Token 2 美元、0.50 美元及 6 美元。通过主流云平台分发，企业可以沿用已有采购、权限和治理体系开展模型评估与部署，降低采用 Grok 的实施门槛。

图表 23: Grok 4.7 benchmark

	Grok 4.7 xHigh	Grok 4.6 High	GPT-5.6 Sol Max	Fable 5.1 Max
Input token price $ per million	**$2**	$2	$4	$10
Output token price $ per million	**$6**	$6	$20	$50
Software engineering CursorBench 4.0	46.3%	40.4%	41.7%	**51.8%**
Software engineering DeepSWE v1.1	71.0%*	65.2%	**72.7%**	70.0%
Electrical engineering EEBench	**64.0%**	53.0%	39.4%	56.4%
Multi-hour office work AA Briefcase v1.1	1,657	1,546	1,487	**1,678**
Multi-hour terminal work Terminal-Bench 4.0	37.6%	20.3%	37.3%	**57.9%**
Legal work Harvey Legal Agent Benchmark	**19.6%**	15.8%	2.5%	6.7%
Clinical reasoning HealthBench Professional	56.7%	48.5%	60.5%	**62.1%**

* high effort

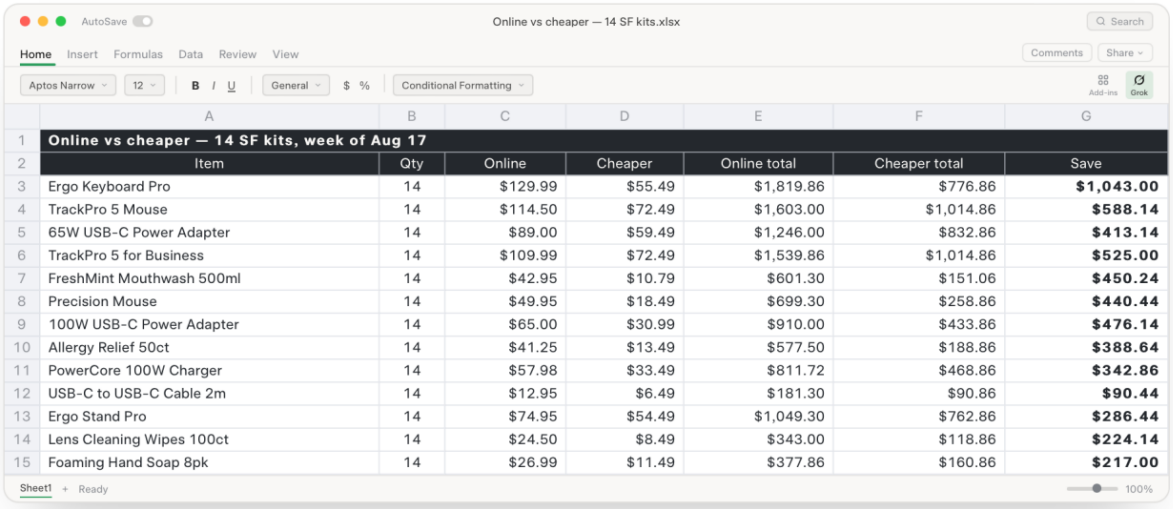
资料来源：xAI，中信建投

下载日志已记录，仅供内部参考

Grok Bot 企业版上线，xAI 加快从模型 API 向企业 Agent 商业化延伸。9 月 3 日，xAI 推出 Grok Bot 企业版，增加访问、网络及审计控制，支持企业对长期运行的 AI Bot 进行统一管理。每个 Bot 拥有独立的云端计算机，可以调用日常软件和网站，自主完成连续工作任务。xAI 披露，已有数千家机构采用 Grok Bot，包括 Supermicro、ServiceTitan 等，使用场景覆盖工程研发、销售、招聘、营销及财务。发布时，Grok 和 Cursor 企业客户可获得两周免费试用，并邀请整个组织参与，推动 Agent 由个人工具向组织级应用扩展。

9 月 4 日，xAI 进一步展示 Haggle Bot 采购案例，Bot 能够结合企业采购数据、历史订单及市场价格，对办公用品进行跨平台比价，并将更低价格的采购方案整理至表格中。以 14 套办公用品采购为例，Ergo Keyboard Pro 单项合计节省超过 1000 美元。xAI 还披露，在一次实际运行中，相关流程将一笔 14629 美元的办公用品订单降至 6143 美元，成本下降 58%。随着 Agent 逐步具备跨系统获取信息、比较方案并执行任务的能力，AI 应用由信息问答和内容生成向采购、销售、财务等业务流程延伸。

图表 24:Grok Bot 推动企业 Agent 向实际业务流程落地



	A	B	C	D	E	F	G
1	Online vs cheaper — 14 SF kits, week of Aug 17						
2	Item	Qty	Online	Cheaper	Online total	Cheaper total	Save
3	Ergo Keyboard Pro	14	$129.99	$55.49	$1,819.86	$776.86	$1,043.00
4	TrackPro 5 Mouse	14	$114.50	$72.49	$1,603.00	$1,014.86	$588.14
5	65W USB-C Power Adapter	14	$89.00	$59.49	$1,246.00	$832.86	$413.14
6	TrackPro 5 for Business	14	$109.99	$72.49	$1,539.86	$1,014.86	$525.00
7	FreshMint Mouthwash 500ml	14	$42.95	$10.79	$601.30	$151.06	$450.24
8	Precision Mouse	14	$49.95	$18.49	$699.30	$258.86	$440.44
9	100W USB-C Power Adapter	14	$65.00	$30.99	$910.00	$433.86	$476.14
10	Allergy Relief 50ct	14	$41.25	$13.49	$577.50	$188.86	$388.64
11	PowerCore 100W Charger	14	$57.98	$33.49	$811.72	$468.86	$342.86
12	USB-C to USB-C Cable 2m	14	$12.95	$6.49	$181.30	$90.86	$90.44
13	Ergo Stand Pro	14	$74.95	$54.49	$1,049.30	$762.86	$286.44
14	Lens Cleaning Wipes 100ct	14	$24.50	$8.49	$343.00	$118.86	$224.14
15	Foaming Hand Soap 8pk	14	$26.99	$11.49	$377.86	$160.86	$217.00

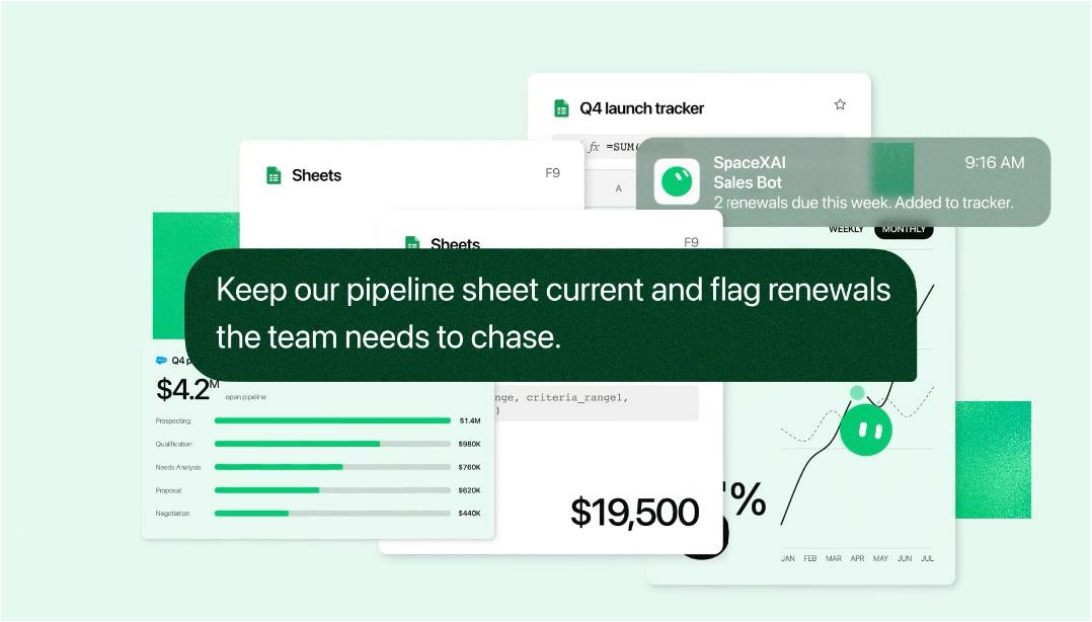
资料来源：xAI，中信建投

Grok Bot 接入 X 生态，拓展企业 Agent 的外部信息来源。8 月 29 日，xAI 宣布 Grok Bot 与 X 进一步打通，用户可通过 X connector 搜索帖子、读取 Timeline 和查看 Mentions；X plugin 支持搜索帖子、查看趋势及管理 Bookmarks 等操作。接入 X 的实时公开信息后，Grok Bot 可将企业内部数据与外部动态结合，为市场研究、舆情跟踪及内容运营提供支持，扩大企业 Agent 的任务覆盖范围。

Team Bots 将共享知识与跨系统操作连接，Grok Bot 进一步向团队执行角色扩展。9 月 28 日，xAI 推出 Team Bots 公测，面向 Teams 和 Enterprise 套餐开放。企业可围绕销售、工程、营销及数据分析角色配置共享文件、技能、插件和 API 凭证，并在 Slack 中协同使用；每名用户的独立对话和记忆保持隔离，团队共同沉淀可复用知识。产品可连接 Salesforce、Notion 和 GitHub 等应用，完成客户简报、工单处理及数据查询。团队成员围绕同一个 Bot 积累上下文，有助于减少重复交接和资料整理，使 Agent 从一次性工具逐步进入持续运行的组织流程。

下载日志已记录，仅供内部参考

图表 25: Team Bots 面向团队共享场景开放公测

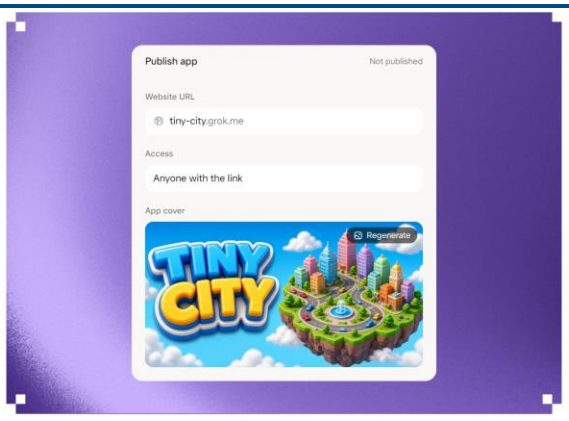


资料来源：xAI，中信建投

Grok Build 扩大开放范围并增加跨绘画记忆，提高长期项目的上下文复用效率。8 月 19 日，xAI 将 Grok Build 向所有套餐开放，同步覆盖 Web、iOS 和 Android 端。用户可以通过自然语言描述创建 App、游戏、网站及 Dashboard，并发布和分享成果。此次开放从此前仅面向 SuperGrok Heavy 用户的 Early Beta 扩展至全部套餐，同时支持与 X 连接及在所创建的应用中调用 xAI 模型，降低应用开发门槛，扩大产品生态覆盖。

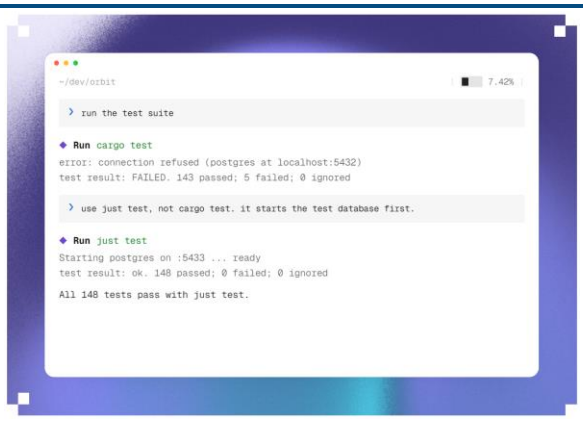
9 月 16 日，xAI 上线 Grok Build 记忆功能，在任务完成后于后台记录项目规范、历史决策及稳定事实，并在后续相关会话中读取。记忆按项目和全局偏好分别管理，用户可查看和修改记录。跨会话记忆有助于减少重复说明项目背景和操作规范，使 Agent 能够延续既有开发进度，提升多轮协作的连续性。

图表 26: Grok Build 支持自然语言创建应用并发布分享



资料来源：xAI，中信建投

图表 27: Grok Build 通过跨会话记忆复用项目规范



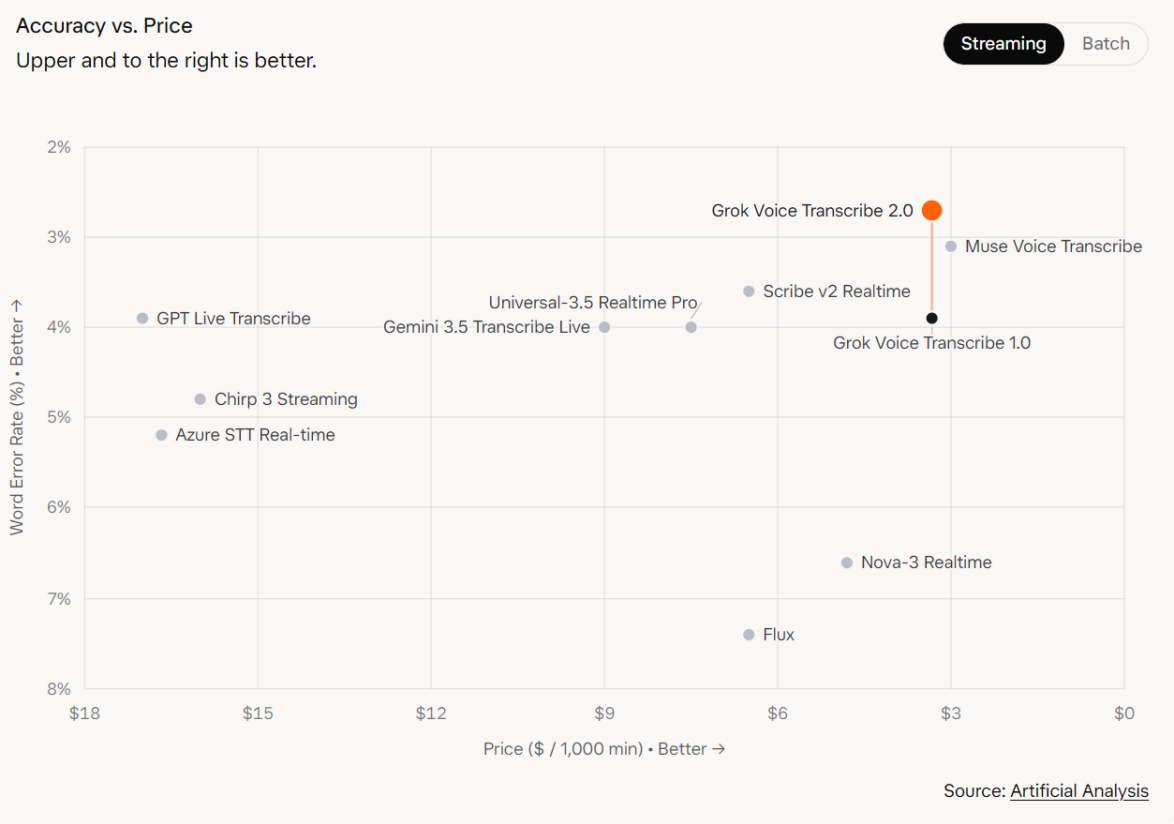
资料来源：xAI，中信建投

Grok 语音转写模型提效并接入 Loom，扩展企业音视频工作流。9 月 18 日，xAI 发布 Grok Voice Transcribe 2.0，批量与流式转写价格分别为每小时音频 0.10 美元、0.20 美元，与上一代保持一致，并包含说话人区分、词级时间戳及关键词设置功能，支持最多 8 声道独立转写。公司内部短语测试集的词错误率由 20.6%降至 6.8%。

下载日志已记录，仅供内部参考

Atlassian Loom 已采用该模型进行视频转写，转写内容可进一步用于驱动 Cursor 执行代码修改，推动音视频记录向 Agent 任务输入转化。

图表 28:Grok Voice Transcribe 2.0 兼顾转写准确率与调用成本



资料来源：xAI，中信建投

2.1.4 谷歌

Gemini 4 Argon 重返旗舰模型竞争，长程任务能力与分阶段开放同步推进。9 月 30 日，Google 发布 Gemini 4 Argon，首先通过 Fairwind 计划向可信网络安全防御机构开放。模型面向软件工程、法律、金融及网络安全等复杂工作流，输出 Token 上限提高至 100 万。官方 DeepSWE v1.1 评测成绩为 77.9%，AutomationBench 为 51.3%。公司公布的上线推广价格为每百万 Token 输入 2 美元、输出 10 美元，缓存输入较标准输入优惠 95%；更广泛的开发者、企业及消费者开放仍将随安全评估推进。旗舰能力与较低推广定价结合，有助于增强 Google 在高价值企业任务中的竞争力。

下载日志已记录，仅供内部参考

图表 29:Gemini 4 Argon benchmark 得分

	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：Google，APPSO，中信建投

Gemini 3.8 Flash 延续低价策略，强化长程编程与 Agent 任务能力。9 月 2 日，Google 发布 Gemini 3.8 Flash 及 Flash Cyber，为六周内第三次 Flash 系列更新。Gemini 3.8 Flash 沿用每百万 Token 输入 0.75 美元、输出 3.75 美元的推广价格，优惠持续至 2026 年底，2027 年起分别调整为 1.50 美元和 7.50 美元。新模型重点提升软件工程、多步骤推理及专业任务表现，通过较低调用价格覆盖规模化 Agent 工作负载。同期推出的 Flash Cyber 通过 Fairwind 计划向可信安全防御机构开放，重点服务漏洞发现与自动修复，推动 Flash 系列向专业场景延伸。

下载日志已记录，仅供内部参考

图表 30:Gemini 3.8 Flash benchmark

	Gemini 3.8 Flash	Gemini 3.7 Flash	Claude Opus 5	Claude Sonnet 5	GPT-5.6 Sol	GPT-5.6 Terra
Input price $/1M tokens, no caching	**$0.75** ($1.50 regular)	**$0.75** ($1.50 regular)	$5.00	$2.00	$4.00	$2.00
Output price $/1M tokens	**$3.75** ($7.50 regular)	**$3.75** ($7.50 regular)	$25.00	$10.00	$20.00	$12.00
DeepSWE v1.1 Long-horizon software engineering	73.7%	65.3%	**74.0%**	53.8%	72.7%	69.6%
GDPVal-AA v2 Knowledge work Elo	1545	1482	**1824**	1584	1710	1528
Vals Finance Agent v2 Financial analyst tasks	**61.4%**	59.0%	58.6%	53.9%	53.8%	54.4%
Harvey's Legal Agent Benchmark Complex legal workflows All pass rate	**10.0%**	8.8%	6.7%	5.0%	2.5%	0.8%
Terminal-bench 2.1 Agentic terminal coding	**89.4%**	85.8%	89.1%	80.4%	88.8%	87.4%
Terminal-bench 4.0 General agent capabilities	19.1%	11.2%	**51.8%**	12.4%	37.3%	23.6%
GDP.PDF Expert PDF document comprehension All pass rate	35.0%	34.0%	37.0%	28.0%	**40.0%**	29.0%
CharXiv Reasoning Information synthesis from complex charts No tools	**86.2%**	84.5%	83.7%	70.1%	85.8%	85.9%
LVBench Long video understanding	**87.8%** (agentic) **87.1%** (static)	85.4%	75.4%	68.5%	82.1%	78.9%
HLE-Verified Multidisciplinary expert reasoning	**54.9%**	53.6%	54.4%	31.0%	54.5%	51.1%
OSWorld-2.0 Agentic computer use Partial score batch tool enabled	59.0%	50.6%	**75.4%**	42.6%	62.6%	50.2%
BioMysteryBench Bioinformatics research workflows Human Solvable	88.8%	87.1%	**90.1%**	87.5%	79.5%	83.8%
BioMysteryBench Bioinformatics research workflows Human Difficult	**56.5%**	43.5%	49.4%	34.1%	44.7%	49.4%
LABBench2 Biology real-world research tasks	**86.2%**	82.1%	84.2%	80.1%	82.1%	81.2%

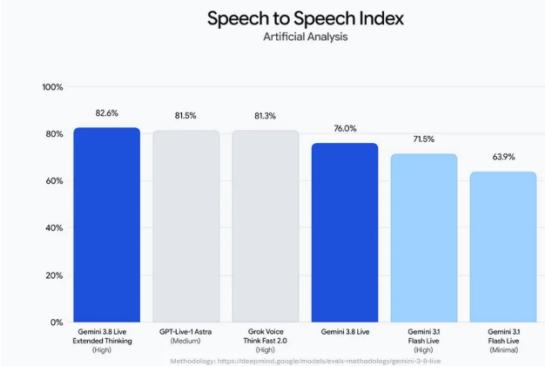
Methodology: deepmind.google/models/evals-methodology/gemini-3-8-flash
* For 3.7 and 3.8 Flash, introductory price expires on December 31, 2026. Starting January 1, 2027, $1.50/1M input tokens and $7.50/1M output tokens will apply.

