仅供内部参考，请勿外传！

传媒

模型能力持续跃迁有望带动 AI 应用生态繁荣，云计算产业步入二次加速期

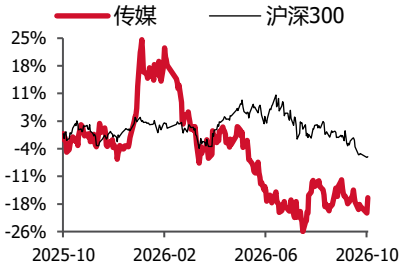
国产开源模型能力快速追赶海外闭源模型，有望加速 AI 在企业端落地。自六月以来，国内多家头部模型厂商密集发布了新一代大模型，呈现出参数规模更大、智能体（Agent）能力更强、开源生态更完善以及多模态应用更深入的趋势。随着国内模型厂商加速模型迭代，国产开源模型能力快速提升，与海外头部闭源模型（如 Anthropic、OpenAI 等）差距持续收窄。相对于闭源模型，开源模型（1）具备更高的性价比；（2）可以更深层次的对模型进行优化，推动模型行为实现自主可控；（3）开源模型可以在公有云、私有云、本地、边缘及混合部署，客户具备多种选择。我们认为随着开源模型能力追赶以及维持较高性价比，顶尖模型 token 价格中枢有望持续下移，进而推动高质量基础模型从可用走向易用，从而加快企业级 AI 部署以及商业化落地，带动需求端从前沿模型实验室扩展到工业、生命科学、金融、公共部门等多个垂直领域。

云计算：北美及国内 AI 云计算行业步入量价齐升阶段。北美/国内 CSP 以及 NeoCloud 公司 26Q2 业绩均已发布，26Q2 Azure/AWS/Google/Coreweave/Nebius/阿里巴巴云业务收入同比增速分别为 43%/37%/82%/112%/454%/45%，均维持加速增长趋势。我们认为云计算行业已经步入量价齐升阶段，主要因（1）开源模型能力快速提升，云厂商可以依据客户对质量、延迟、成本等方面要求为客户提供对应的模型，模型分层策略带动产业话语权由头部模型厂商向下游 CSP 迁移；（2）AI 实际使用成本（cost/Mtok）大幅下降，在成本端具备经济效益之后，杰文斯悖论将带动云需求总量步入爆发增长阶段；（3）开源模型参数量持续攀升，例如 KIMI 7 月 17 日发布的 K3 模型参数量已经达到 2.8 万亿，单次部署的资源门槛同步抬高，企业自建算力不经济，因此更倾向于租赁云厂商基础设施，带动云厂 IaaS+MaaS 收入共振。2026 H1 国内外主要云厂商相继上调算力服务价格，价格战成为过去式。北美云厂商、NeoCloud 及国内云厂商云业务收入增速持续加速，量价齐升逻辑得到报表验证。

AI 应用：Vibe Coding 后，关注 AI4S、FDE、AI 办公等应用端机遇。参照移动互联网时代，AI 应用将成为 Token 价格下降以及模型能力提升直接受益方。26H1 下游变现场景主要集中在 Vibe Coding，随着模型能力提升，AI 有望赋能更多下游应用场景（1）AI4S：AI 用于加速科学研究和发现的过程，目前已经在生命科学、物理学、材料学等多领域落地，打造“AI 预测—自动化验证—数据回流”的完整数据闭环；（2）FDE：模型落地企业业务需要跨过系统对接、权限梳理、流程改造、合规审计、组织适应等难题，FDE 加速模型与企业业务流程落地，面向客户，FDE 把散落在员工脑中的业务经验和流程规则转化为 AI 可调用的知识库、工作流和应用能力；面向平台，FDE 把行业场景、系统接口、测试方法和失败案例沉淀为可复用的产品组件，持续迭代产品，从而降低后续 FDE 项目交付成本与周期；（3）AI 办公：企业以及员工为效率付费意愿动力强，AI 提高工作效率，用户付费意愿较

强于大市（维持评级）

行业走势



作者


分析师 侯宾
SAC：S1070522080001
邮箱：houbin@cgws.com


分析师 纪涵宇
SAC：S1070526010004
邮箱：jihanyu@cgws.com


分析师 黄俊峰
SAC：S1070524120004
邮箱：huangjf@cgws.com

相关研究

1、《GPT-6 Astra 发布，前沿模型能力持续提升》2026-09-15
2、《北美部分软件公司业绩超预期，关注港股互联网及 AI 应用公司底部机会》2026-06-02
3、《传媒互联网板块延续低配，关注后续估值修复》2026-05-12

高；此外，我们认为 AI 在办公场景的高频调用，能持续获取企业私域数据（如审批流、文档），用以微调模型，形成用户使用越多--模型越懂企业--用户越离不开的正向数据飞轮循环。

投资建议：随着模型能力提升以及 AI 使用成本的快速下行，我们认为云计算以及 AI 应用产业景气度正处于上行趋势。云计算：杰文斯悖论将带动云需求总量步入爆发增长阶段，云计算有望维持量价齐升趋势；AI 应用：企业级 AI 应用以及 C 端 Agent 