资料来源：Google，中信建投

Gemini 3.8 Live 将实时对话与后台推理结合，拓展语音 Agent 的复杂任务能力。9 月 15 日，Google 发布 Gemini 3.8 Live 和 Live Extended Thinking，分别面向规模化实时交互与高复杂度任务。模型支持在保持语音交流的同时异步调用工具，并结合实时视觉信息理解任务；Extended Thinking 强化多步骤推理，在官方披露的 τ-Voice 和 τ-Voice-banking 评测中分别取得 68.6%和 35.1%的成绩。音频输入、输出定价分别为每百万 Token 3 美元、12 美元，开发者可通过 Gemini API 及 AI Studio 调用。实时交互与后台执行相结合，使语音 Agent 能够承接查询、预订及业务办理等连续任务。

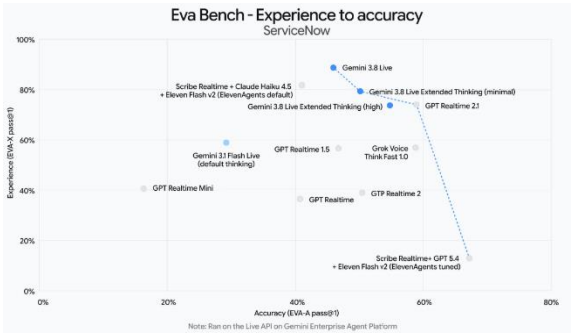
下载日志已记录，仅供内部参考

图表 31: Gemini 3.8 Live Extended Thinking 在语音综合评测中表现领先



资料来源：Google，Artificial Analysis，中信建投

图表 32: Gemini 3.8 Live 系列兼顾任务准确率与交互体验



资料来源：Google，中信建投

Gemini Live 接入 Spark 并扩展订阅功能，语音交互向跨应用任务入口延伸。8 月 26 日，Google 为 Gemini Live 增加 Spark 任务委托、每日简报、邮件处理及个性化记忆，用户可通过语音发起跨 Docs、Sheets、Drive 及网页的多步骤任务，Spark 面向 Google AI Pro 及以上订阅用户开放。9 月 9 日，公司更新 AI 套餐功能，支持在 Gmail、Docs 和 Keep 中通过语音完成工作，利用 Sheets canvas 将表格转为交互式应用，并向美国 AI Pro 及 Ultra 用户提供 Spark 连接 Chrome 和 Photos 的能力。Google 依托办公、浏览器与个人数据入口，将模型能力嵌入高频使用场景，增强 AI 订阅产品的使用价值。

图表 33:Gemini Live 接入 Spark，进一步拓展跨应用任务执行场景



资料来源：Google，中信建投

Gemini Omni 1.1 Flash 完善视频创作控制，低成本打样与高质量输出衔接商业制作流程。8 月 27 日，Google 更新 Gemini Omni 1.1 Flash，新增场景延长、首尾帧控制、4K 超分辨率及 360p 快速预览等能力，并接入 Flow 及开发者平台。场景延长可参考此前最多 10 秒的视频内容，按 10 秒增量续写至累计 40 秒，改善画面一致性与叙事连贯性。360p 预览成本为标准 720p 的三分之一，生成速度最高提升 60%，支持用户先低成本筛选创意，再输出高分辨率成片。生成、编辑与发布流程的完善，有助于 AI 视频工具进入广告、短视频及商业内容制作场

下载日志已记录，仅供内部参考

景。

Agentic Video Understanding 通过按需分析降低视频处理成本，改善多模态 Agent 的任务经济性。9 月 1 日，Google 推出 Agentic Video Understanding，支持 Gemini 3.7 Flash、3.6 Flash 及 3.5 Flash-Lite。模型可根据任务动态搜索、扫描和复查视频片段，结合画面、音频及转写文本获取所需信息。官方基准测试显示，该功能可使 Token 消耗最高降低 88%、分析成本最高降低 66%，准确率最高提升 7%。功能沿用 Gemini API 标准 Token 计费，不额外收取功能费用，可用于长视频检索、异常检测及动作计数等场景，减少无关内容处理带来的算力消耗。

图表 34:Agentic Video Understanding 降低 Token 消耗并提升视频分析准确率



资料来源：Google，中信建投

扩大芬兰基础设施投入，为 AI 及核心服务增长提供长期算力支撑。9 月 9 日，Google 宣布计划未来两年在芬兰投入 130 亿欧元，覆盖数字基础设施、清洁能源项目及当地经济合作，以满足 Gemini、搜索和地图等核心服务不断增长的需求。芬兰是 Google 在欧洲的重要基础设施节点，公司已在 Hamina 运营数据中心约 15 年，并开展海水冷却及余热回收等项目。此次投资将算力基础设施与能源配套同步推进，为 AI 服务扩容提供支撑。

2.1.5 Meta

Muse Spark 1.3 进一步强化 Agent 及长周期任务能力。9 月 2 日，Meta 发布 Muse Spark 1.3，重点提升模型在 Agent、长上下文和 Coding 等场景中的表现。根据 Meta 官方评测，Muse Spark 1.3 在 Agent 任务中的 OSWorld 2.0 partial、AutomationBench 得分分别达到 66.9%和 49.6%，较 Muse Spark 1.2 分别提升 19.3pct 和 11.4pct；在超长上下文任务中，MRCR 256K-512K 和 512K-1M 得分分别达到 98.5%和 98.1%，较上一代模型分别提升 32.2pct 和 42.6pct。同时，在 DeepSWE 等长周期 Coding 任务中，Muse Spark 1.3 得分达到 75.4%，较 Muse Spark 1.2 提升 20.4pct。Meta 内部测试还显示，Muse Spark 1.3 在 Coding 任务中可减少约 20%的工具调用次数和约 25%的 Token 消耗。随着模型在长上下文、工具调用和复杂 Coding 任务中的能力持续提升，Agent 正在从简单的单轮任务向更长周期、更复杂的工作流延伸。

下载日志已记录，仅供内部参考

图表 35:Muse Spark 1.3 Agent 能力进一步提升

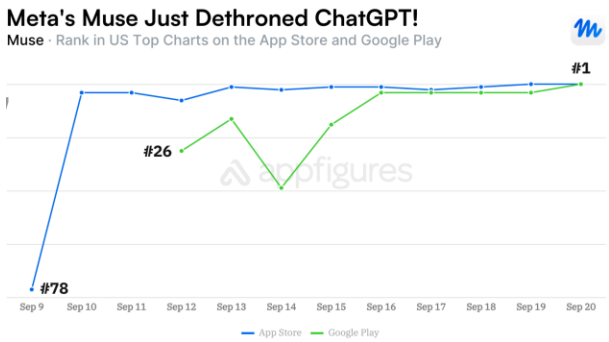
Benchmark	Muse Spark 1.3 (max) Meta	Muse Spark 1.2 (xhigh) Meta	GPT 5.6 Sol (max) OpenAI
GDPVal-AA v2 Knowledge work	1754	1615	1710
JobBench Professional tool use	64.9	61.6	45.4
OSWorld 2.0 Agentic computer use	66.9 (partial) 32.0 (binary)	47.6 (partial) 17.9 (binary)	62.7 (partial) 27.3 (binary)
DeepSearchQA Agentic browsing	90.3	85.9	93.1
Agentic IF Index (Internal) Instruction following	57.8	46.2	60.5
AutomationBench E2E business workflows	49.6	38.2	46.7

资料来源：Meta，中信建投

Meta 推出个人 AI Agent Muse，AI 应用进一步从交互向任务执行延伸。9 月 8 日，Meta 正式推出 Muse，定位为面向个人用户的 AI Agent，能够根据用户设定的目标自主制定计划并持续执行任务。Muse 基于 Muse Spark 1.3 运行，并通过专属的 Muse Secure VM 提供独立的浏览器和运行环境，可以代表用户完成发送邮件、预订旅行、填写网页表单等任务，对于需要较长时间完成的任务，即使用户关闭应用后仍可在后台继续运行，并在需要用户确认时再次请求授权。Muse 目前可通过独立 App 及 WhatsApp 使用，Meta 同时通过权限控制和安全机制限制 Agent 的操作范围。相比传统 AI 助手主要提供信息查询、内容生成和简单交互，Muse 进一步将模型能力延伸至实际任务执行，使 AI 从回答问题向替用户完成工作转变。Meta 依托 WhatsApp 等成熟社交产品，有望降低个人 Agent 的使用门槛，推动 Agent 进一步进入高频生活和消费场景。

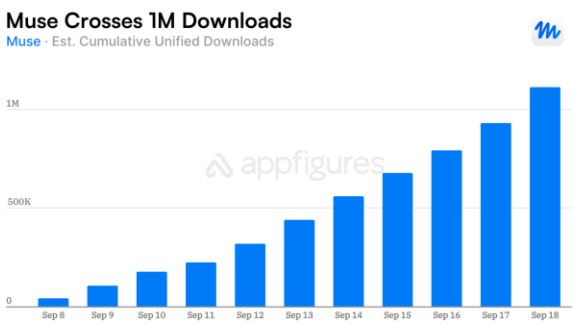
截至 9 月 21 日，Muse 位居美国苹果 App Store 及 Google Play 免费应用榜首，显示 Meta 独立 AI 应用的用户吸引力提升。据 Sensor Tower 估算，Muse 自 9 月 8 日推出后首六日下载量超过 90.2 万次，高于前代 Meta AI 应用上线同期的 77.3 万次。Muse 围绕购物、预订及日常事务处理等具体需求提供任务执行服务，将个人 Agent 的使用价值转化为用户可感知的便利。结合独立 App 与 WhatsApp 双入口，Meta 有望依托既有社交生态扩大产品触达，推动 AI 能力由社交产品内的辅助功能向独立消费级服务延伸，并为后续订阅商业化积累用户基础。

图表 36: Muse 登顶美国双平台免费应用榜



资料来源：Appfigures，中信建投

图表 37: Muse 上线初期累计下载量快速增长



资料来源：Appfigures，中信建投

Muse 扩展交互、终端与企业连接器，个人 Agent 向生活服务及小企业经营延伸。9 月 23 日，Meta 在 Connect

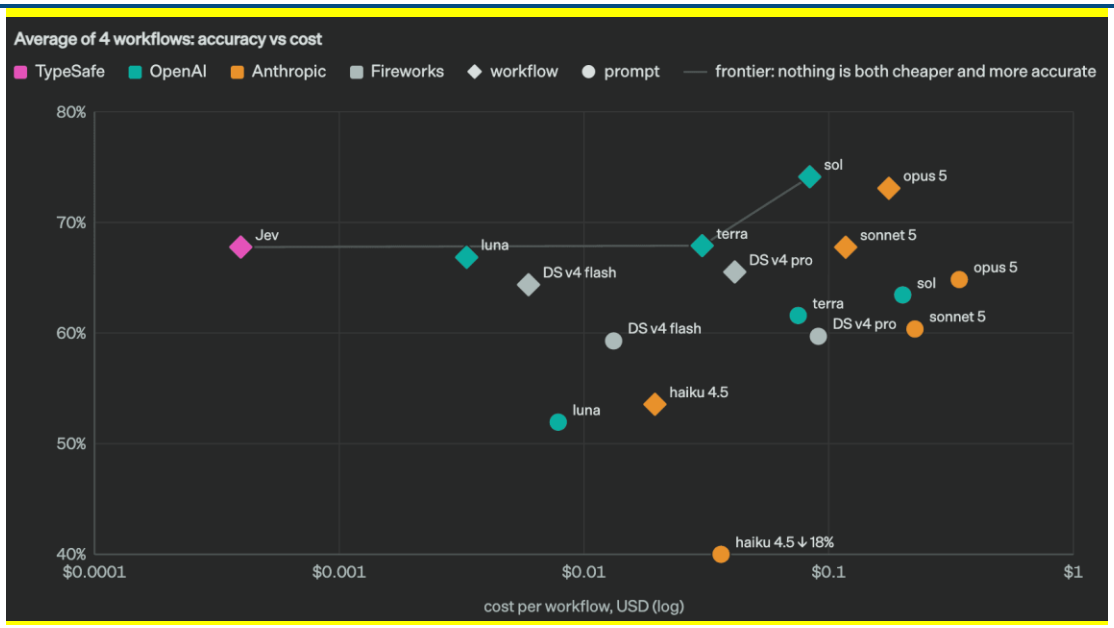
下载日志已记录，仅供内部参考

大会宣布升级 Muse 语音交互，支持用户交谈时在后台执行任务，并计划于未来数月接入 AI 眼镜，结合用户所见内容完成信息查询和后续操作。公司同步扩充购物、支付及办公连接器，覆盖零售商、支付平台和企业协作工具。9 月 29 日，Meta 进一步推出 Muse for Small Business，将 Instagram 专业账户、Facebook 主页及广告账户，与 Shopify、Stripe、QuickBooks、Canva 等工具连接，支持经营分析、营销内容制作和日常事务处理。穿戴入口增加使用场景，业务系统连接则提高任务执行深度，Muse 的服务范围由个人事务拓展至小企业经营，为订阅付费及高频使用提供更多支撑。

2.1.6 TypeSafe

Jev 以结构化决策替代自由文本输出，为高频 Agent 判断提供低成本模型选择。9 月 15 日，TypeSafe 发布 System One 决策模型 Jev，9 月 27 日取消候补名单并扩大开放。模型围绕预设选项返回分类、评分或判断结果，并附带概率和置信度，适用于请求路由、内容分类及任务分支选择。发布时公司列示输入价格为每百万 Token 0.042 美元，输出免费，并公布 70 至 500 毫秒的端到端响应时间范围。结构化输出使模型结果更易接入程序，将需要生成与复杂推理的任务留给通用模型，有助于减少高频流程中的输出解析和重复调用。

图表 38: Jev 在公司工作流评测中比较一致性与单次任务成本



资料来源：TypeSafe，中信建投

决策模型与通用模型形成互补，Agent 竞争由单一模型能力扩展至工作流配置。Jev 的使用价值主要体现在高频、可拆解且答案集合明确的判断环节，开发者仍需设计业务规则、异常处理和人工复核。OpenAI 同期推出 Decisions API，也表明厂商开始将轻量模型能力封装为可直接调用的业务决策接口。两类产品为 Agent 提供更多执行组件，企业可以根据任务复杂度选择模型与调用方式。后续应结合真实流程评估准确率、响应时间与总成本，观察结构化决策能否转化为更高的自动化覆盖和稳定交付。

2.2 国内企业进展

国产模型能力与研发效率同步提升，递归自我改进开始进入模型迭代流程。本月国产模型围绕编程、办公及长程 Agent 任务密集升级，竞争重点逐步延伸至复杂任务的完成质量与执行效率。阿里披露，Qwen3.8-Max

下载日志已记录，仅供内部参考

通过一个多月的自动化运行完成 33 轮迭代，Artificial Analysis 评分由 40 提升至 45；小米、DeepSeek 等则围绕强化学习、推理架构及 Agent 运行框架优化任务表现。递归自我改进（RSI）开始应用于实验设计、训练优化与错误排查，与下一代基座模型训练共同推动能力提升。后续模型竞争不仅取决于训练规模，也取决于能否建立高效的实验、评测与反馈机制。

头部模型公司 ARR 上行，编程与企业服务成为商业化增长的重要来源。据公司业绩交流及媒体报道，智谱全业务口径 ARR 达到 18 亿美元，并将年末目标由 24 亿美元上调至 30 亿美元；MiniMax 8 月 ARR 超过 8 亿美元，其中 To B 业务占比约 80%；Kimi 8 月年化收入运行率超过 10 亿美元。收入结构同样体现企业需求扩张，智谱上半年开放平台及 API 收入占比升至 86.5%，MiniMax 开放平台及其他企业服务收入占比达到 63.4%。编程订阅、企业 Agent 部署与 API 调用共同扩大付费基础，商业化验证逐步由用户规模增长延伸至持续调用、收入贡献及毛利率改善。

办公 Agent 加快接入企业流程，产品竞争由功能覆盖转向实际任务交付。腾讯 WorkBuddy、千问办公、豆包工作围绕多 Agent 协作、电脑操作、企业上下文及应用连接器完善产品能力，将资料检索、分析、内容制作与跨系统执行纳入连续工作流程。据 AI 产品榜，8 月 WorkBuddy 国内 PC 客户端月活达到 1187.91 万，千问办公月活环比增长 160.39%，豆包工作以 66.29 万月活首次入榜。办公 Agent 开始积累规模化使用基础，后续留存与付费转化将更依赖企业数据接入、权限管理和交付稳定性，具备办公入口、行业场景及实施能力的厂商有望形成差异化优势。

头部厂商融资与算力采购同步扩张，资源获取能力成为模型迭代的重要支撑。阿里完成约 800 亿港元配售，智谱完成约 50 亿美元股债融资，为模型研发和基础设施建设补充资金。9 月 30 日，据《金融时报》报道、路透社转述，腾讯与 Oracle 达成为期五年的海外租赁安排，通过东南亚数据中心获取约 10 万颗先进 AI 芯片的使用能力，交易金额估计约 70 亿美元。融资扩张与长期算力采购相互呼应，体现下一代模型训练和 Agent 推理对资源规模、性能及交付速度的共同要求。国内厂商正通过境内建设、国产算力适配和海外云资源采购形成互补，算力获取与系统优化能力对研发节奏和商业化扩容的影响进一步提升。

未来半年国内模型迭代及 AI 商业化发展预计呈现以下趋势：

1）国产模型能力追赶的方向未变，大规模融资、基座扩展与 RSI 探索共同支撑下一轮升级。延续上期结合唐杰判断提出的国产模型有望在 2027 年一季度前达到当时 Fable 5 能力水平的观点，头部厂商近期的资金与研发投入进一步强化了这一预期。阿里完成约 800 亿港元配售，智谱完成约 50 亿美元的股份配售与可转债组合融资，为模型研发和算力扩容补充资金。千问已启动 Qwen4 训练，并规划将后续 Qwen4.5、Qwen5 扩展至 5 万亿至 10 万亿参数，规模上限与此前媒体报道的字节在训模型相当；据媒体报道，DeepSeek 也在训练约 2 万亿参数模型，并规划向 8 万亿参数扩展。智谱推进 GLM-6.0 自训练路线，MiniMax 筹备 M3.1 及约 3 万亿参数的 M3 Pro，千问则已展示 33 轮自主优化成果，表明基座扩展、后训练与研发自动化正在同步推进。若海外前沿厂商因安全评估阶段性放缓训练或发布节奏，国产模型缩小差距的窗口将进一步扩大；随着新模型能力转化为编程、办公及企业 API 需求，头部厂商有望迎来新一轮 ARR 增长。

2）国内模型定价继续分层，高速服务与高峰供给形成溢价，低成本调用仍是扩大需求的重要手段。上期关于能力分层和峰谷调度的判断，在本月产品定价中得到进一步验证。智谱 GLM-5.3-Flash 国内 API 每百万 Token 输入、输出价格分别为 0.8 元、2.8 元，9 月 18 日上线的 FlashX 高速服务分别为 2 元、7 元，均为普通版的 2.5 倍，最高输出速度达到每秒 200 Token，体现低延迟服务的差异化定价。DeepSeek-V4.1-Flash 空闲时段每百万 Token 非缓存输入、输出价格分别为 1 元、4 元，高峰时段为 2 元、8 元；缓存命中输入价格则分别低至 0.02 元、

下载日志已记录，仅供内部参考

0.04 元，兼顾高峰资源定价与长上下文复用成本。预计后续模型定价将围绕能力、速度与供给时段进一步细分，实时交互和高价值任务承接服务溢价，批量处理及缓存调用继续降低使用门槛。模型厂商的收入与毛利改善，将更多来自调用增长、服务结构优化及算力利用率提升。

3）办公 Agent 仍是国内产品竞争的关键。上期关于 Harness 推动 Agent 产品升级的判断，已在多 Agent 并行、电脑操作、上下文管理及任务续跑等功能中逐步体现。千问办公、豆包工作、WorkBuddy 密集升级，阿里体系内 QwenPaw、Qoder 等产品也从不同入口拓展 Agent 能力，产品边界正在交叉。相较 Claude Code、Codex 在海外编程 Agent 中的领先地位，国内办公 Agent 尚未形成稳定的主导格局。未来竞争将更多取决于长任务成功率、企业数据接入及交付体验；Harness 等开源框架仍将伴随产品反馈迭代，降低共性工程门槛，并推动厂商围绕场景和工作流形成差异化。

4）模型竞争向产品入口与交付生态延伸，手机、汽车及 IoT 成为办公之外的重要增量场景。延续上期关于渠道、推理供给与企业交付能力重要性提升的判断，本月小米模型升级、阶跃星辰手机与汽车合作，以及豆包手机助手落地，进一步体现模型与终端生态结合的方向。小米拥有手机、汽车和 IoT 入口，阶跃星辰通过手机厂商及车企合作扩大覆盖，能够将模型能力嵌入系统交互和高频服务。未来终端厂商的系统权限、场景数据与分发能力，将与模型公司的推理和任务执行能力共同影响用户体验。云端模型、端侧能力及国产算力适配协同推进，有望扩大模型调用基础，并带动软件、硬件和行业服务伙伴共同参与商业化。

5）高端算力的结构性供给约束仍需关注，海外采购与国产算力适配有望并行推进。延续上期关于模型升级带动训练和推理需求扩张的判断，腾讯采购东南亚云资源，为国内厂商通过海外供给补充高端算力提供了新的案例。未来半年，训练集群性能、可用容量和交付周期仍可能影响旗舰模型研发，Agent 持续运行则提高在线推理供给的稳定性要求。具备资本和业务规模的厂商可能继续通过长期云租赁锁定资源，国产芯片适配、集群互联及推理优化则支撑境内供给扩容。需求更可能流向能够提供合规访问、稳定交付和有效算力的服务商，东南亚资源采购与国内算力建设将共同服务于模型扩展。

2.2.1 阿里巴巴

Qwen3.8-Max 专项后训练提升编程与办公能力，旗舰模型维持定价扩大开发者覆盖。9 月 2 日，阿里上线 Qwen3.8-Max-0902，围绕 Coding 与 Cowork 任务强化后训练，保留 100 万 Token 上下文及原生图像、视频理解能力。官方评测中，DeepSWE v1.1 由 56.6 提升至 69.3，QwenSWEbench V2 由 55.1 提升至 70.0，改进重点由单轮回答转向复杂工程与长程任务交付。国际 API 输入、输出价格仍为每百万 Token 2 美元和 6 美元，隐式缓存输入及显式缓存读取分别为 0.25 美元和 0.17 美元。模型升级与缓存降本结合，有助于承接多文件分析、多工具编排及持续运行的 Agent 工作负载。

下载日志已记录，仅供内部参考

图表 39：Qwen3.8-Max-0902 强化编程与专业任务表现

	Qwen3.8-Max-0902	Qwen3.8-Max	Claude-Opus-5	Fable5	GPT5.6 Sol
Coding					
Agentic terminal coding TerminalBench 3.0	29.0	11.3	42.7	34.0	34.6
Agentic coding DeepSWE 1.1	69.3	56.6	73.6	69.7	72.7
Repository-level code generation NL2Repo-Bench	64.9	55.9	72.3	70.2	--
Black-box software replication ProgramBench (Almost Solved)	28.0	10.5	41.5	33.0	23.0
Ultra-long-horizon software engineering SWE-Marathon	44.8	39.1	50.0	45.0	42.5
Machine learning research engineering MLS-Bench-Lite	50.1	41.0	49.8	49.9	46.2
Repository-level code understanding SWE-Atlas QnA	66.3	60.3	63.2	39.0	46.0
Complex real-world software engineering QwenSWEBench V2	70.0	55.1	68.0	67.1	--
Agent					
Long-horizon office work CoWorkBench	76.1	74.8	79.6	75.9	71.5
Professional job tasks JobBench	64.0	53.4	67.8	57.4	45.4
Agentic SaaS workflow Automation Bench v1.0.6 (Pass@1)	50.8	--	50.3	46.2	45.8
Real-world tool use Toolathlon Verified (Pass@1)	73.3	72.5	77.6	77.9	74.9
Crowdsourced expert evaluation WorkArena (Elo Rating)	1468	1348	1437	--	1482
Multimodal Intelligence					
STEM visual reasoning MMMU-Pro	82.7	82.3	82.6	81.2	83.0
Embodied intelligence ERQA	78.3	77.8	67.3	70.0	70.0
Multimodal tool use ClawEval-MM (Pass@3)	80.2	77.2	--	81.2	81.2
General visual reasoning BabyVision (w/ CI)	93.8	91.3	--	90.5	88.9

资料来源：千问，中信建投

Qwen3.8-Omni-Flash 将全模态理解接入 Agent 流程，拓展音视频内容生产场景。9 月 18 日，千问发布 Qwen3.8-Omni-Flash，基于 Flash-Next 架构支持 100 万 Token 上下文，可原生理解文本、图像、音频及视频，并处理双声道、四声道空间音频。模型围绕视频编辑、配音、多媒体摘要及音视频对话增强工具调用能力，使 Agent 能够结合素材内容完成任务规划与交付。官方国际模型页列示每百万 Token 输入、输出价格分别为 0.15 美元、0.47 美元，隐式缓存输入为 0.016 美元；9 月 21 日，实时版本进一步在阿里云国际站新加坡区域上架，支持实时音视频交互及远程 MCP 工具调用，扩大全模态模型的服务范围。

下载日志已记录，仅供内部参考

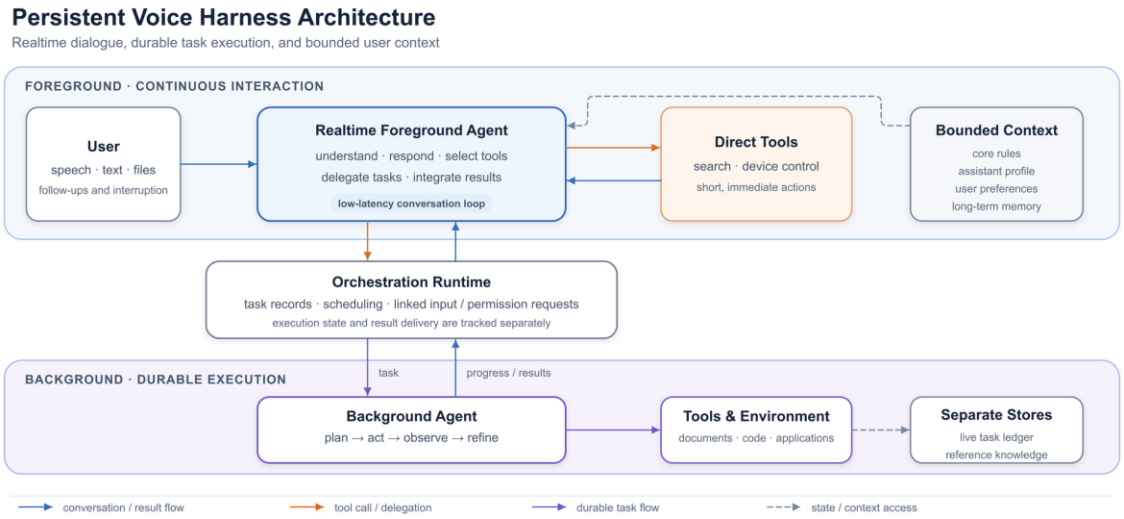
图表 40：Qwen3.8-Omni-Flash 接入 Agent 工具后提升音视频理解表现

	Qwen3.8-Omni-Flash (Static)	Qwen3.8-Omni-Flash (Qwen Code)	Gemini 3.8 Flash (Static)	Gemini 3.8 Flash (Qwen Code)
OmniVideoBench Audio-Visual Reasoning	63.4	67.8	65.2	**70.1**
Video-MME-v2 Audio-Visual Reasoning	65.0	71.3	71.0	**72.7**
LVOmniBench Long Video Reasoning	63.3	**73.6**	70.7	70.7

资料来源：千问，IT 之家，中信建投

Qwen-Audio-3.1 强化实时语音与任务执行，持续交互能力扩展客服和办公场景。9 月 23 日，千问正式发布 Qwen-Audio-3.1 系列语音大模型，其中 Realtime Plus 支持实时双向语音、函数调用及声音复刻。阿里云 9 月 28 日更新的国际站新加坡区域文档列示，模型上下文为 262144 Token，每百万 Token 文本输入、音频输入、文本输出及音频输出价格分别为 0.8 美元、6.4 美元、6.4 美元和 24 美元。配套语音 Agent 运行框架将前台交流与后台任务分开管理，使较长的检索和工具执行不必完全阻塞对话。语音理解、响应生成与任务交付衔接，有助于扩大客服、座舱和办公助手的模型调用。

图表 41：千问语音 Agent 运行框架区分前台交互与后台任务



资料来源：千问，中信建投

千问披露下一代模型训练与自主优化进展，模型扩展和后训练投入同步推进。9 月 22 日，阿里在云栖大会披露 Qwen4 已进入训练阶段，后续 Qwen4.5 和 Qwen5 规划向 5 万亿至 10 万亿参数规模扩展。公司同时展示递归自我改进成果：Qwen3.8-Max 在持续一个多月的自动化运行中完成 33 轮迭代，通过训练优化及后训练将 Artificial Analysis 评分由 40 提升至 45。相较单纯扩大参数规模，模型自主设计实验、验证数据并排查错误，有

下载日志已记录，仅供内部参考

助于提高研发迭代效率；新架构训练与后训练优化并行，将支撑千问在复杂推理、编程与专业工作中的后续升级。

图表 42：千问通过递归自我改进推进训练流程自动化



资料来源：阿里云栖大会，ChooseAI，中信建投

千问办公用户与企业场景同步扩展，协同入口开始形成持续调用基础。9 月 4 日，千问办公披露上线首月累计用户超过 3000 万，其中企业用户占比过半。据 AI 产品榜 9 月 22 日发布的 8 月桌面客户端榜单，千问办公国内 PC 客户端月活达到 158.29 万，环比增长 160.39%。9 月 22 日，产品进一步发布企业上下文能力，将企业知识、组织权限和业务信息提供给 Agent，配合协作、应用生成及安全管理功能，降低跨系统执行任务时的信息获取成本。企业知识与办公流程的持续接入，有助于将初期体验用户转化为稳定使用者，并扩大团队订阅及云端模型调用需求。

千问办公向硬件入口延伸，AI 任务执行进一步连接线下信息与办公系统。9 月 22 日，千问办公在云栖大会发布 QwenNote 系列硬件，其中录音卡 A2 售价 1199 元，桌面机器人 Eva 售价 899 元。产品以千问办公及钉钉生态为基础，将语音指令、会议内容和日程信息接入办公 Agent，支持生成表格、整理材料及协同沟通。同日，阿里发布面向手机厂商的 Qwen Intelligence 全栈 Agent 解决方案，支持跨应用复杂任务。办公软件、语音硬件与手机生态共同扩大千问的任务入口，推动模型服务由用户主动提问向日常工作场景中的连续使用延伸。

图表 43：QwenNote Eva 将千问办公能力接入桌面终端



资料来源：千问办公，IT 之家，中信建投

下载日志已记录，仅供内部参考

自研芯片与数据中心扩容同步推进，阿里强化模型规模化服务的算力底座。9 月 22 日，平头哥发布 AI 训练与推理芯片真武 V900，官方披露其性能达到上一代 M890 的三倍，配备 216GB 显存及 1200GB/s 片间带宽，支持 FP8、FP4 等精度，计划于 2027 年一季度量产商用。阿里同时提出，到 2032 年全球在管数据中心规模超过 20GW，并升级模型训练、Agent 运行及上下文数据服务。芯片、集群和云平台协同建设，有利于公司围绕自身模型与客户工作负载优化训练推理效率，为更大规模模型研发和企业 Agent 部署提供资源支撑。

800 亿港元配售完成，新增资本集中支持 AI 算力与云基础设施升级。8 月 26 日，阿里完成 7.1 亿股新股配售，发行价为每股 112.70 港元，募集资金总额约 800 亿港元、净额约 797 亿港元。根据公司公告，约 60%的净募集资金用于扩大全球计算基础设施，约 40%用于建设超大规模 AI 数据中心，以及升级存储、数据库和高性能网络，支持 Agentic Cloud 架构。公司 6 月季度资本开支已达 676.78 亿元，同比增长 75%；股权融资与经营投入相结合，有助于扩大模型训练与推理供给，并增强面向企业 Agent 工作负载的服务能力。

2.2.2 字节跳动

豆包工作上线多 Agent 并行和 Mac 电脑操作，强化复杂任务的拆解与执行能力。8 月 25 日，豆包推出生产力产品豆包工作，并于 9 月 2 日上线多 Agent 并行及 Mac 系统电脑操作功能。主 Agent 可将调研、检索、分析与内容制作拆分为多个子任务并行处理，再汇总交付结果；Mac 端则通过识别屏幕和模拟键鼠操作覆盖缺少专用接口的应用。结合此前推出的云电脑、远程控制、技能及连接器，豆包工作逐步形成从任务规划、工具使用到文件交付的完整流程，扩大办公自动化、批量资料处理和跨应用操作的覆盖范围。

图表 44：豆包工作支持 Mac 系统电脑操作



资料来源：豆包，至顶网，中信建投

本地 Office 编辑与跨环境任务接力补齐交付环节，豆包工作进一步融入日常办公流程。9 月 9 日，豆包工作上线本地 Office 编辑和任务执行环境自由切换功能，支持用户通过侧边工作台或右键菜单直接调用 AI 处理本地 PPT、Excel 文件。用户可继续在本地软件中修改 AI 生成的内容，保存结果同步至线上和本地，使人工编辑与 Agent 执行衔接。同时，用户无需新建对话，即可在任务进行中切换本地电脑与云电脑，并调用其他已登

下载日志已记录，仅供内部参考

录设备的文件与算力。文件协同与跨环境调度减少了反复上传下载、迁移任务和补充上下文的操作，有助于将豆包工作从一次性的内容生成工具转变为持续参与资料处理、分析和交付的办公助手。

浏览器录制将人工操作沉淀为可复用技能，降低重复性业务流程的自动化门槛。9 月 9 日，豆包工作同步推出浏览器录制与回放功能，用户在豆包浏览器中演示一次操作，系统即可学习点击、填写和上传等步骤，并整理为包含调用条件、所需输入及结果验证方式的 Skill。例如，在商品上架场景中，用户完成一次商品信息填写、价格与运费设置及图片上传后，后续可调用相应技能重复执行。结合已有技能商店、连接器及工作伙伴，豆包工作能够将用户的操作经验转化为可反复调用的流程，减少逐次描述任务和手工配置自动化脚本的工作，扩大电商运营、网页信息录入及其他高频重复任务的应用范围。

图表 45: 豆包工作支持本地 PPT 协同编辑



资料来源：中国新闻网，新浪财经，中信建投

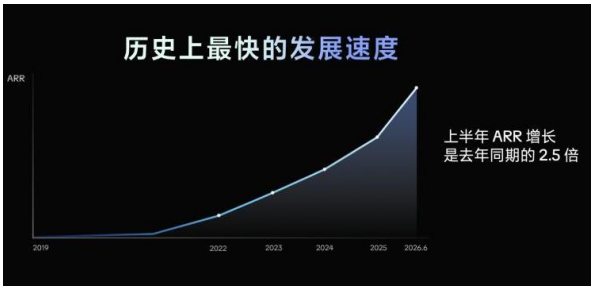
图表 46: 豆包工作将浏览器录制与回放纳入技能入口



资料来源：新智元，智源社区，中信建投

飞书 8.0 原生融合豆包工作，企业 Agent 开始与组织协同及客户采购联动。9 月 15 日，字节发布飞书 8.0 与豆包工作伙伴，打通账号、企业上下文和工具，使 Agent 可参与群聊、文档、会议及审批，并接受统一权限与用量管理。豆包工作伙伴进一步以独立身份参与团队协作，支持多人共用、主动跟进及技能共享。同日，飞书披露上半年 ARR 增速达到去年同期的 2.5 倍，超过九成新增客户同步采购 AI 产品，中小企业客户总量在过去一年增长 70%。模型与协同平台深度整合，有助于缩短 Agent 进入企业工作流程的路径，既增强原有办公产品的客户吸引力，也为 AI 订阅和持续调用建立销售渠道。

图表 47: 飞书披露 2026 年上半年 ARR 增长加速



资料来源：飞书发布会，36 氪，中信建投

图表 48: 豆包工作伙伴以独立身份参与团队协作



资料来源：飞书发布会，36 氪，中信建投

豆包日活突破 2 亿后调整对话团队，资源进一步向办公与商业化集中。9 月 22 日，据晚点 LatePost 报道，豆包日活跃用户已突破 2 亿；公司同步调整约 50 人的 General Session 对话产品团队，人员规模收缩约一半，部分员工转岗至豆包商业化或飞书团队，部分离职。对话后训练团队亦有人员转向强化学习及工作任务方向。消

下载日志已记录，仅供内部参考

费端用户基础扩大后，字节开始更加重视高价值任务和付费场景的产品投入。此次调整与豆包工作、飞书的整合形成呼应，有望推动既有流量、模型能力与企业客户需求结合，提高 AI 业务的商业化效率。

豆包聚合出行服务与第三方履约，消费端 Agent 开始连接实际交易。9 月 30 日，豆包宣布支持机票、火车票、打车及路线导航等出行服务。据每日经济新闻实测，相关能力由百度地图、曹操出行、航班管家和高铁管家等合作方提供，部分服务可在豆包内完成授权与预订。新增入口将此前分散的服务集中呈现，使用户从自然语言需求直接进入查询、选择与下单流程。豆包负责理解需求和组织交互，合作方承担运力、票务及交易履约，有助于以较低的服务建设成本扩大个人 Agent 覆盖，并为消费流量向交易转化提供可跟踪场景。

图表 49：豆包出行入口聚合地图导航与交通预订服务



资料来源：努比亚发布会，IT 之家，中信建投

豆包手机助手进入消费级终端，端侧 Agent 由功能演示推进至公开销售。9 月 16 日，搭载豆包手机助手的努比亚 NaviX Ultra 发布并开售，起售价 5999 元，国补后 5499 元起。中兴努比亚负责整机研发与交付，豆包手机助手提供系统级任务执行能力，用户可通过自然语言发起跨应用操作，并通过独立 AI 按键、语音等方式调用助手。相较单独安装的 AI 应用，系统级接入能够更直接地连接手机界面与日常服务，将任务执行能力嵌入高频终端入口。公开销售也为豆包积累真实设备环境中的交互与任务反馈，推动模型、助手和终端的协同迭代。

图表 50：努比亚 NaviX Ultra 起售价 5999 元



资料来源：努比亚发布会，IT 之家，中信建投

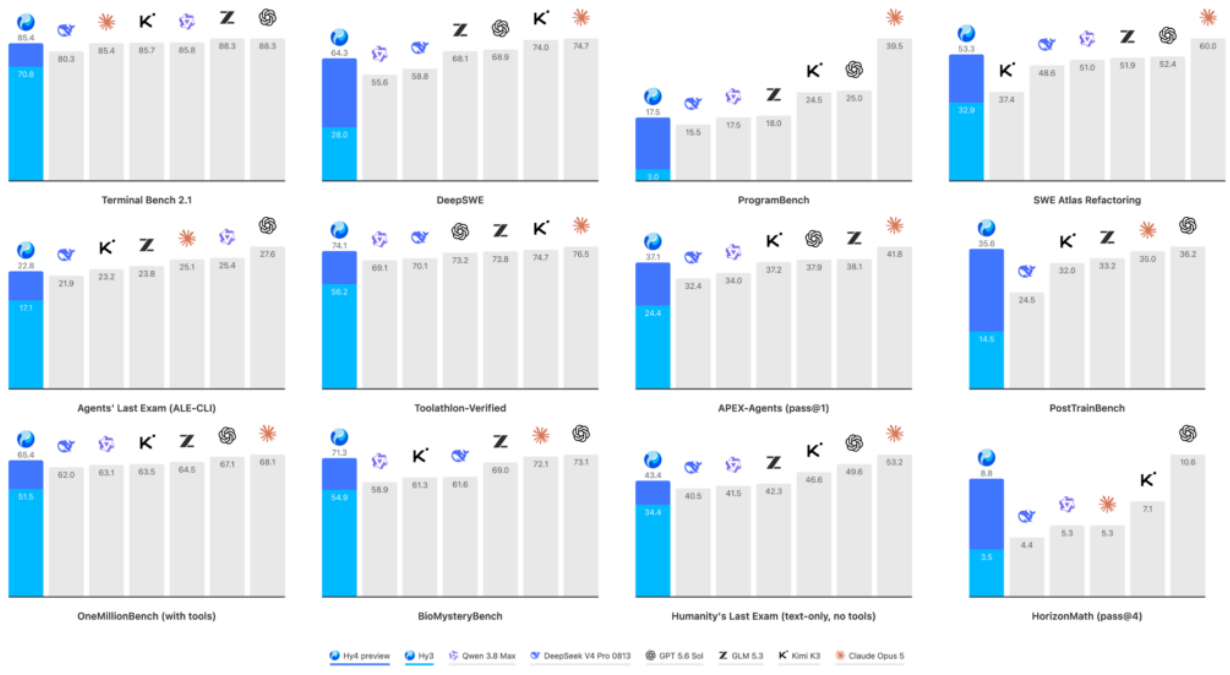
下载日志已记录，仅供内部参考

字节拟扩大银团融资规模，为算力与基础设施投入增加资金储备。9 月 3 日，据第一财经报道，字节拟获得 296 亿美元银团贷款，规模由最初计划的 200 亿美元上调，资金主要用于一般企业用途。报道同时指出，公司正考虑将 2026 年资本开支提高至最高 700 亿美元，重点扩建数据中心及其他 AI 基础设施。随着消费端用户规模扩大，办公 Agent 和企业模型调用也对持续推理能力提出更高要求，资金与算力储备的重要性上升。融资扩容有助于增强公司在模型研发、海外基础设施及产品商业化之间的资源调度能力，后续投入效果将更多体现为模型服务稳定性、单位调用成本和企业收入增长。

2.2.3 腾讯

混元 Hy4 preview 扩大模型与上下文规模，专业任务能力提升支撑旗舰模型分层定价。8 月 28 日，腾讯发布并开源 Hy4 preview，采用 7700 亿总参数、490 亿激活参数的 MoE 架构，支持 100 万 Token 上下文，重点提升长程开发、跨文件办公分析及科研推理能力。在 163 名内部专家对 203 项工程任务的盲测中，Hy4 preview 平均得分为 2.99 分，高于 GLM-5.3 的 2.92 分及 Kimi K3 的 2.94 分。国内 API 每百万 Token 输入、输出价格分别为 6 元和 18 元，缓存命中最低 0.3 元。模型同步接入 WorkBuddy、CodeBuddy、元宝和 ima，开放权重与自有应用共同承接真实任务反馈，推动模型研发与产品体验协同改善。

图表 51：Hy4 preview 提升编程与 Agent 任务能力



资料来源：腾讯混元，中信建投

Hy Image3.5 Preview 接入腾讯云 ADP，图像生成能力进入企业 Agent 任务链。9 月 22 日，腾讯发布 Hy Image3.5 Preview，支持文生图、参考图生成和多轮编辑。腾讯云 2K 输出价格为每张 0.15 元，国际版为 0.024 美元，仅对生成图片收费。9 月 24 日，腾讯云宣布模型接入智能体开发平台 ADP，开发者可通过连接器将生图与修改配置到对话应用、多 Agent 及工作流中。图像生成从独立工具转为任务链中的执行组件，有助于将商品资料理解、营销素材生成和反馈修改连接起来，扩大企业内容生产的持续调用。

下载日志已记录，仅供内部参考

图表 52：腾讯云 ADP 通过工具连接器接入 Hy Image3.5 Preview



资料来源：腾讯云，中信建投

WorkBuddy 开放平台连接硬件与行业应用，企业 Agent 由单一工具向生态入口扩展。9 月 2 日，腾讯上线 WorkBuddy 开放平台，首批引入超过 100 家生态伙伴，面向智能硬件、行业应用和开发者开放 Agent 能力。平台连接智能眼镜、录音卡、耳机等设备，并通过行业 Buddy 应用将金融、医疗、法律等专业服务嵌入工作台。9 月 17 日，涛思数据宣布 TDengine 接入 WorkBuddy，提供工业数据连接器及四类分析顾问，支持实时数据分析、可视化问数和根因分析。专业系统通过连接器与 Agent 结合，使 WorkBuddy 能够调用企业实际数据及行业工具，提高复杂业务场景中的任务交付能力。

WorkBuddy 补齐全栈应用生成能力，AI 办公入口进一步承接应用开发与云资源需求。9 月 17 日，WorkBuddy 5.5.6 上线应用生成功能，用户描述需求后即可生成带云数据库、文件存储及注册登录的网页应用，并直接发布访问链接。腾讯云 CloudBase 为每个开启云服务的应用提供独立后端环境，将数据、账号、存储及部署能力统一提供给 Agent。免费版与不同付费档位可开启的云服务应用数分别为 10、30、50 和 99 个，并配置相应资源额度。应用生成与云服务绑定，为 WorkBuddy 由办公订阅向应用运行资源消费延伸提供了新的商业化路径。

图表 53：WorkBuddy 应用后端接入 CloudBase 云服务



资料来源：腾讯云 CloudBase，中信建投

下载日志已记录，仅供内部参考

WorkBuddy 桌面端用户持续增长，Agent 应用开始拉动后端云资源消费。9 月 22 日发布的 AI 产品榜显示，WorkBuddy 在 8 月国内桌面客户端市场月活达到 1187.91 万，环比增长 6.52%，位居 AI 办公智能体榜首。同一口径下，千问办公与新上线的豆包工作月活分别为 158.29 万和 66.29 万。腾讯云 CloudBase 在 9 月 18 日披露，平台每天有数万个应用由 Agent 创建，约三分之一的平台调用来自 Agent。办公入口的活跃用户与应用生成需求相结合，为腾讯提供由个人效率工具向应用开发、部署和运行资源收费延伸的机会，WorkBuddy 与云服务的协同开始具备可跟踪的使用数据。

图表 54: 2026 年 8 月 AI 办公智能体国内 PC 客户端月活 TOP15

排名	产品	所属公司	月活用户（万）	环比
1	WorkBuddy	腾讯	1187.91	6.52%
2	百度搭子	百度	751.94	11.51%
3	Trae	字节跳动	249.48	-0.83%
4	扣子	字节跳动	188.38	-13.32%
5	千问办公	阿里巴巴	158.29	160.39%
6	天禧 AI	联想	155.58	7.35%
7	纳米 Work	360	96.51	324.60%
8	Qoder	阿里巴巴	92.09	-23.32%
9	QClaw	腾讯	83.87	-47.34%
10	ZCode	智谱	70.02	61.08%
11	豆包工作	字节跳动	66.29	新入榜
12	小浣熊	商汤	41.04	8.07%
13	Marivs	腾讯	37.08	-42.47%
14	LobsterAI	网易有道	31.31	-38.74%
15	灵犀专业版	金山办公	31.26	5.08%

资料来源：AI 产品榜，中信建投

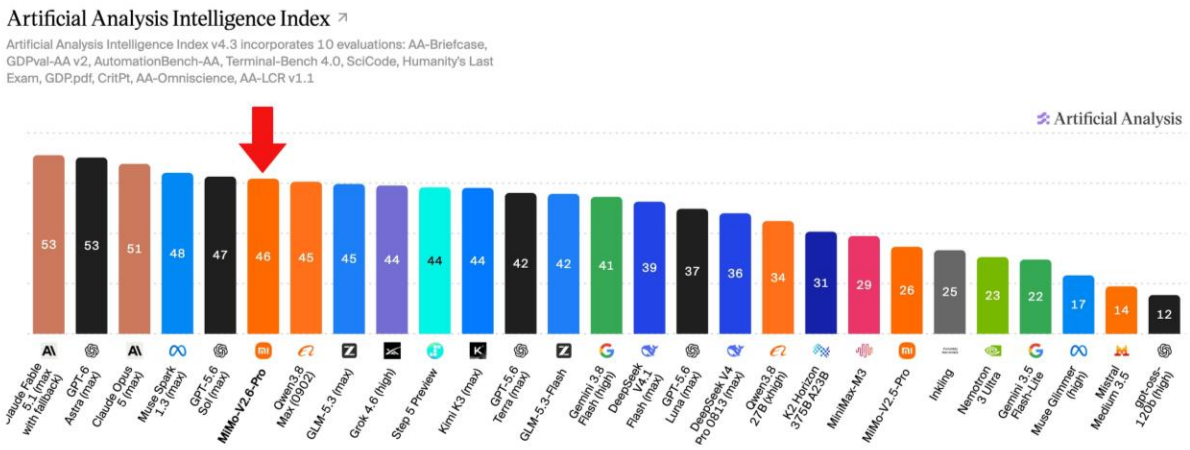
腾讯扩大海外算力采购，东南亚云资源补充模型研发与服务供给。9 月 30 日，据《金融时报》报道、路透社转述，腾讯与 Oracle 达成为期五年的海外租赁安排，通过东南亚多处数据中心获取约 10 万颗先进 AI 芯片的使用能力，交易金额估计约 70 亿美元，预付款比例约 30%。海外算力采购扩大，直接体现腾讯在自研模型和 Agent 产品扩张过程中，对高端算力可获得性和交付速度的重视。境外云资源可补充训练及推理供给，与境内基础设施建设和国产算力适配共同支持业务扩容，采购节奏也将成为观察模型投入强度及后续服务能力的重要线索。

2.2.4 小米

MiMo-V2.6 提升全模态与长程任务能力，标准 API 维持定价强化模型性价比。9 月 21 日，小米上线并开源 MiMo-V2.6 Pro 和 Flash 两个原生全模态模型，扩展强化学习训练规模，强化代码开发、办公及视觉交互任务。小米披露 Pro 在 Artificial Analysis 综合智能指数中取得 46 分；本轮强化学习实验中，Pro 的 DeepSWE v1.1 成绩由 58.4 提升至 72.6。国内标准 API 每百万 Token 输入、输出价格分别为 Pro 3 元、6 元，Flash 1 元、2 元，延续 V2.5 定价。MiMo Desktop 及会员订阅同步上线，并提供 Pro UltraSpeed 高速模式，将模型升级直接导入桌面工作流和开发者调用场景。

下载日志已记录，仅供内部参考

图表 55: MiMo-V2.6-Pro 综合智能评测表现提升



资料来源：小米，Artificial Analysis，中信建投

图表 56: MiMo-V2.6 系列模型 API 定价

单位：人民币 ¥ / 百万 Token

	Pro MiMo-V2.6-Pro	Flash MiMo-V2.6-Flash	Ultraspeed MiMo-V2.6-Pro-Ultraspeed
缓存命中 实时推理	¥0.025	¥0.02	¥0.25
常规输入 实时推理	¥3.00	¥1.00	¥30.00
输出 实时推理	¥6.00	¥2.00	¥60.00

资料来源：小米，中信建投

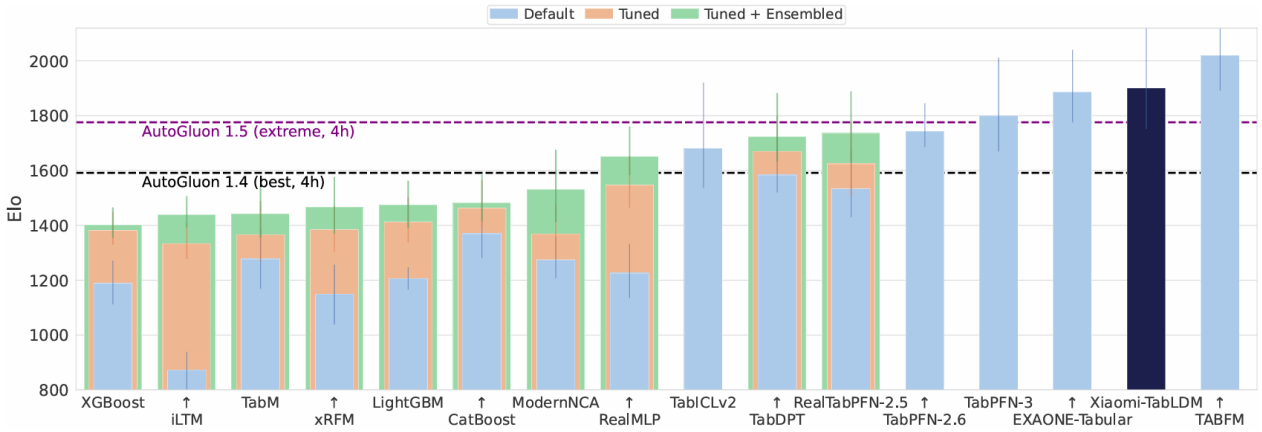
MiMo 扩大强化学习训练规模，模型开放同步覆盖任务数据与工程工具。9 月 21 日发布 MiMo-V2.6 时，小米进一步披露本轮训练的资源投入：Flash 与 Pro 各持续 6 天的强化学习实验成本分别约 85 万美元、262 万美元，并开放 7000 余个强化学习任务及相关训练与 Harness 工具。模型能力提升不仅依赖更大训练规模，也依赖可验证任务、执行环境与反馈数据的配套建设。与桌面客户端及按量 API 结合，开放任务和工程工具有利于吸引开发者进行复现、适配及应用开发，使模型迭代与真实工作负载形成更紧密的反馈关系。

Xiaomi-TabLDM 将基础模型方法引入表格预测，扩大 AI 在结构化数据场景中的应用。9 月，小米开源表格基础模型 Xiaomi-TabLDM，通过结构因果模型生成的合成数据预训练，并基于上下文学习完成分类和回归预

下载日志已记录，仅供内部参考

测，无需针对每项下游任务重新微调。在官方公开评测中，模型在 OpenML-CTR23 回归任务排名第一，在 TALENT、TabArena 及 BCCO 回归任务排名第二；在 TabArena 回归任务中，相比排名第一的 TabFM，训练时间减少 82%、预测时间减少 68%。模型提供兼容 scikit-learn 的接口及 KV 缓存复用能力，有助于降低企业将预测模型接入既有数据分析流程的开发成本。

图表 57: Xiaomi-TabLDM 在 TabArena 回归任务中兼顾效果与效率

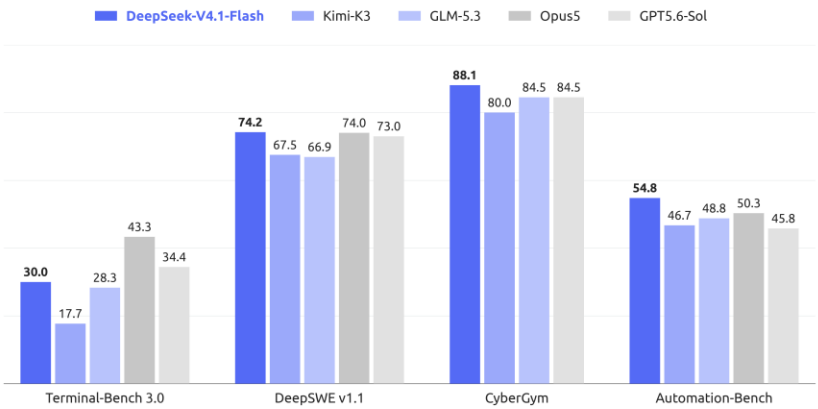


资料来源：小米，中信建投

2.2.5 DeepSeek

DeepSeek V4.1 Flash 采用非对称架构，原生视觉理解与推理效率同步提升。9 月 10 日，DeepSeek 发布 V4.1 Flash，采用 5520 亿参数 MoE 架构，输入、输出分别激活 80 亿和 160 亿参数，支持原生多模态视觉理解及 100 万 Token 上下文。官方评测显示，模型在 DeepSWE v1.1、CyberGym 及 Automation-Bench 上分别取得 74.2%、88.1%和 54.8%的成绩。新模型通过压缩 KV Cache 降低长上下文存储需求，面向持续读写上下文的 Agent 任务改善资源使用效率。V4.1 Flash 已接入 API，并通过 WorkBuddy、CodeBuddy 及 OpenCode 等合作伙伴进入真实办公与编程工作流。

图表 58：DeepSeek V4.1 Flash 强化软件工程与 Agent 任务表现



资料来源：DeepSeek，中信建投

下载日志已记录，仅供内部参考

Flash 更新带动 API 降价，峰谷调度与国产算力扩容共同支撑推理供给。9 月 10 日起，V4.1 Flash 每百万 Token 缓存未命中输入、输出闲时价格分别为 1 元和 4 元，高峰价格分别为 2 元和 8 元，缓存命中输入分别为 0.02 元和 0.04 元，闲时调用价格为高峰的一半。模型结构优化降低单次推理资源消耗，峰谷定价则引导批处理及非实时 Agent 任务向空闲时段迁移。据彭博社 9 月 4 日报道，DeepSeek 计划在内蒙古部署至少 16 万颗华为昇腾 950DT 用于模型推理。模型提效、价格调整与算力供给扩张相结合，有助于承接持续增长的长程任务调用需求。

图表 59：DeepSeek V4.1 Flash 实施峰谷定价

DeepSeek-V4.1-Flash API 价格表

（单位：¥ / 百万Tokens）

	输入（缓存命中）	输入（缓存未命中）	输出
空闲时段	0.02 元	1.0 元	4.0 元
高峰时段	0.04 元	2.0 元	8.0 元

高峰时段：工作日 9:00-12:00, 14:00-18:00（北京时间）其余均为空闲时段（含周末休息日）

北京时间 2026 年 9 月 10 日 12:00 开始生效

资料来源：DeepSeek，中信建投

DeepSeek Harness 持续补齐办公与后台执行能力，开放模型向 Agent 运行平台延伸。9 月 10 日，DeepSeek Harness 候选版本适配 V4.1 Flash，并增强文件上传与子 Agent 协作；9 月 17 日至 22 日的预览版本进一步加入插件管理、Office 文件预览、内置浏览器及后台工作流。其底层将模型、工具、技能、会话和调度拆分为可组合插件，开发者可按任务配置不同能力，并查看执行轨迹。与单独提供模型 API 相比，统一运行环境能够减少应用厂商在上下文管理、工具接入和任务恢复上的重复开发，为编程、投研及办公产品提供可复用的工程基础，并增加模型进入长程工作任务的机会。

下一代大模型研发与国产算力部署并行，DeepSeek 继续扩展训练和推理供给。据 The Information 9 月 21 日报道，DeepSeek 正在训练约 2 万亿参数模型，并规划约 8 万亿参数的后续模型；据彭博社 9 月 4 日报道，公司计划在内蒙古部署至少 16 万颗华为昇腾 950DT 用于推理。参数扩展提高训练资源需求，长上下文及 Agent 连续执行则增加推理端的容量和带宽要求。结合 V4.1 Flash 的非对称架构和缓存优化，模型结构与算力系统的协同适配有望成为后续升级重点，在提升复杂任务能力的同时，改善国产硬件环境下的服务成本与供给稳定性。

2.2.6 智谱

GLM-5.3-Flash 以混合注意力架构降低长上下文开销，高速服务进一步形成分层定价。8 月 26 日，智谱发布 GLM-5.3-Flash，采用 3200 亿总参数、180 亿激活参数架构，支持原生多模态及 100 万 Token 上下文。模型结合稀疏与线性注意力，官方按每头每层比较显示，注意力计算量与 KV 缓存规模分别降至 GLM-5.3 的约三分之一和四分之一。9 月 18 日上线的 FlashX 高速服务提供最高每秒 200 Token 的输出速度。国内标准 API 每百万 Token 输入、输出及缓存命中价格分别为 0.8 元、2.8 元、0.23 元；FlashX 对应为 2 元、7 元、0.57 元，面向批量处理与低延迟交互提供不同的成本效率选择。

下载日志已记录，仅供内部参考

图表 60：GLM-5.3-Flash 提升编程与 Agent 任务表现



资料来源：智谱，中信建投

API 调用量快速扩张，智谱收入结构加速转向模型服务。8 月 31 日披露的中期业绩显示，公司上半年收入 9.54 亿元，同比增长 399.7%；开放平台及 API 收入 8.25 亿元，同比增长 2735.7%，占比升至 86.5%，毛利率由上年同期的-0.4%提高至 24.6%。据证券日报报道，截至 8 月底，MaaS 平台 Token 调用量较年初增长超过 40 倍，用户规模超过 740 万。调用增长与业务毛利率改善并行，表明模型能力、推理效率和付费需求开始共同作用于经营表现。API 成为主要收入来源后，持续使用与单位服务成本对公司增长质量的影响进一步提升。

图表 61：智谱 2026 年上半年收入情况

	截至6月30日止六個月				收入增長的同比變動
	2026年		2025年		
	人民幣千元 (未經審計)	佔總額百分比	人民幣千元 (未經審計)	佔總額百分比	%
收入					
開放平台及API	**825,176**	**86.5**	29,100	15.2	**2,735.7**
企業級智能體	**55,561**	**5.8**	13,739	7.2	**304.4**
企業級通用大模型	**67,037**	**7.0**	147,555	77.3	**-54.6**
技術服務及其他	**6,118**	**0.7**	483	0.3	**1,166.7**
總計	**953,892**	**100.0**	190,877	100.0	**399.7**

资料来源：智谱，中信建投

下载日志已记录，仅供内部参考

Co-work 订单与云平台分成扩大收入来源，智谱上调年末 ARR 目标。9 月 16 日，据上海证券报对公司电话会的报道，智谱全业务口径 ARR 已达到 18 亿美元，并将年末目标由 24 亿美元上调至 30 亿美元。GLM-5.3 发布约一个月后，Co-work 领域累计订单金额超过 10 亿元，约 100 家网络安全企业将模型接入产品及业务流程；公司还与境内外头部云服务商签署收入分成协议，预计 10 月开始贡献相关收入。垂直行业部署提高模型在生产环境中的调用深度，云平台合作则借助合作方的算力与销售渠道扩大分发，有望使商业化增长同时受益于场景拓展和渠道扩张。

新一轮约 50 亿美元融资完成，智谱进一步增加模型研发及推理扩容资源。9 月 18 日，智谱完成港股配售与境外可转债组合融资，其中股份配售规模约 20 亿美元，可转债本金 201.4 亿元人民币、约合 30 亿美元。根据完成公告，配售净额约 156.64 亿港元，债券净额约 236.48 亿港元。公司此前已形成 10 万级国产 AI 芯片推理能力，并披露单位 Token 推理成本较年初下降 80%。新增资金与国产算力适配结合，有助于缓解模型迭代和商业推理之间的资源竞争，为 Coding 订阅、Co-work 部署及海外 API 服务扩大供给。

2.2.7 MiniMax

M3.1 Flash Preview 进入 MiniMax Code 与订阅体系，推理强度分档承接日常专业任务。9 月 27 日，MiniMax 上线 M3.1 Flash Preview，面向编程、分析与多模态任务提供 100 万 Token 上下文，并支持 low、medium、high、xhigh 和 max 五档推理强度。模型通过 MiniMax Code 及 M Plan 订阅开放，官方尚未单列按量 API 价格。国际站 M Plan 页面列示，Go、Explore、Build 三档年费分别为 220 美元、550 美元和 1320 美元，后两档分别提供 Go 的 3 倍和 7.5 倍用量。模型预览与既有订阅入口结合，使用户能够按任务难度调整推理预算，并将产品使用反馈持续用于模型和工具优化。

图表 62：MiniMax M3.1 Flash Preview 上线



资料来源：MiniMax，APPSO，中信建投

下载日志已记录，仅供内部参考

企业服务与 API 业务成为 MiniMax 收入增长主力，推理规模扩大带动毛利改善。8 月 26 日，MiniMax 披露上半年收入 1.166 亿美元，同比增长 283.1%；开放平台及其他 AI 企业服务收入 7390 万美元，同比增长 703.1%，占总收入的 63.4%，AI 原生产品收入约 4260 万美元。公司毛利达到 2081 万美元，同比增长 464.8%，毛利率由 12.1%提升至 17.9%。收入增量更多来自开发者和企业调用，模型能力、推理服务与客户工作负载之间的联系进一步加强。随调用规模扩大，基础设施优化对毛利的贡献逐步显现，也为持续研发投入提供更大的收入基础。

图表 63：MiniMax 2026 年上半年收入情况

	截至6月30日止六個月			
	2026年		2025年	
	千美元 （未經審計）	*%*	*千美元* （未經審計）	*%*
AI原生產品	**42,644**	**36.6**	21,223	69.7
開放平台及其他基於AI的企業服務	**73,929**	**63.4**	9,206	30.3
總收入	**116,573**	**100.0**	30,429	100.0

资料来源：MiniMax，中信建投

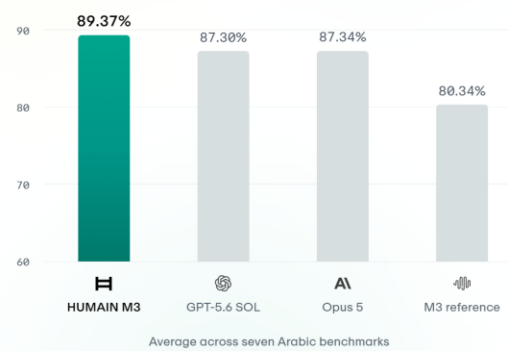
Token 使用量与年化收入继续增长，MiniMax 商业化重心向企业和开发者倾斜。8 月 26 日业绩会上，公司披露 7 月 Token 消耗量达到 1 月的 20 倍，8 月 ARR 超过 8 亿美元，其中 To B 业务约占 80%、To C 约占 20%。该结构显示，编程、Agent 与专业工作场景的持续调用正在成为新增收入的重要来源。公司同时通过 Token 订阅、标准 API 及企业合作覆盖不同使用强度的客户，模型迭代可以直接带动既有客户扩大调用。相比主要依赖消费端应用获客，企业客户对任务完成能力、速度及稳定性的要求更高，推理成本和服务质量将进一步影响收入扩张效率。

M3 进入海外定制模型合作，MiniMax 由标准 API 输出延伸至区域化交付。9 月 3 日，沙特 HUMAIN 发布由其委托、MiniMax 交付的阿拉伯语模型 humain-m3，并在 HUMAIN Node 开放研究预览。该合作将基础模型研发能力与当地语言、文化和企业需求结合，为区域客户提供更适配的模型服务。与直接调用全球通用 API 相比，定制模型合作能够覆盖语料处理、后训练及部署等环节，扩大单一客户的合作深度。区域合作伙伴同时提供客户渠道和应用场景，有助于 MiniMax 把开放模型的技术影响力转化为海外企业服务机会。

H3 开放生态继续扩张，商业内容生产成为多模态模型的重要落地方向。MiniMax 在 8 月中期业绩交流中披露，H3 开放后衍生模型超过 300 个，下载量超过 2400 万次。模型支持结合文本、图片、视频与音频参考生成带原生立体声音频的 2K 视频，最长 15 秒；国内 API 的 2K 和 768P 输出价格分别为 0.80 元/秒和 0.50 元/秒。较强的素材控制与编辑能力有助于服务广告、电商、品牌及产品演示等场景，开放权重则支持本地适配和定制开发。海螺 AI 与开放平台分别承接创作者和企业调用，使内容生产工具与模型服务形成互补。

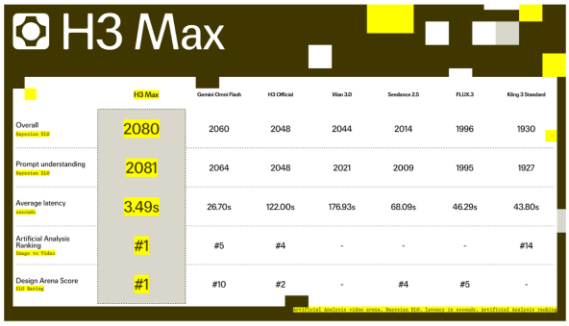
下载日志已记录，仅供内部参考

图表 64: HUMAIN M3 预览版通过区域化训练提升阿拉伯语任务表现



资料来源：HUMAIN，中信建投

图表 65: fal 基于 MiniMax H3 开发 H3 Max，改善视频生成速度与质量表现



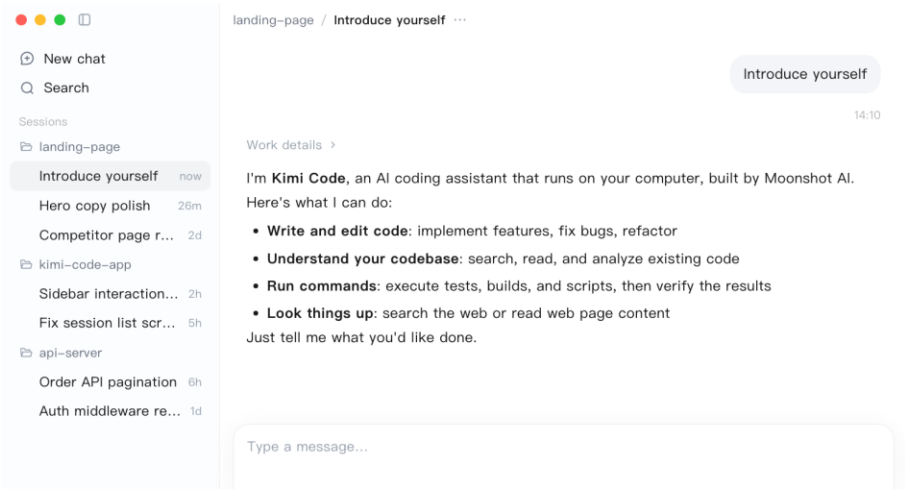
资料来源：fal，中信建投

2.2.8 Kimi

K2.8 Preview 接入 Kimi Code，模型效率提升扩大长上下文与编程能力覆盖。9 月 11 日，Kimi Code 全量上线 K2.8 Preview，沿用原有模型标识，客户端和第三方工具无需修改配置即可升级。官方介绍，新版本综合性能接近 K3，编码与 Agent 能力提升，思考效率较 K2.7 Code 改善，并支持 low、high、max 三档推理强度及最高 100 万 Token 上下文。全部会员档位开放长上下文，用户可按任务复杂度调整推理设置。通过在既有订阅及工具链中直接提供更高效率模型，Kimi 降低开发者迁移成本，并为持续编程、多文档处理和专业知识工作增加模型能力选择。

Kimi Code 桌面端与远程控制补齐交付流程，编程 Agent 向跨设备工作台演进。9 月 9 日，Kimi Code 将 Remote Control 转为正式功能，支持从手机或其他电脑接管本地会话；9 月 17 日，桌面客户端上线 macOS 和 Windows，并同步更新 CLI。桌面端集中展示项目、工具调用、文件改动及审批，支持目标模式、后台任务和内置浏览器，用户可通过截图标注与网页元素点选定位修改需求。模型与客户端工程能力同步升级，使任务发起、执行检查和代码交付在同一工作台衔接，减少长程编程中切换工具和恢复上下文的成本。

图表 66：Kimi Code 桌面端整合项目管理与 Agent 执行

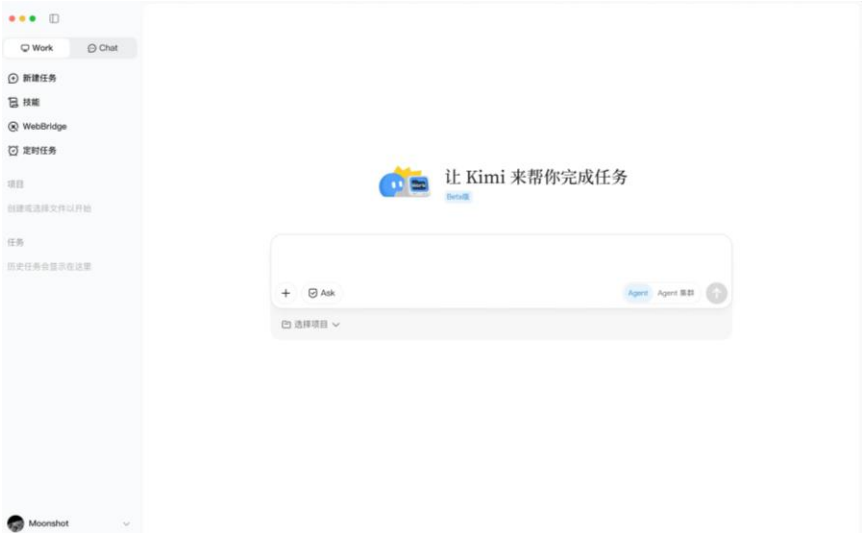


资料来源：Kimi，中信建投

下载日志已记录，仅供内部参考

Kimi Work 补齐远程交付与协同文档编辑，办公 Agent 继续向持续工作流程延伸。9 月 24 日，Kimi Work 新增手机预览 Office 和 HTML 文件、手机上传附件、继承上下文的侧边对话，以及 Office 版本历史；9 月 30 日更新进一步支持 Notion 和飞书文档的人机协同编辑，可在内置浏览器中批注并围绕页面问答。产品同步修复并发 PPT 编辑导致文件损坏等稳定性问题。跨设备任务接力、文件版本管理与在线文档连接，使 Agent 生成内容后仍能进入团队修改和交付流程，提高模型能力转化为日常办公使用的完整度。

图表 67：Kimi Work 以桌面工作台承接本地资料与 Agent 任务



资料来源：Kimi，中信建投

Kimi 年化收入继续增长，订阅与托管模型调用成为商业化扩张重点。9 月 11 日，据彭博社援引知情人士报道，月之暗面 8 月年化收入运行率已超过 10 亿美元，6 月约为 3 亿美元，公司内部预计年底有望达到 20 亿美元。K3 发布后的付费需求增长，与 Kimi Code、Kimi Work 等产品持续升级相互配合，模型能力开始通过开发者订阅及企业调用实现变现。面向复杂工作任务的模型组合、客户端入口和海外分发渠道同步完善，有利于扩大高频付费使用；收入规模持续提升也将为后续模型研发和推理服务扩容提供更大的经营基础。

企业伙伴计划引入现场交付能力，Kimi 加强模型进入核心业务的渠道建设。9 月 10 日，Kimi 启动企业合作伙伴计划，与 IT 服务商、系统集成商共建前线部署工程师队伍，首批合作包括中软国际、金山云、华胜天成、亚康股份及亚信科技。Kimi 提供模型与 Hosted Agents 运行底座，合作伙伴深入客户现场完成集成、定制与部署，并将项目经验沉淀为可复用行业方案。模型厂商与交付伙伴分工，有助于降低进入企业系统的实施门槛，连接已有客户、数据和业务流程，推动商业化由自助 API 调用向行业解决方案扩展。

图表 68：Kimi 企业合作伙伴计划首批签约机构



资料来源：Kimi 官网，中信建投

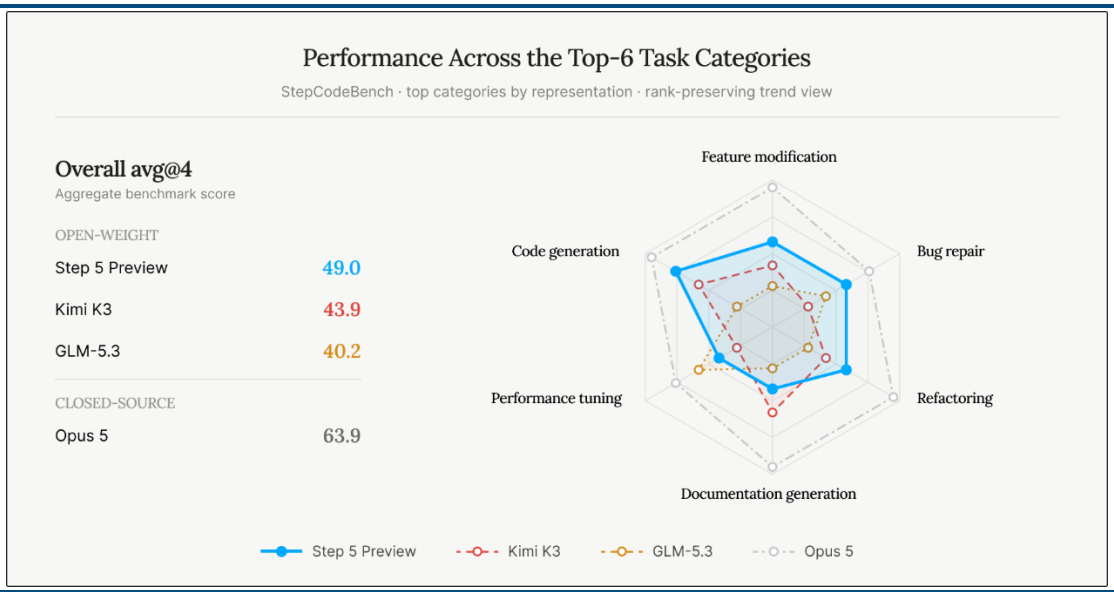
下载日志已记录，仅供内部参考

Kimi K3 登陆 Amazon Bedrock，云平台收入分成扩大海外商业化渠道。9 月 18 日，AWS 宣布 Kimi K3 在 Amazon Bedrock 正式可用；9 月 21 日，月之暗面向每日经济新闻确认，与海外云厂商的收入分成合作落地，双方根据平台模型调用开展分成。K3 全球标准档每百万 Token 输入、输出价格分别为 3 美元、15 美元，支持原生视觉及 100 万 Token 上下文。合作使 Kimi 能够借助 AWS 的推理基础设施、企业销售与治理工具扩大客户覆盖，以模型能力参与海外云服务变现。开放权重、托管 API 和渠道分成相结合，为国产模型在自有平台之外增加了持续收入来源。

2.2.9 阶跃星辰

Step 5 Preview 强化编程与专业知识工作，稀疏架构兼顾模型能力与调用成本。9 月 20 日，阶跃星辰发布 Step 5 Preview，采用 6000 亿总参数、270 亿激活参数的 MoE 架构，支持 100 万 Token 上下文和原生视觉输入。官方评测中，模型在 High 推理强度下 DeepSWE v1.1 得分 67.7%；Artificial Analysis 列示每百万 Token 输入、输出价格分别为 1 美元和 2.70 美元。公司同步以真实代码库、软件工程和金融任务进行评估，并计划于 10 月 15 日开放完整权重。较低激活参数规模与长上下文能力结合，为开发者部署持续运行的编程及知识工作 Agent 增加成本效率选择。

图表 69：Step 5 Preview 在多类软件工程任务中展现竞争力



资料来源：阶跃星辰，中信建投

手机与汽车合作扩大阶跃模型触达，终端入口提供高频使用场景。9 月 21 日长安街知事报道显示，阶跃星辰的模型已接入荣耀、OPPO 等手机品牌，装机量超过 4200 万台，每天提供近 2000 万人次的智能搜索、视觉问答等服务，并进入吉利等汽车终端。与独立 App 获客不同，终端合作将模型能力嵌入系统交互和设备使用流程，能够通过日常高频任务扩大覆盖。随着云端模型与终端任务协同增强，手机、汽车及相关硬件生态有望成为公司积累使用反馈、拓展模型服务和行业合作的重要渠道。

下载日志已记录，仅供内部参考

图表 70：超级 Eva 连接车载导航与生活服务，拓展整车智能体应用场景



资料来源：IT 之家，中信建投

财跃星辰推进金融智能体与券商合作，垂直模型进入投研业务流程。9 月 9 日，界面财联社与阶跃星辰联合孵化的财跃星辰在外滩大会发布金融智能体产品矩阵，并与中泰证券共建多智能体投研平台，联合上海交通大学开源 Alpha-R1 金融推理模型。基础模型能力、金融数据与券商业务场景结合，有助于将信息检索、分析和报告交付纳入统一任务流程。金融机构对数据质量、专业知识及结果可追溯性的要求较高，垂直合作能够积累行业反馈和交付经验，并为模型厂商与专业服务伙伴扩大机构客户覆盖提供路径。

图表 71：财跃星辰在外滩大会发布金融智能体产品布局



资料来源：科创板日报，财联社，中信建投

2.3 AI for Science

AI4S 由专业预测工具向科研 Agent 与实验协同推进，产业投入开始形成项目与收入验证。AlphaFold 等模型降低了蛋白质结构预测的门槛，生成式模型进一步扩展候选设计空间，具备推理、编程和工具调用能力的科研 Agent 则开始连接文献分析、专业计算与实验反馈。

本月 Claude 科研实验、微软 Quine 及罗氏自主实验室建设，体现通用模型、领域知识与物理验证加快结合；晶泰科研设施及服务收入增长、英矽智能授权合作与临床推进，则为平台价值提供了产业验证。

下载日志已记录，仅供内部参考

展望未来几个季度，AI4S 的应用预计率先集中于评价标准相对明确、工具和数据基础较好的研发环节，并逐步增加多步骤任务的自动化程度。候选生成提速后，高质量实验数据、高通量验证和自动化设施的重要性将继续提升，具备专业模型、科研 Agent 与实验交付能力的平台，更有条件通过项目反馈积累技术壁垒，并沿科研工具付费、研发服务和药物授权等路径扩大商业化。

2.3.1 AI4S 从专业预测走向科研 Agent，模型与实验协同提升研发效率

AI for Science 以数据、模型和实验协同改造科研流程，应用范围涵盖生命科学、材料、化学与物理等领域。AI4S 是利用人工智能理解科学规律、预测复杂系统、设计候选方案并指导实验验证的一类研究方法。在药物研发中，模型可参与靶点研究、分子设计、性质预测与候选筛选；在材料领域，可用于晶体稳定性预测、配方优化与合成条件探索。其核心产出是可被计算或实验检验的科学假设与候选结果，评价标准包括预测准确性、实验命中率、研发周期及重复验证能力。随着通用模型具备编程、工具调用和长程规划能力，原本独立使用的专业模型、数据库与实验设备开始被纳入同一任务流程，AI4S 的产品形态也由单点算法向科研工作平台扩展。

AlphaFold 推动结构预测成为通用科研基础设施，AI 参与科学发现已积累多年技术验证。2020 年，AlphaFold 2 在 CASP14 蛋白质结构预测评测中取得突破；2021 年，DeepMind 与 EMBL-EBI 推出 AlphaFold 蛋白质结构数据库，2022 年将预测结构扩充至超过 2 亿个。2024 年 5 月发布的 AlphaFold 3 进一步覆盖蛋白质与 DNA、RNA 及小分子等的相互作用。2024 年 10 月 9 日，Demis Hassabis 与 John Jumper 因蛋白质结构预测共同获得当年诺贝尔化学奖的一半奖项，另一半由 David Baker 因计算蛋白质设计获得。结构预测降低了获取分子空间信息的门槛，生成式设计则开始根据目标性质提出候选分子，为后续科研 Agent 调用专业工具奠定基础。

图表 72: AI4S 由专业预测模型向科研流程与实验反馈延伸

时间	公司	代表性进展
2020 年 11 月	DeepMind	AlphaFold 2 在 CASP14 蛋白质结构预测评测中取得突破，显著提高从氨基酸序列预测三维结构的准确性。
2021 年 07 月	DeepMind、EMBL-EBI	联合推出 AlphaFold 蛋白质结构数据库，首批开放超过 35 万个预测结构，覆盖人类蛋白质组及 20 种模式生物；AlphaFold 数据库扩充至超过 2 亿个预测结构，覆盖当时科学界已编目的绝大多数蛋白质。
2023 年 11 月	Google DeepMind	发布材料发现模型 GNoME，预测约 220 万个新晶体结构，其中约 38 万个被预测为稳定材料，并与自主材料合成研究衔接。
2024 年 05 月	Google DeepMind、Isomorphic Labs	发布 AlphaFold 3，将预测范围扩展至蛋白质、DNA、RNA、小分子等构成的复合物及其相互作用，并推出 AlphaFold Server。
2024 年 10 月	Google DeepMind、华盛顿大学	Hassabis 与 Jumper 因蛋白质结构预测共同获得当年诺贝尔化学奖的一半奖项，Baker 因计算蛋白质设计获得另一半。
2025 年 02 月	Google Research	推出基于 Gemini 2.0 的 AI co-scientist，通过多 Agent 生成、评议和优化科研假设，并在药物重定位等方向开展实验验证。
2025 年 05 月	Google DeepMind	发布 AlphaEvolve，将大模型生成代码与自动评估、进化搜索结合，用于数学算法发现及数据中心、芯片设计和模型训练优化。
2026 年 07 月	晶泰控股	推出 XtalPi Science 及 Genius Agent 矩阵，整合科学模型、专业工具与机器人实验，形成研究目标拆解、实验执行和数据反馈体系。
2026 年 08 月	Anthropic	披露 Claude Science 蛋白设计实验：调用专业模型完成设计与筛选，1,320 个设计产生 354 个实验命中，覆盖 15 个披露靶点中的 14 个。
2026 年 09 月	OpenAI	宣布 GPT-Rosalind 结束研究预览，通过可信访问计划向全球符合条件的机构开放，覆盖 ChatGPT Enterprise、Codex 及 API 等科研入口。
2026 年 09 月	英矽智能	宣布 AI 参与靶点发现与分子设计的 Rentosertib 在中国 III 期临床试验完成首例患者给药，研发成果进入更大规模临床验证。

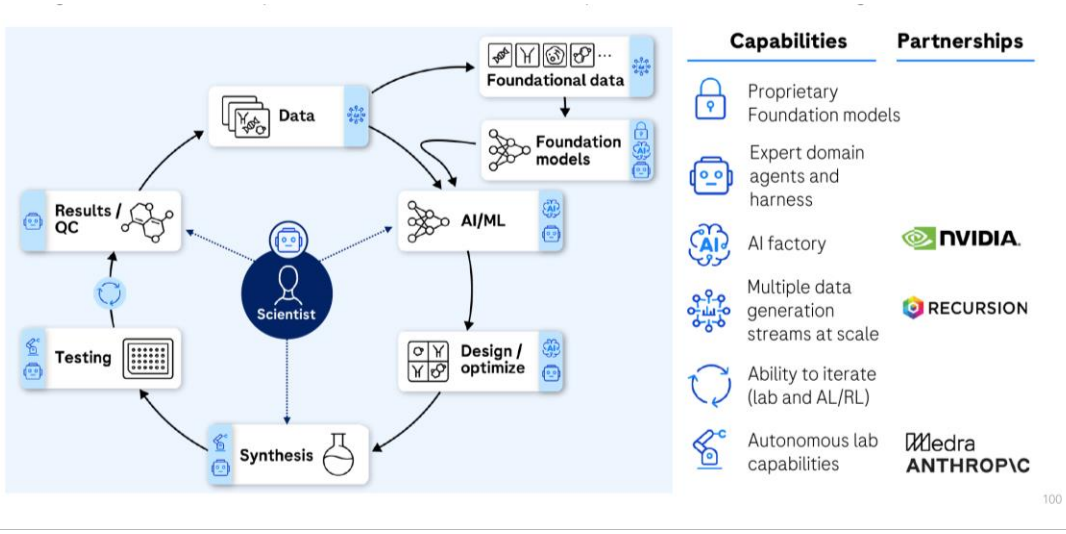
下载日志已记录，仅供内部参考

2026 年 09 月	Anthropic、诺和诺德	宣布合作，将 Claude 及 Claude Science 用于药物发现中的科学推理与研发工作流，并加强 AI 辅助软件开发。
2026 年 09 月	礼来、金斯瑞、Twist	TuneLab 引入金斯瑞湿实验服务与 Twist 抗体表征数据，将 AI 预测、候选筛选与实验数据生成进一步连接。
2026 年 09 月	NVIDIA 及研究合作方	发布 Proteina-Complexa 配套湿实验预印本，系统披露超过 100 万个候选结合物的验证结果，研究整体覆盖 133 个靶点。
2026 年 09 月	罗氏	在 Pharma Day 披露 Lab-in-the-Loop 体系及自主实验室建设进展；2025 年四季度至 2026 年二季度，40%的研发管线决策记录了 AI 或计算方法的贡献。
2026 年 09 月	Microsoft Research	发布实验性研究系统 Quine，将生物学多模态世界模型与科研编排框架结合，连接专业工具、文献与湿实验反馈，并开放研究者参与计划。

资料来源：Google DeepMind，诺贝尔奖官网，Microsoft Research，中信建投

科研 Agent 连接专业模型与实验反馈，AI4S 逐步形成完整研发体系。早期模型主要解决结构预测、活性评分等局部问题，跨工具衔接与实验决策仍由科学家逐步完成。近期系统则由通用模型拆解目标、调用专业算法并分析结果，再将候选交由真实实验检验；新增数据用于修正假设、调整下一轮设计，并支持专业模型优化。与 AI Coding 可快速运行测试不同，科研反馈涉及样本、设备、实验周期和数据标准，物理验证环节决定了迭代速度。随着计算端能够提出更多候选，高通量合成、检测与自动化实验的重要性相应提升，产业竞争也扩展到科学数据、专业工具、任务编排与实验设施的协同能力。

图表 73：罗氏以模型、数据、科研 Agent 与自主实验设施构建研发反馈体系



资料来源：罗氏，中信建投

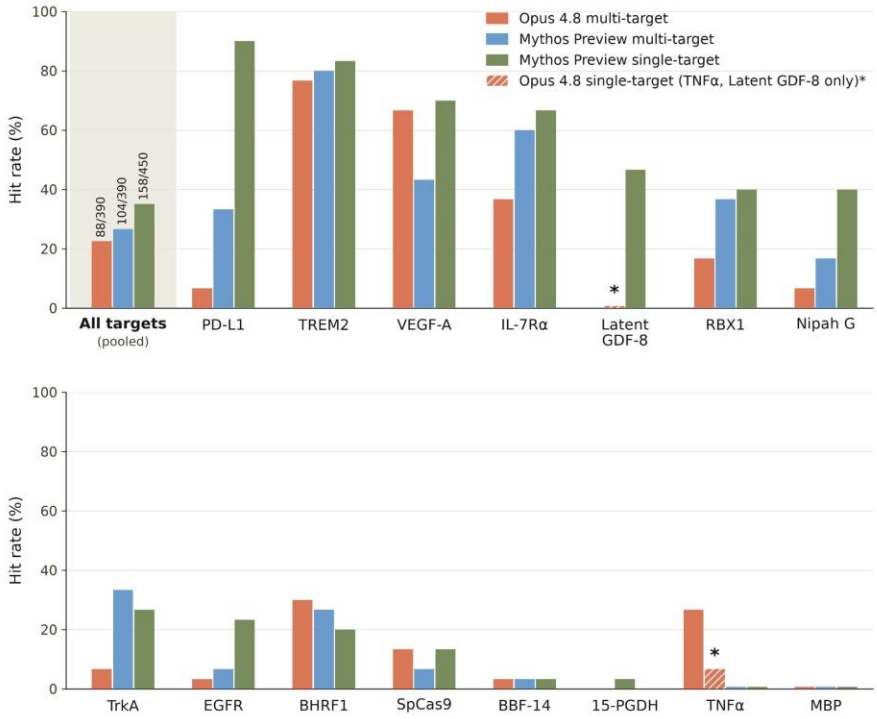
2.3.2 AI4S 近期产业进展

Claude Science 完成蛋白设计的跨工具任务验证，通用 Agent 开始承担连续科研工作。8 月 18 日，Anthropic 披露 Claude 调用现有专业蛋白模型完成设计、计算筛选与多轮优化，再交由 Adaptyv Bio 及 Twist 进行独立实验验证。在披露的 15 个靶点中，14 个获得有效结合物，1,320 个设计产生 354 个实验命中；不同运行配置下，单个设计的命中率为 22%-35%。相关任务在 24-48 小时连续运行中完成，表明模型能够组织多项专业计算并交付可测试结果。科研 Agent 的增量价值体现在任务规划、工具衔接和迭代效率上，为科学家同时推进更多候选方案提供了新的执行方式。

下载日志已记录，仅供内部参考

诺和诺德合作与生命科学验证计划同步推进，Anthropic 将科研能力接入真实药物研发流程。9 月 16 日，诺和诺德宣布与 Anthropic 合作，围绕具体药物发现问题测试 Claude 及 Claude Science，同时加强 AI 辅助软件开发。9 月 17 日推出的 Life Sciences Verification Program 面向科研机构、初创企业和药企开放申请，经研究资质、安全标准及伦理监督审核后提供适配生命科学任务的模型访问，早期阶段已接入数十家机构。9 月 23 日，Anthropic 又披露 Claude 参与新型酶系统研究，计算阶段使用约 950 个 Agent 连续工作 21 小时、消耗约 2.1 亿 Token，并由研究人员开展实验验证。企业合作、受控访问与自建研究能力结合，推动科研 Agent 由单次展示向可持续使用的机构工作流发展。

图表 74：Claude 在不同配置下完成蛋白设计并获得湿实验命中结果

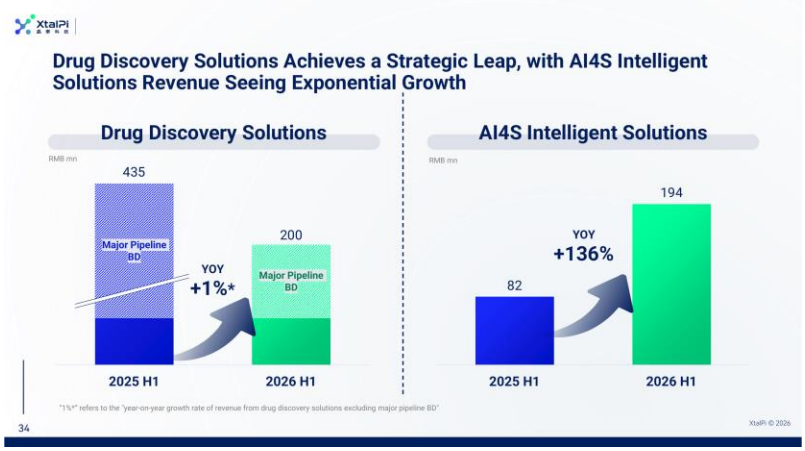


资料来源：Anthropic，中信建投

晶泰 AI4S 智慧解决方案收入翻倍，机器人实验室与科研服务需求开始形成规模。8 月 19 日披露的中期业绩显示，晶泰控股 2026 年上半年收入 3.936 亿元，受上年大额管线授权首付款形成的高基数影响，同比下降 23.9%；按公司剔除重大管线 BD 影响的口径，收入同比增长 73.8%。其中，AI4S 智慧解决方案收入 1.935 亿元，同比增长 136.4%，占总收入约 49.2%。公司已向礼来交付化合物存储管理系统，与 JW Pharmaceutical 合作的自主合成及工艺研发系统于 4 月完成交付，智能合成工作站已复制部署至 13 家客户。实验设备与科研服务增长，使平台收入来源由药物项目交付扩展至客户研发基础设施建设，并向钙钛矿、锂电池和分子筛等材料场景延伸。

下载日志已记录，仅供内部参考

图表 75：晶泰 AI4S 智慧解决方案上半年收入增长 136.4%



资料来源：晶泰控股，中信建投

Genius Agent 与自主实验平台协同，晶泰将专业模型能力转化为可执行的科研任务。公司 7 月推出 XtalPi Science 及 Genius Agent 矩阵，并在中期材料中展示由 Agent 调度、Scientific AI、Physical AI 和高质量数据构成的四层架构。Agent 负责拆解研究目标、调用专业模型与工具，并连接机器人实验执行；实验数据回流后支持预测和流程优化。公司披露的 Agentic HTE 案例将反应条件优化周期由 3-4 周缩短至 6 天，同时保留关键节点的化学家确认。9 月 17 日，旗下 LCC 与斯坦福大学签署药物发现合作，将疾病生物学成果接入 PACE 平台的设计、合成、测试与分析流程；9 月 22 日，晶泰披露与 Viva-Thera 合作在 7 个月内完成一项抗病毒小分子临床前候选物提名。平台向外部科研项目开放，使实验设施与算法能够在更多项目中复用。

图表 76：晶泰 Agentic HTE 将任务编排、机器人执行与人工确认结合

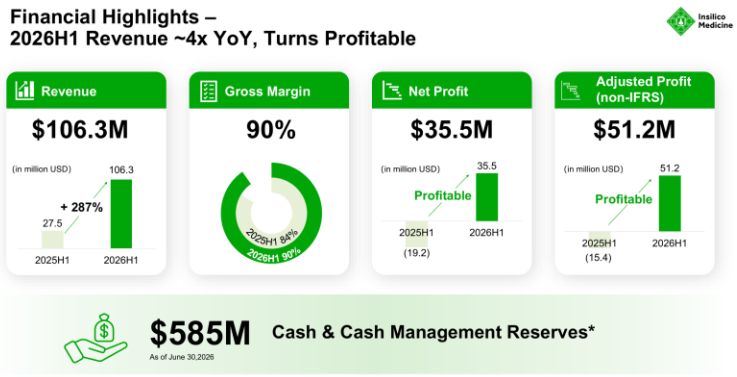


资料来源：晶泰控股，中信建投

收入与经营现金流同步改善，AI 研发平台的授权及合作成果进入财务兑现阶段。8 月 26 日披露的中期业绩显示，英矽智能 2026 年上半年收入 1.063 亿美元，同比增长 287.2%；IFRS 净利润 3553.7 万美元，经调整净利润 5122.7 万美元，经营活动净现金流转正至约 7900 万美元。药物发现及管线开发收入约 1.031 亿美元，占总收入约 97%，授权交易和研发合作是本期增长的重要来源。据业绩演示材料，截至 8 月 14 日，2026 年已宣布新增合同的潜在总价值约 73 亿美元，累计约 110 亿美元，相关对价包括首付款及后续研发、商业化里程碑。现金回款与项目推进为平台扩展提供资源，也使研发效率的评价由模型指标延伸至合作交付和药物资产价值。

下载日志已记录，仅供内部参考

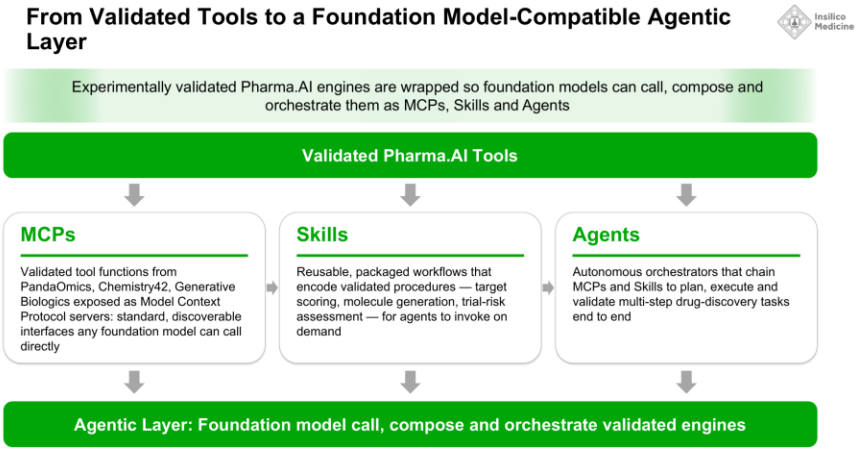
图表 77：英矽智能上半年收入增长 287%并实现盈利



资料来源：英矽智能，中信建投

Pharma.AI 向通用模型开放专业工具，英矽智能将科研 Agent 能力与临床管线并行推进。9 月 24 日，公司预告秋季 Pharma.AI 更新，通过 MCP、Skills 与 Agents 连接 PandaOmics、Chemistry42 等既有引擎，使通用模型能够调用专业工具、组合多步骤任务并输出分析结果。业绩材料展示了与 Claude Code、Codex、Cursor 等入口的衔接，以及由专业工具到标准接口、可复用工作流和 Agent 编排的分层架构。临床方面，9 月 10 日，公司宣布 AI 发现靶点并设计分子的 Rentosertib 在中国 III 期试验完成首例患者给药。工具平台对外开放扩大研发人员的使用入口，内部及合作管线则通过临床研究检验药物研发成果，两条路径共同增加平台积累专业数据和重复使用能力的机会。

图表 78：Pharma.AI 通过 MCP、Skills 与 Agents 连接专业模型和通用科研入口

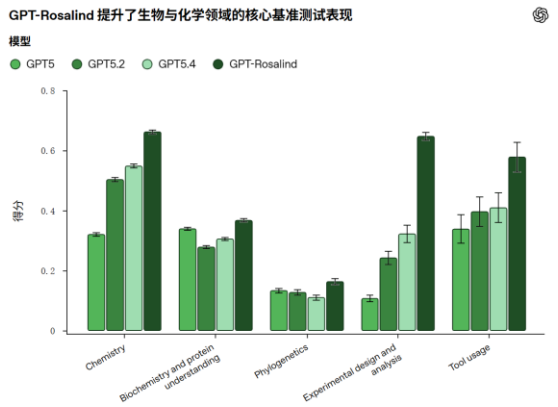


资料来源：英矽智能，中信建投

GPT-Rosalind 扩大机构开放范围，生命科学模型开始形成面向企业研发的收费服务。9 月 11 日，OpenAI 宣布 GPT-Rosalind 结束研究预览，通过可信访问计划向全球符合条件的机构开放，覆盖 ChatGPT Enterprise、Codex 及 API。模型将生物学推理与科学工具调用结合，用于文献证据梳理、药物化学、基因组学及实验规划，并可通过 Codex 中的生命科学研究插件连接专业工具与数据。10 月 5 日起，标准 API 输入、缓存输入和输出价格分别为每百万 Token 5 美元、0.5 美元和 25 美元。模型开放与计费机制落地，使专业科研任务能够进入企业软件和算力预算；机构身份审核、权限控制及内部数据管理则成为研发部署的重要配套。

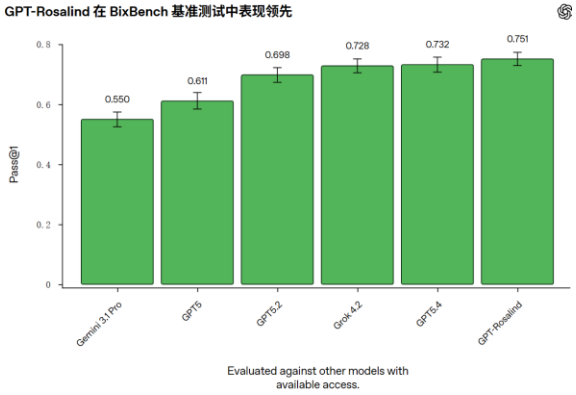
下载日志已记录，仅供内部参考

图表 79: GPT-Rosalind 提升生化领域的核心基准测试表现



资料来源：OpenAI，中信建投

图表 80: GPT-Rosalind 在 BixBench 基准测试中表现领先



资料来源：OpenAI，中信建投

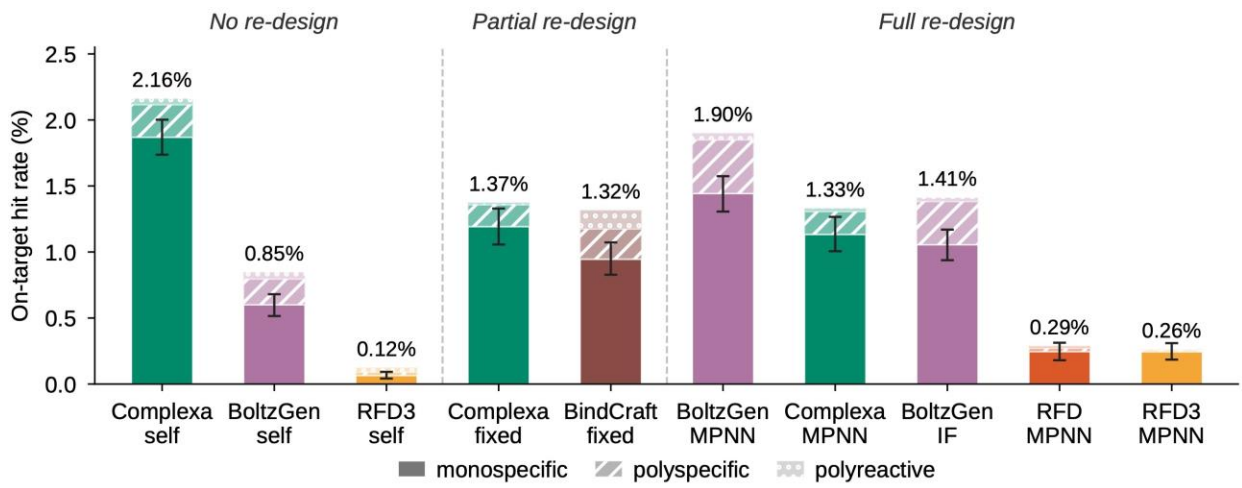
礼来 TuneLab 引入金斯瑞与 Twist，专业预测平台补齐高质量实验数据生成能力。9 月 16 日，金斯瑞宣布加入 Lilly TuneLab，为参与企业提供湿实验服务；Twist 同日宣布合作，提供抗体表征数据并支持 AbLab 等可开发性预测模型。TuneLab 于 2025 年推出，向生物技术企业开放基于礼来多年专有研发数据训练的模型，通过合作方贡献数据及联邦学习改进预测能力。本轮扩展将模型筛选与后续实验服务连接，帮助客户把计算候选转化为可比较的生物学证据。随着候选生成提速，蛋白表达、纯化、表征及标准化数据交付成为研发流程中更重要的环节，AI 模型与实验服务商的合作也更趋紧密。

字节 AI 制药业务独立融资，产业资本支持科研平台转向长期项目开发。9 月 16 日，据路透社报道，字节跳动拆分的 Anew Labs 完成 2.9 亿美元首轮外部融资，投后估值约 15 亿美元，由 HSG、IDG 资本及高瓴等领投，融资后字节仍持有约 56%股份。AI 药物研发涉及较长的实验、临床和合作周期，独立主体有利于形成适配研发业务的资本安排与组织机制。字节保留控股地位，同时引入外部投资者，体现头部互联网公司对 AI4S 的投入正在由内部探索转向专业平台建设；后续能力积累将更多依赖算法、实验数据与药物项目的共同推进。

英伟达公布百万级蛋白设计的湿实验论文，生成模型扩展开始对应更大规模的实验验证。9 月 18 日发布的 Proteina-Complexa 配套预印本系统披露大规模实验结果，整体研究覆盖 133 个靶点、超过 100 万个候选结合物，其中 127 个靶点组成统一筛选集合，115 个靶点获得靶向结合命中。该模型将生成式预训练与推理阶段搜索结合，探索通过增加计算提高候选质量。合作方 Manifold Bio 已于 3 月披露百万级筛选规模，本期新增的是系统化论文与验证结果。研究表明，计算端候选供给扩张需要高通量实验承接，模型优劣也开始通过实验命中、选择性和可重复性等指标评价，湿实验数据对下一轮模型训练的价值随之上升。

下载日志已记录，仅供内部参考

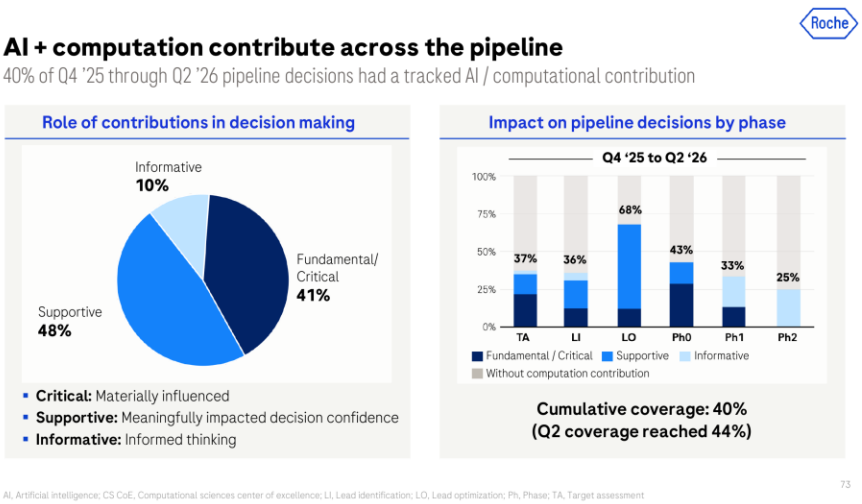
图表 81：Proteina-Complexa 以湿实验结果比较不同蛋白设计方法



资料来源：NVIDIA，中信建投

罗氏推进自主实验室建设，AI 已参与研发管线决策与资源配置。9 月 28 日，罗氏在 Pharma Day 披露 Lab-in-the-Loop 体系，将数据生成、基础模型、科研 Agent 与实验设施纳入连续迭代。2025 年四季度至 2026 年二季度，40%的研发管线决策记录了 AI 或计算方法的贡献，二季度覆盖率达到 44%；公司预计 TargetNexus 在 2026 年四季度参与 80%的研究组合决策。该 Agent 整合靶点生物学、药物发现、临床与组合信息，为研究优先级提供支持。罗氏同时推进 Autonomous Lab Accelerator，逐步连接实验执行、分析与报告。大型药企从引入单项工具走向重构研发流程，有助于将模型调用、自动化设备和实验数据投入转化为持续性的研发能力建设。

图表 82：罗氏 40%的阶段性研发管线决策记录了 AI 或计算方法贡献



资料来源：罗氏，中信建投

微软 Quine 结合生物学世界模型与科研编排框架，探索跨尺度推理和实验反馈。9 月 29 日，Microsoft Research 发布实验性研究系统 Quine，将基因组、蛋白质、化学、RNA、细胞状态与生物成像等信息联合建模，并通过 harness 连接推理模型、文献、科学工具与研究人员。在与 Broad Institute 开展的胰腺癌研究中，系统对数千种化合物进行预测排序，优先候选随后获得湿实验验证；从缩小搜索范围到形成待验证候选清单的计算筛

下载日志已记录，仅供内部参考

选在一个周末内完成。微软同步开放 Quine Fellows 项目，为研究者提供模型与计算资源。其技术路径强调不同模态之间的信息关联，实验结果既检验候选，也用于改进后续预测和任务流程。

Isomorphic Labs 强化模型、数据与实验一体化，AI 药物设计向完整研发体系推进。9 月 29 日，公司披露其研发路径，将生物与物理环境建模、前沿推理和生成式设计 Agent 纳入统一平台，覆盖结构、结合、分子性质与毒性等问题，并同步建设数据生成和物理实验设施。公司此前于 5 月完成 21 亿美元 B 轮融资，计划用于扩展 IsoDDE 药物设计引擎、全球业务及临床方向的候选管线。与主要提供通用模型访问的厂商相比，Isomorphic 更强调在同一主体内连接模型研发、药物设计和临床推进，将算法改进直接应用于自有及合作项目。这一路径也提高了对专有实验数据、跨学科团队和长期资本的要求。

三、行情回顾

3.1 海外软件与 A 股 AI 应用板块表现复盘

3.1.1 IGV 指数表现复盘

2026 年，AI 软件行情由替代担忧转向商业化验证，市场从关注模型能力转向关注软件企业 AI 变现能力。

1）1 月至 4 月上旬：IGV 由 1 月 7 日的 106.24 美元回落至 4 月 10 日的 74.67 美元，回撤 29.7%。Anthropic 推出 Cowork 及企业工作流插件，Agent 逐步进入销售、法律、财务和数据分析等场景，引发市场对传统 SaaS 席位需求、产品差异化及定价能力的担忧。AI 替代部分人工操作、通用入口整合软件功能的预期，推动市场下调软件企业长期增长预期，板块估值承压。

2）4 月中旬至 6 月初：IGV 由 4 月 10 日的 74.67 美元升至 6 月 1 日的 107.70 美元，涨幅 44.2%。数据平台业绩改善为 AI 需求提供验证，推动软件板块修复。5 月 27 日，Snowflake 披露 FY2027 第一财季产品收入 13.3 亿美元，同比增长 34%，并上调全年产品收入指引。Agent 落地带来的数据调用、计算消耗和治理需求，使按使用量收费的数据平台具备较清晰的收入增长路径，市场开始重新评估基础设施型软件在 AI 产业链中的价值。

3）6 月初至 7 月底：IGV 由 6 月 1 日的 107.70 美元回落至 6 月 25 日的 84.76 美元，跌幅 21.3%，随后震荡修复，7 月底收于 94.58 美元。前期快速上涨后，科技股整体调整与风险偏好回落对软件估值形成压制，市场对 AI 投入回报和收入兑现节奏的分歧扩大。软件企业的 AI 付费转化、存量客户扩张及利润率表现成为业绩观察重点，数据基础设施需求增长与应用端商业化进度之间的差异，推动板块内部重新定价。

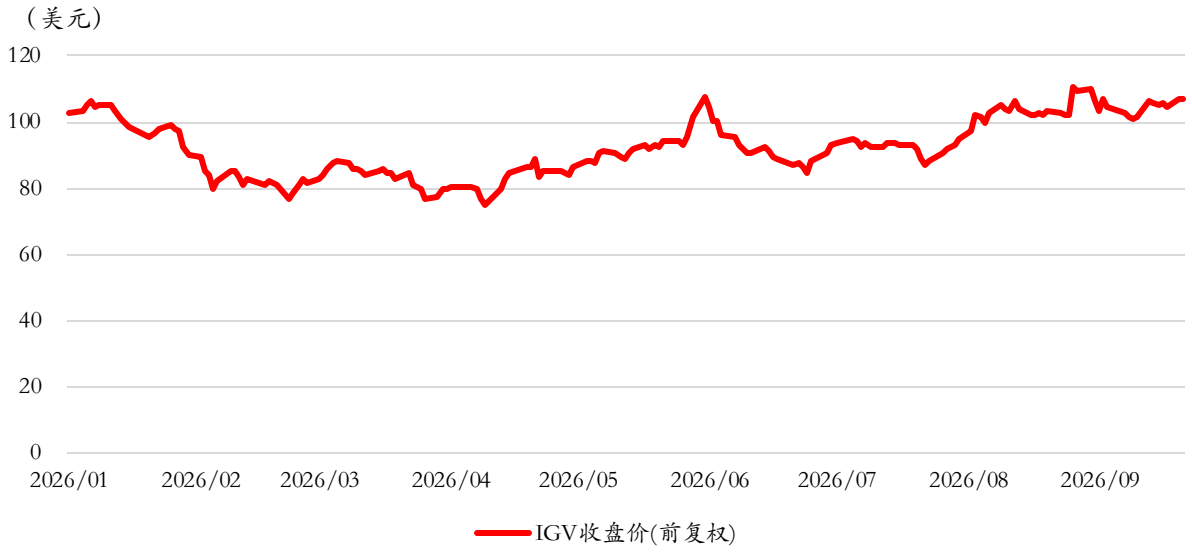
4）8 月：IGV 由 7 月底的 94.58 美元升至 8 月底的 109.98 美元，单月上涨 16.3%。应用端 AI 收入增长为板块修复提供新增证据。8 月 26 日，Salesforce 披露 FY2027 第二财季 Agentforce ARR 超过 15 亿美元，同比增长超过 240%（该季度统计范围纳入 Slackbot 及 Headless 360），并上调全年收入指引。成熟软件厂商依托客户基础、企业数据和业务流程拓展 AI 产品，强化了市场对软件平台通过 AI 交叉销售、产品升级和使用量增长实现增收的预期。

5）9 月：IGV 由 8 月底的 109.98 美元回落至 9 月 10 日的 101.20 美元，随后修复至 9 月 23 日的 108.10 美元，月末收于 106.48 美元，9 月下跌 3.2%。前期上涨后的估值消化与 AI 替代风险讨论带来波动，数据平台业绩增长则继续提供支撑。9 月 2 日，Snowflake 披露 FY2027 第二财季产品收入 14.9 亿美元，同比增长 37%，并将全年产品收入增长指引由 31%上调至 36%。9 月中旬，围绕前沿 AI 发展节奏的讨论带动科技板块内部轮动，

下载日志已记录，仅供内部参考

软件替代压力缓和的预期推动估值修复，IGV 于 9 月 14 日上涨 5.0%；下旬利率压力延续，软件估值再度消化，月末表现仍呈现结构性分化。

图表 83: IGV 指数 2026 年 1 月至今走势



资料来源：Wind，中信建投

3.1.2 申万计算机指数表现复盘

2026 年以来，申万计算机指数经历“国产 AI 催化上涨-模型冲击板块估值-AI 应用预期修复-业绩验证压力释放”的行情演绎，市场关注点逐步由主题交易转向商业化落地。

1）1 月上旬至中旬：申万计算机指数由 1 月 5 日的 5352.35 点升至 1 月 14 日的 6151.34 点，涨幅 14.9%。智谱、MiniMax 相继登陆港股，国产大模型产业的资本市场关注度提升，AI 应用成为年初科技行情的重要方向。办公、企业服务及行业智能体等应用场景的商业化预期升温，带动市场对软件企业产品升级和新增需求的预期，板块快速上涨。

2）1 月中旬至 4 月上旬：指数由 6151.34 点震荡回落至 4 月 3 日的 4719.83 点，回撤 23.3%。年初 AI 应用主题快速上涨后，市场对商业化兑现节奏的分歧扩大。海外 Agent 向企业工作流延伸，使传统软件功能壁垒、开发交付模式及收费方式面临重新评估；国内软件企业的客户采购、项目交付和 AI 收入贡献也进入验证阶段。主题预期与当期经营表现之间的差距，带动相关板块估值调整。

3）4 月上旬至 5 月中旬：指数由 4 月 3 日的 4719.83 点回升至 5 月 13 日的 5461.19 点，涨幅 15.7%。国产模型迭代重新提振产业预期，DeepSeek 于 4 月 24 日发布 V4 预览版，模型能力提升及国产算力适配进展，带动市场关注推理算力、模型部署和行业应用机会。对于软件和 IT 服务企业，模型落地涉及的数据治理、系统集成及业务流程改造，为新增项目需求提供了空间，推动板块阶段性修复。

4）5 月中旬至 7 月中旬：指数由 5 月 13 日的 5461.19 点回落至 7 月 17 日的 4183.77 点，跌幅 23.4%。6 月以来海外科技资产波动向 A 股传导，科技成长板块风险偏好下降，资金对业绩确定性和估值水平的要求提高。

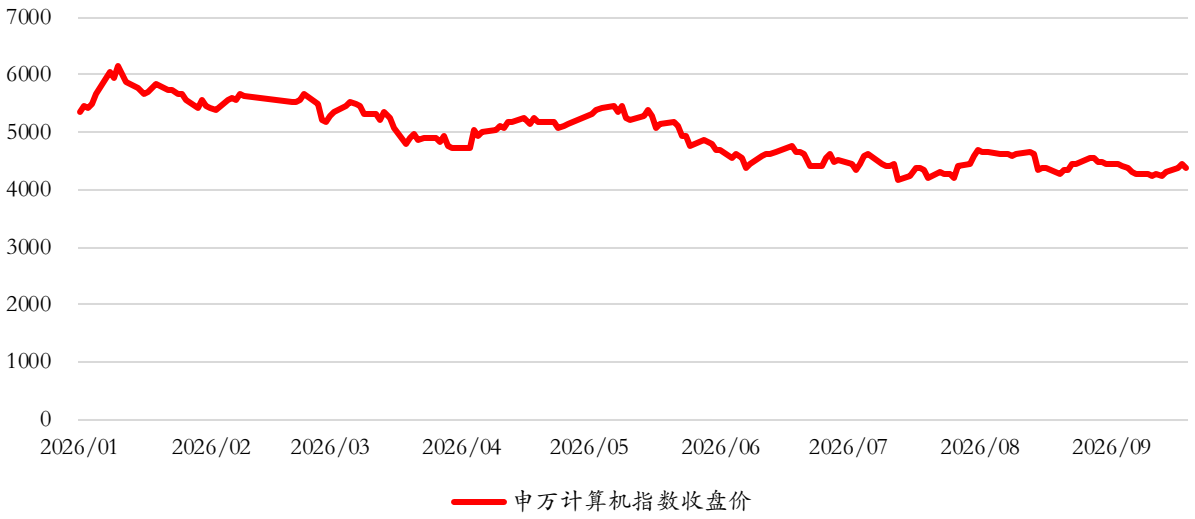
下载日志已记录，仅供内部参考

随着中报窗口临近，算力需求增长、软件订单转化及盈利兑现节奏成为比较重点，前期AI应用预期逐步在部分标的上得到经营数据检验，板块估值有所消化。

5）7月下旬至9月：指数由7月17日的4183.77点反弹至8月5日的4690.46点，涨幅12.1%，随后震荡回落，9月30日收于4192.48点，9月下跌7.7%。中报呈现总量盈利修复与企业间分化并存的特征，算力基础设施、数据服务与应用软件的增长节奏存在差异，业绩改善对板块的支撑集中于部分方向。政策端，工信部8月底启动人工智能应用服务商培育专项行动，9月发布“人工智能+软件”相关部署，围绕智能体软件、行业交付及应用服务形成催化。数据治理、模型集成和持续运营需求带来业务拓展空间，客户采购和收入兑现节奏则影响板块修复的持续性。

下载日志已记录，仅供内部参考

图表 84: 申万计算机指数 2026 年 1 月至今走势



资料来源：Wind，中信建投

3.1.3 申万传媒指数表现复盘

2026 年，AI 应用预期与内容产业周期共同驱动板块行情。

1）1 月上旬至中旬：申万传媒指数由 1 月 5 日的 853.48 点升至 1 月 14 日的 1021.75 点，涨幅 19.7%。AI 应用热度上升，生成式 AI 入口带来的流量迁移推动市场关注 GEO、AI 营销及内容生产。广告营销和数字内容企业的 AI 业务布局密集进入市场视野，投资者对营销效率提升、内容制作降本及商业模式拓展的预期增强，传媒板块表现出较高的主题弹性。

2）1 月中旬至 3 月下旬：指数由 1021.75 点震荡回落至 3 月 23 日的 754.36 点，跌幅 26.2%。前期 AI 营销、数字内容等方向快速上涨后，交易热度逐步回落，市场开始评估新业务收入规模及利润贡献。3 月广告营销、游戏和影视等多个方向同步调整，板块经历较广泛的估值消化。AI 应用的用户增长、客户付费和实际经营贡献，逐步成为影响个股定价的重要因素。

3）3 月下旬至 5 月上旬：指数由 3 月 23 日的 754.36 点反弹至 5 月 7 日的 836.25 点，涨幅 10.9%。AI 视频生成及内容工具迭代，带动市场重新关注影视制作、广告素材和数字内容领域的应用空间。年报及一季报窗口叠加游戏新品、广告经营和内容供给预期，使板块交易重点由年初集中于 AI 概念，逐步延伸至产品周期与经营改善，推动阶段性修复。

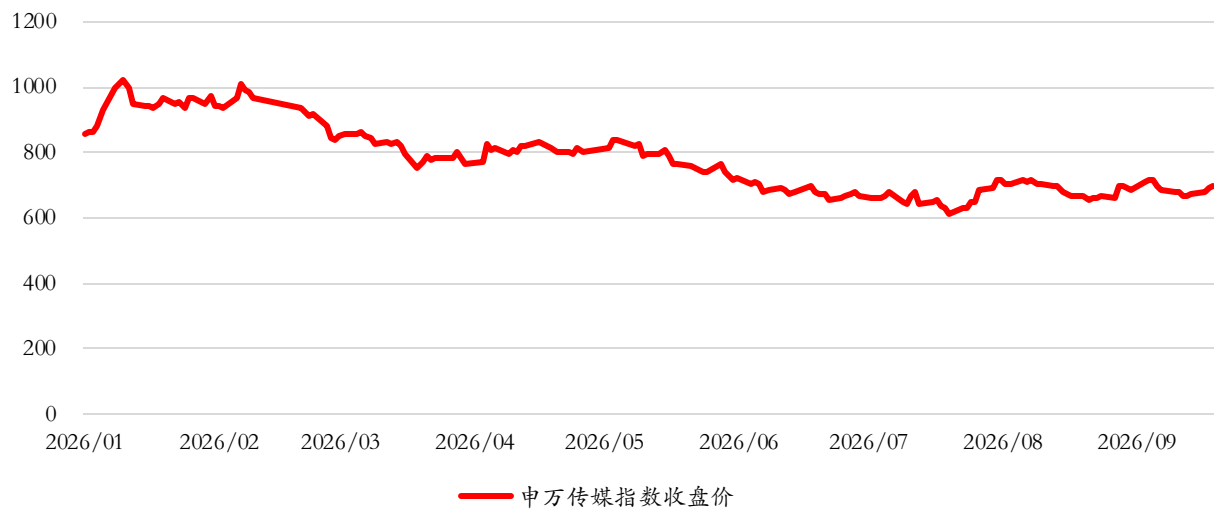
4）5 月上旬至 7 月下旬：指数由 5 月 7 日的 836.25 点回落至 7 月 24 日的 608.60 点，跌幅 27.2%。科技成长板块整体波动加大，AI 应用主题的风险偏好回落，传媒板块估值承压。行业内部，游戏产品周期、广告客户投放及影视内容表现各有差异，市场对收入增长持续性和利润兑现的要求提高，内容产品表现及商业化进展对个股走势的影响增强。

5）7 月下旬至 9 月至今：指数由 7 月 24 日的 608.60 点反弹至 8 月 5 日的 717.22 点，涨幅 17.8%，8 月底

下载日志已记录，仅供内部参考

收于 691.69 点。9 月 8 日回升至 721.62 点后再度调整，月末收于 664.55 点，9 月下跌 3.9%，8-9 月累计下跌 3.2%。游戏景气与 AI 影视落地提供阶段性催化：8 月国内游戏市场实际销售收入同比增长 8.08%，自研游戏海外收入增长 20.87%，支撑新品周期和出海需求；《后西游记》播出，将 AIGC 应用拓展至长剧生产与发行，提升市场对 AI 内容商业化、IP 开发效率和成本优化的关注。月末板块回落，市场继续围绕产品表现及商业化贡献定价。

图表 85: 申万传媒指数 2026 年 1 月至 9 月走势



资料来源：Wind，中信建投

3.2 海外软件与 A 股 AI 应用板块估值复盘

3.2.1 IGV 指数估值复盘

2023 年以来，随着 ChatGPT 引发生成式 AI 浪潮，海外软件板块迎来估值重塑，IGV 指数 PS-TTM 整体呈现先升后降趋势。

1）2023 年初至 2025 年中，生成式 AI 带动市场上调软件企业长期增长预期，IGV 综合 PS-TTM 由 2023 年 1 月 3 日的 6.39 倍升至 2025 年 7 月 3 日的 11.06 倍，累计提升 73.1%。云端数据处理、企业工作流改造及 AI 产品商业化预期共同推动估值扩张，市场逐步给予具备数据基础、客户入口和 AI 应用能力的软件平台更高的收入估值。

2）2025 年下半年至 2026 年 6 月，软件企业 AI 变现节奏与 Agent 替代风险成为估值定价的重要因素。通用模型向企业工作流延伸，市场重新评估传统软件的席位需求、产品壁垒及收费模式，IGV 综合 PS-TTM 震荡回落，于 2026 年 6 月 25 日降至 6.16 倍，较 2025 年 7 月高点下降 44.3%。

3）2026 年 6 月低点后，IGV 综合 PS-TTM 逐步修复，8 月底达到 9.37 倍，略高于 1 月 2 日的 9.28 倍；9 月 30 日为 8.59 倍，较 6 月低点提升 39.4%，较 1 月 2 日低 7.4%。软件企业 AI 产品收入和数据平台需求增长为估值提供支撑，板块已修复至年初附近。

下载日志已记录，仅供内部参考

当前估值支撑主要集中于网络安全和部分 AI 平台公司，成熟软件平台更多为低点以来的估值修复。9 月 30 日，Palo Alto Networks、CrowdStrike、Fortinet 和 Datadog 的 PS-TTM 分别为 28.31、50.24、17.42 和 24.79 倍，均高于年初；Palantir 为 73.02 倍，较 6 月低点明显修复，但仍低于年初的 102.69 倍。上述五家公司合计期末代理权重 34.15%，构成组合高估值的重要支撑。基本面方面，CrowdStrike FY2027 第二财季末 ARR 同比增长 25%至 58.4 亿美元，企业安全需求增长为相关公司的估值溢价提供了经营支撑。

企业软件平台方面，Salesforce、微软和 ServiceNow 的 PS-TTM 分别由 6 月 25 日的 2.87、8.23 和 6.61 倍回升至 9 月 30 日的 4.30、11.48 和 9.41 倍，但仍较年初的 5.89、11.96 和 12.08 倍低 27.0%、4.1%和 22.1%。Salesforce 披露 Agentforce ARR 超过 15 亿美元，同比增长超过 240%（该季度统计范围纳入 Slackbot 及 Headless 360），强化了市场对存量客户 AI 升级及产品增收的预期；微软 Copilot 付费席位增长亦验证应用商业化。固定 10 月 1 日权重回溯共同正 PS 样本，Salesforce、微软、ServiceNow 和 Adobe 合计贡献约 50%的 6 月 25 日至 9 月 30 日 PS 升幅，安全及 AI 平台维持溢价、传统软件龙头估值修复的特征仍然明显。

图表 86: 2023 年至今 IGV 指数 PS-TTM 估值变化



资料来源：Wind，中信建投

2023 年以来，IGV 指数 PE-TTM 整体呈现先升后降、低位修复的走势，市场关注重点逐步由 AI 赋能软件行业的长期增长空间，转向 AI 商业化落地及软件企业盈利兑现。

1）2023 年初至 2024 年初，AI 赋能软件增长及经营效率提升的预期推动 IGV 综合 PE-TTM 扩张，由 2023 年 1 月 3 日的 40.69 倍升至 2024 年 2 月 9 日的 59.24 倍。此后估值震荡调整，软件企业的收入兑现、盈利增长及 AI 投入回报逐步成为市场定价重点。

2）2025 年下半年至 2026 年上半年，AI 替代风险讨论与科技板块风险偏好波动加大，IGV 综合 PE-TTM 回落至 2026 年 6 月 25 日的 27.89 倍，较 2024 年高点下降 52.9%。同期，成分股盈利变化及正利润样本范围调整也影响综合估值。

下载日志已记录，仅供内部参考

3）2026 年 6 月低点后，IGV 综合 PE-TTM 由 6 月低点的 27.89 倍修复至 8 月底的 40.52 倍，9 月 30 日为 38.11 倍，较 6 月低点提升 36.7%、较 1 月 2 日的 39.82 倍低 4.3%。微软、Salesforce 和 Adobe 的 PE-TTM 分别由 6 月 25 日的 20.93、15.33 和 10.64 倍回升至 28.48、19.55 和 12.82 倍，但仍较年初的 33.50、32.91 和 19.57 倍低 15.0%、40.6%和 34.5%。大权重软件公司的估值修复为组合提供支撑，其中微软恢复程度相对较高，Salesforce 和 Adobe 与年初仍有明显差距。

网络安全及 AI 平台的高估值权重，对 IGV 整体估值形成明显支撑。剔除 Palo Alto Networks、Palantir、CrowdStrike、Fortinet 和 Datadog 并重新归一化后，IGV 9 月 30 日综合 PE 由 38.11 倍降至 24.92 倍。结合微软、Salesforce 和 Adobe 的 PE 仍低于年初，IGV 整体估值接近年初，反映出网络安全及部分 AI 平台维持较高估值、传统软件企业修复程度不一的结构性分化。

图表 87: 2023 年至今 IGV 指数 PE-TTM 估值变化



资料来源：Wind，中信建投

3.1.2 申万 TMT 板块估值复盘

ChatGPT 于 2022 年底发布后，生成式 AI 成为全球科技产业核心投资方向，A 股市场围绕 AI 应用、AIGC 及国产替代展开估值重估。

1）2023 年，计算机板块受 AI 软件、云计算及国产化需求推动，市场对 AI 赋能软件产业的长期成长空间形成较强预期，PE-TTM 由年初的 50.57 倍升至 6 月 20 日的 78.52 倍；传媒板块受 AIGC 内容生成、游戏创新及内容生产效率提升预期催化，PE-TTM 由 32.67 倍升至同期的 53.86 倍。市场主要围绕人工智能技术突破带来的增量需求和商业模式变化展开定价，软件企业的客户入口、数据积累及传媒公司的内容资产成为估值重估的重要依据。

2）进入 2024 年，AI 产业逐步进入商业化验证阶段，市场更加关注应用落地速度、客户付费意愿及企业盈利兑现。计算机、传媒板块的 AI 业务仍处于产品探索和客户导入阶段，收入贡献与前期增长预期之间存在时间

下载日志已记录，仅供内部参考

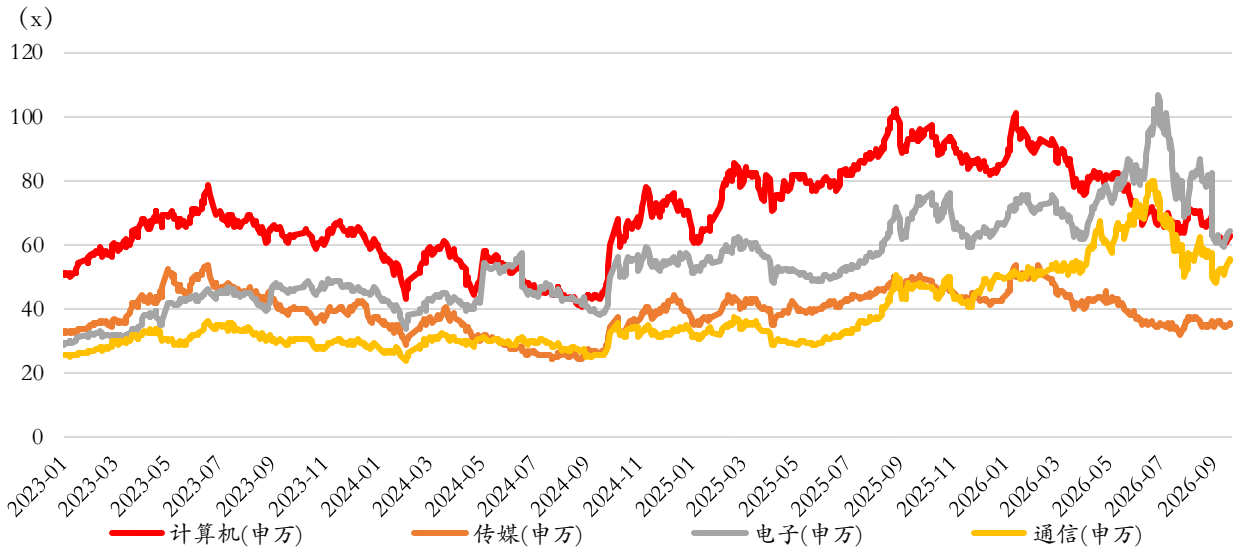
差，估值经历明显调整，计算机 PE-TTM 于 8 月降至 40.37 倍，传媒于 7 月降至 24.40 倍。与此同时，全球 AI 基础设施建设推进，算力芯片、服务器、光模块及高速互联需求逐步体现，市场对硬件产业链的关注提升。四季度 TMT 板块估值普遍修复，计算机、传媒年底 PE 分别回升至 67.45 倍和 37.90 倍，全年呈现先调整、后修复的走势。

3）2025 年至 2026 年上半年，AI 基础设施投入与国产模型、应用创新共同驱动 TMT 行情，板块估值扩张节奏逐步分化。计算机、传媒在 2025 年仍经历明显的估值上行，PE-TTM 年度高点分别达到 102.14 倍和 50.59 倍；进入 2026 年后，市场对应用端收入增长和盈利兑现的要求提高，两板块估值逐步回落。电子、通信则受益于 AI 服务器、半导体、光模块及高速互联需求增长，估值扩张延续至年中，电子 PE-TTM 于 6 月 30 日达到 106.44 倍，通信于 6 月 25 日达到 79.92 倍。算力产业链较明确的订单需求及业绩增长，为市场上调增长预期提供了支撑。

4）2026 年 7 月至 9 月，TMT 板块经历估值消化与结构性修复。电子 PE-TTM 由 6 月底的 106.44 倍降至 7 月底的 71.03 倍，8 月底回升至 82.19 倍，9 月 30 日为 58.55 倍；通信由 6 月底的 76.15 倍回落至 7 月底的 51.93 倍，最新为 48.82 倍。前期较高估值、科技板块风险偏好波动及中报盈利更新共同影响 PE 走势。其中，电子 9 月 PE 下降 28.8%，PS 同期下降 7.7%，估值变化亦受到盈利分母及样本变化影响。计算机、传媒最新 PE 分别为 59.69 倍和 34.10 倍，较 2025 年末下降 30.4%和 21.5%，应用端整体估值恢复仍相对有限。

与海外软件相比，A 股计算机、传媒估值距离年初仍有较大差距。IGV 于 8 月底已基本恢复至年初水平，9 月 30 日 PE-TTM 为 38.11 倍，较 2025 年末下降 6.2%，降幅明显小于申万计算机和传媒。海外网络安全及部分 AI 平台维持较高估值，微软、Salesforce 等软件龙头的 AI 收入与客户付费进展也为修复提供支撑。国内计算机企业的 AI 收入兑现更多涉及客户预算、项目实施、验收及回款，传媒则同时受到游戏产品周期、广告投放和内容表现影响，行业内部经营改善的节奏差异较大。海外软件估值修复提供了商业化参考，A 股应用端的估值回升仍主要围绕具备实际收入贡献和盈利改善的公司展开。

图表 88: 申万 TMT 四大行业 PE 走势对比



资料来源：Wind，中信建投

ChatGPT 发布后，生成式 AI 成为市场重要交易方向，A 股 TMT 板块围绕 AI 应用、内容生产及算力基础

下载日志已记录，仅供内部参考

设施展开估值重估，随着产业落地推进，各环节收入增长预期和估值表现逐步分化。

1）2023 年，计算机、传媒板块受 AI 应用、AIGC 及软件生态发展预期推动，PS-TTM 估值率先提升。计算机、传媒 PS 分别由年初的 2.67 倍和 2.46 倍升至 6 月 20 日的 3.77 倍和 3.89 倍。市场主要交易 AI 技术变革带来的长期成长空间，软件及内容应用方向成为早期资金关注重点。企业客户、数据资源、内容版权及流量入口的潜在商业价值受到重估，市场对 AI 提升软件付费能力、降低内容生产成本和拓展应用场景形成较强预期。同期，电子、通信估值亦有提升，但 AI 基础设施需求向产业链收入传导的过程还在起步阶段。

2）2024 年至 2025 年，AI 应用商业化验证与算力基础设施需求增长共同影响板块估值。计算机、传媒在 2024 年经历调整与修复，年底 PS 分别为 3.25 倍和 2.84 倍；2025 年，国产模型进展及应用创新再次带动估值扩张，年度高点分别达到 4.03 倍和 3.68 倍。与此同时，全球云厂商加大 AI 资本开支，服务器、半导体、光模块及高速互联需求增长，推动市场提高对硬件产业链收入规模和增长持续性的预期。至 2025 年底，电子、通信 PS 分别升至 3.66 倍和 2.36 倍，算力基础设施成为 TMT 估值扩张的重要方向。

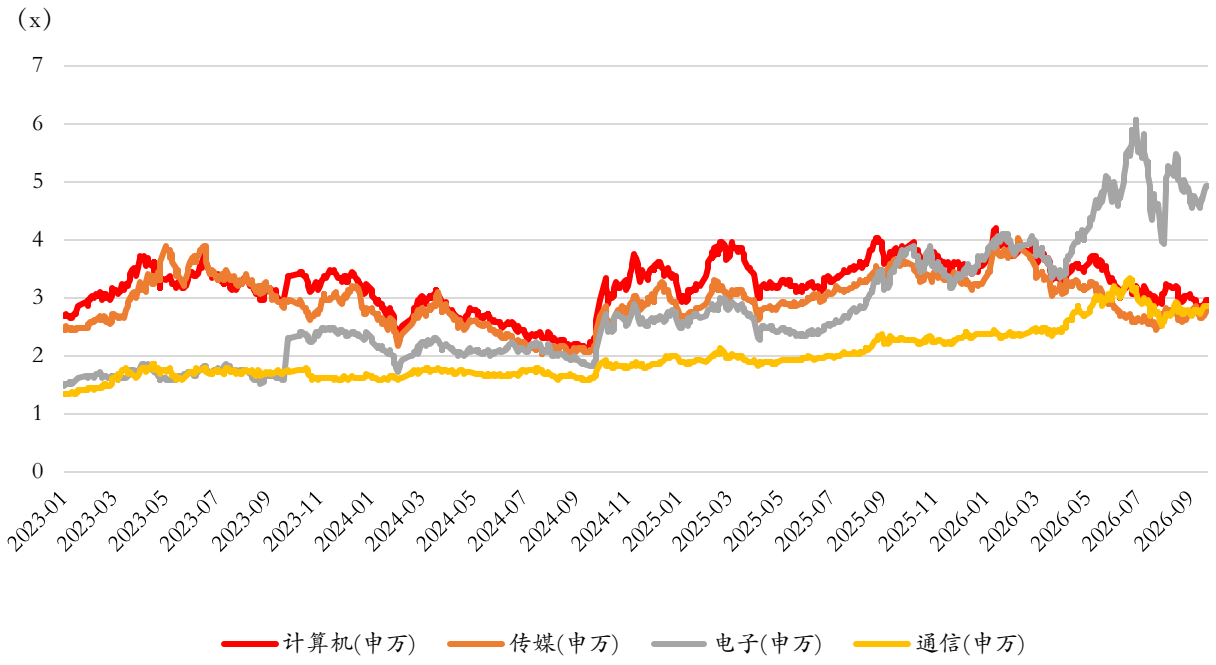
3）2026 年上半年，硬件产业链与应用端估值走势进一步分化。电子板块受 AI 服务器、半导体及相关配套需求推动，PS 于 6 月 30 日达到 6.06 倍；通信板块受光模块和高速互联需求支撑，PS 于 6 月 22 日达到 3.33 倍。相比之下，计算机、传媒在年初 AI 应用行情后叠加模型吞噬软件的叙事下逐步进入预期消化阶段，6 月底 PS 分别回落至 3.16 倍和 2.59 倍。市场更加关注软件企业 AI 产品付费、项目订单转化，以及传媒企业游戏流水、广告收入和内容变现，应用端估值支撑逐步落实到具体业务的收入贡献。

4）2026 年 7 月至 9 月，电子、通信在 7 月调整后出现修复，9 月再度回落，计算机、传媒恢复程度相对有限。电子 PS 由 6 月底的 6.06 倍降至 7 月底的 4.09 倍，8 月底回升至 4.89 倍，9 月 30 日为 4.51 倍；通信由 6 月底的 3.17 倍降至 7 月底的 2.57 倍，8 月底回升至 2.77 倍，最新为 2.64 倍。前期快速上涨后的估值消化、市场风险偏好变化与产业需求验证交织，硬件板块波动较大，但最新 PS 仍分别较 2025 年末高 23.4%和 11.9%。计算机最新 PS 为 2.80 倍，较 2025 年末下降 21.6%；传媒为 2.68 倍，较 2025 年末下降 18.6%。9 月两板块 PS 分别下降 7.6%和 1.6%，此前阶段性修复未延续至全月。

与 IGV 相比，A 股计算机、传媒的 PS 仍明显低于年初，估值修复相对滞后。IGV 于 8 月底达到 9.37 倍，接近 2025 年末的 9.55 倍，9 月 30 日为 8.59 倍，较 2025 年末下降 10.1%，降幅小于申万计算机和传媒。与前文 PE 估值分析类似，海外网络安全及部分 AI 平台维持较高收入估值，微软、Salesforce 等软件企业的 AI 付费和产品增收进展，也为修复提供了经营依据。国内计算机企业的收入兑现更多涉及客户预算、项目交付及验收，传媒企业则受到游戏产品周期、广告投放和内容表现影响，AI 业务对收入的拉动程度存在较大差异。海外软件估值较年中低点明显修复，A 股应用端仍主要围绕商业化进展较快、收入增长较明确的方向展开。

下载日志已记录，仅供内部参考

图表 89: 申万 TMT 四大行业 PS 走势对比



资料来源：Wind，中信建投

ChatGPT 于 2022 年底发布后，生成式 AI 成为全球科技产业重要投资方向，A 股市场围绕 AI 应用、大模型及国产替代展开交易。2023 年以来，TMT 板块成交活跃度整体提升，市场关注点由早期的软件和内容应用预期，逐步扩展至算力基础设施需求及各环节的业绩兑现，板块内部交易重心有所变化。

1）2023 年，计算机、传媒等 AI 应用相关板块交易活跃，推动 TMT 成交额占比快速提升。TMT 成交额占比由 1 月 3 日的 20.1%升至 6 月 13 日的 42.5%，全年日均占比为 30.9%。市场主要围绕 AI 技术突破及产业发展空间展开交易，软件、游戏、内容生成及广告营销等方向受到集中关注，企业客户、数据资源和内容资产的潜在价值成为重要交易线索。

2）2024 年，AI 产业逐步进入商业化验证阶段，市场更加关注应用落地速度、客户付费及企业盈利兑现，TMT 交易热度随产业催化和市场风险偏好变化而波动，全年成交额日均占比为 30.8%。计算机、传媒的 AI 业务仍处于产品探索和客户导入阶段，主题催化与经营兑现之间存在时间差；与此同时，AI 基础设施建设带动服务器、光模块及半导体需求，电子、通信在 TMT 交易中的重要性逐步提升。2025 年，国产模型进展、AI 应用创新与全球算力投资共同带动科技板块活跃，TMT 成交额日均占比升至 32.7%。

3）2026 年以来，TMT 交易热度进一步上升，成交额日均占比由 1 月的 34.6%升至 6 月的 45.7%和 7 月的 46.9%，7 月 10 日达到 52.1%的阶段高点。AI 服务器、半导体、光模块及高速互联需求增长，使硬件产业链成为交易重点。从结构看，电子成交额占沪深市场的日均比例由 1 月的 17.2%升至 7 月的 31.7%，通信由 5.6%升至 7.8%；同期计算机由 7.9%降至 5.6%，传媒由 3.9%降至 1.8%。TMT 整体交易热度上升主要由电子、通信带动，应用端的成交占比相对回落。

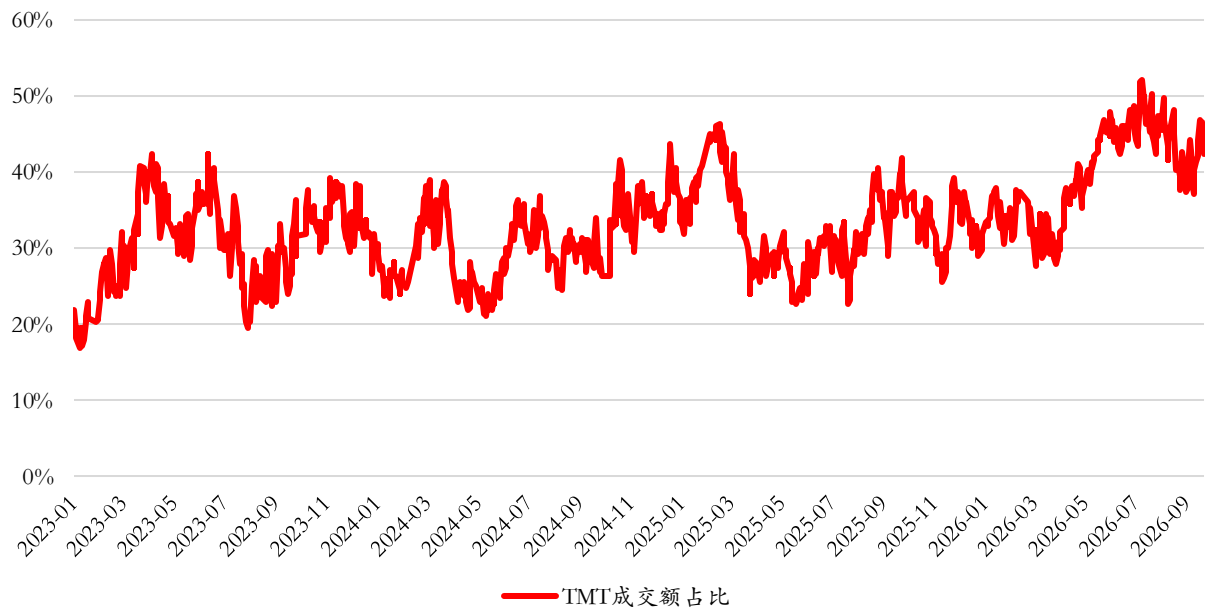
4）8 月至 9 月，TMT 交易热度较 7 月高位有所下降，但仍维持较高水平。8 月和 9 月 TMT 成交额日均占比分别为 43.9%和 41.5%，高于 2023-2025 年的年度日均水平。前期硬件行情快速上涨后，市场对资本开支持续

下载日志已记录，仅供内部参考

性、订单交付和盈利增长的预期反复调整，板块在回撤与修复中保持较高成交活跃度。9 月电子、通信成交额日均占比分别为 25.8%和 8.1%，合计 33.9%，仍占据 TMT 交易的主要部分，与两板块 PS 高于年初的估值表现相互呼应。

应用端交易活跃度的恢复呈现分化。传媒成交额日均占比由 8 月的 2.2%升至 9 月的 2.7%，AI 影视落地、游戏新品及出海经营表现带来阶段性关注；计算机则由 5.7%降至 4.9%，市场对软件订单、AI 付费及盈利兑现的要求仍较高。9 月 30 日 TMT 当日成交额占比为 38.3%，其中电子、通信分别为 24.6%和 6.8%，计算机、传媒分别为 4.5%和 2.4%。结合估值表现，海外 IGV 仍接近年初估值，A 股计算机、传媒的估值与成交占比仍低于年初，国内应用端的市场关注更多集中于具体产品和业绩催化，硬件产业链仍是 TMT 交易的核心方向。

图表 90: 2023 年至今 TMT 板块成交额占 A 股成交额比例变化分析



资料来源：Wind，中信建投

3.3 八月至今 AI 板块表现分析

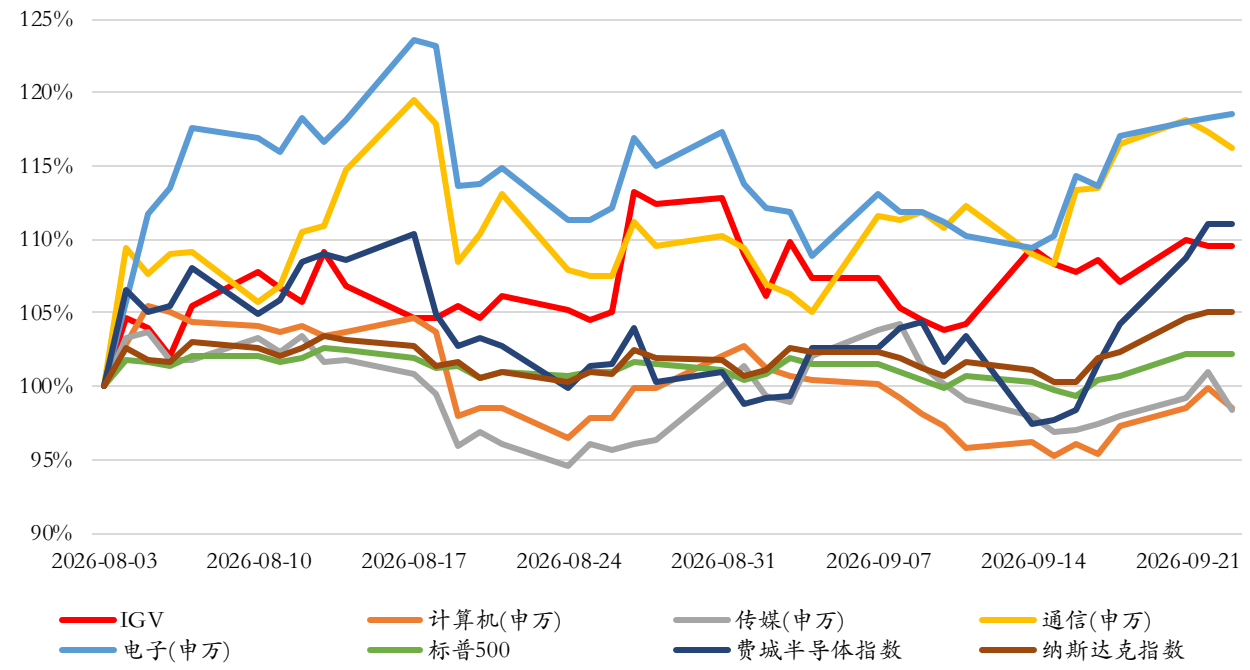
8-9 月，AI 产业链在利率变化与产业预期调整中呈现明显分化。以 7 月 31 日收盘为基期，截至 9 月 30 日，IGV 和费城半导体指数分别上涨 12.6%、11.6%，纳斯达克综合指数、标普 500 分别上涨 5.9%、2.2%。A 股通信、电子同期上涨 4.7%、3.2%，计算机、传媒分别下跌 5.2%、3.2%。硬件产业链与海外软件的修复力度相对更强。

海外市场的波动同时受到宏观利率和模型发展预期影响。通胀压力推动美联储于 9 月 16 日加息 25 个基点，贴现率与融资成本上升，使市场更加关注科技企业增长持续性和资本开支回报。9 月中旬，前沿模型开发节奏与安全风险讨论升温，算力需求预期受到扰动，部分软件公司则受益于替代压力缓和的预期。随后硬件板块再度修复，9 月以来费城半导体指数上涨 9.5%，IGV 下跌 3.2%，科技板块内部轮动较为明显。

下载日志已记录，仅供内部参考

A 股方面，电子、通信在 8 月上半月率先上涨，随后经历回撤，9 月中下旬重新走强。海外算力投入、国内模型迭代及国产算力建设继续支撑产业预期，市场围绕订单增长、交付能力和估值水平反复定价。计算机、传媒则以局部修复为主，海外应用端的积极变化尚未转化为国内板块的普遍上涨，个股表现更依赖产品落地和经营改善。

图表 91：8 月 AI 产业链相关指数日涨跌幅对比



资料来源：Wind，中信建投

海外 AI 交易的核心分歧，逐步集中于收入增长与资本投入能否持续匹配。模型厂商年化收入快速扩张后，高基数下的增速持续性、客户留存及推理成本成为影响远期预期的重要变量。对模型收入增速放缓的担忧，可能传导至云厂商资本开支及上游算力需求预期，加息则进一步提高市场对现金流和投资回报的要求。与此同时，近期软件财报提供了企业需求持续扩张的证据：Snowflake 于 9 月披露的季度净收入留存率为 126%，剩余履约义务达到 90 亿美元，同比增长 30%，显示存量客户扩张和合同储备仍有支撑。市场正在结合模型增长预期与企业实际需求，对 AI 产业链各环节重新定价。

模型能力突破与开发节奏放缓，对产业链不同环节产生不同影响。模型快速进步有助于扩大 AI 需求，同时加剧部分软件功能被替代的担忧；开发节奏放缓的预期，则可能短期压制算力扩张预期，并缓解传统软件的竞争压力。尽管行业围绕前沿开发安全性展开讨论，头部厂商仍在密集发布模型、优化成本和完善 Agent 工具，9 月 22 日 OpenAI 与 Anthropic 均推出新模型。技术供给持续改善，使市场在新增需求、软件替代和安全治理等不同逻辑之间反复定价，形成硬件与软件的阶段性轮动。

海外软件基础设施层的受益路径相对清晰，企业平台则通过与模型合作拓展业务空间。Agent 进入企业后，数据访问、检索计算、身份权限和安全治理需求随之增加，网络安全及数据平台能够承接新增工作负载。MongoDB 最新季度 Non-GAAP 经营利润率由上年同期的 15%升至 24%，自由现金流由 6990 万美元增至 1.38

下载日志已记录，仅供内部参考

亿美元，体现部分数据平台的增长已向盈利和现金流传导。企业平台方面，Salesforce 披露 Agentforce 与 Slack 在 FY2027 第二财季完成 32 亿个智能体工作单元，环比增长 97%，显示 Agent 实际使用规模扩大。其与 Anthropic 的合作进一步连接模型推理和企业数据、权限及业务流程，为软件平台通过业务执行和持续服务实现增收提供了路径。

映射到 A 股而言，海外软件行情的参考意义更多在于识别 AI 落地所需的环节。工信部 8 月底启动人工智能应用服务商培育专项行动，提出到 2026 年底全国服务商资源池内企业数量突破 2000 家、2027 年底不少于 3000 家，并鼓励扩大大模型、智能体及 Token 服务采购。政策支持与企业部署需求结合，有望带动数据治理、系统连接、安全保障和行业流程改造，传媒侧则围绕内容生产效率、营销效果及产品变现形成机会。硬件端较明确的需求传导与应用端分化的商业化进程，共同影响本轮 A 股算力链和应用板块的表现差异；客户采购、实际使用和利润贡献，将决定国内应用端估值修复的持续性。

下载日志已记录，仅供内部参考

四、投资建议

投资方向：继续看多以下方向，尤其是算力服务、2B 类的 AI 收入占比较高的应用公司、国产算力中芯片与超节点方向。

1）重视以阿里/字节为代表的国内互联网大厂和模型公司的 Capex 投资和生态圈玩家，包括国产芯片链企业、isv 厂商，以及以大厂为主要客户的基础设施服务供应商等，推荐海光信息、智谱（H 股）、MiniMax(H 股)等。

2）Pre-AI 的收入率先落地。国内企业数据治理相对海外落后，AI 应用落地需要更多前期准备工作。OA+ERP 作为集成企业数据的平台入口直接受益，推荐泛微网络、金蝶国际、用友网络、鼎捷数智；Pre-AI 环节推荐达梦数据、海天瑞声、深信服、汉得信息、第四范式等。

3）部分细分垂直场景 AI 收入更快。其中创收关注工业、医疗、教育、市场营销侧产品，推荐中控技术、合合信息、金山办公、拓尔思、税友股份、视源股份、京北方、国能日新、晶泰控股、创业慧康等；降本关注 AI-coding、多模态生成，推荐万兴科技。

4）本地推理逐步起量，利好服务和云厂商。推荐浪潮信息、深信服、神州数码、海光信息等。

5）推荐关注端侧 AI 和相关产业链，推荐虹软科技、中科创达等。

下载日志已记录，仅供内部参考

五、风险提示

（1）AI 产业商业化落地不及预期：目前各环节 AI 产品的商业化模式尚处于探索阶段，如果各环节产品的推进节奏不及预期，或对相关企业业绩造成不利影响；

（2）市场竞争风险：海外 AI 厂商凭借先发优势，以及较强的技术积累，在竞争中处于优势地位，如果国内 AI 厂商技术迭代不及预期，经营状况或将受到影响；同时，目前国内已有众多企业投入 AI 产品研发，后续可能存在同质化竞争风险，进而影响相关企业的收入；

（3）政策风险：AI 技术的发展直接受各国政策和监管影响。随着 AI 在各个领域的渗透，政府可能会进一步出台相应的监管政策以规范其发展。如果企业未能及时适应和遵守相关政策，可能面临相应处罚，甚至被迫调整业务策略。此外，政策的不确定性也可能导致企业战略规划和投资决策的错误，增加运营的不确定性；

（4）地缘政治风险：在全球地缘政治环境的波动下，尤其美国对中国的出口限制或将直接影响国内企业算力芯片的获取，进而影响其产品研发和市场竞争力。同时，地缘政治风险也可能导致 AI 产品开拓海外市场面临障碍，影响相关企业的营收情况。

下载日志已记录，仅供内部参考

分析师介绍

应瑛

中信建投证券计算机行业首席分析师，伦敦国王学院硕士，8年计算机行业研究经验。2021年加入中信建投，深入覆盖AI产业链、医疗信息化、工业软件、云计算、网络安全等细分领域。

研究助理

李楚涵

lichuhan@csc.com.cn

下载日志已记录，仅供内部参考

评级说明

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

分析师声明

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

法律主体说明

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

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

一般性声明

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

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

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

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

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

中信建投证券研究发展部

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

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

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

中信建投（国际）

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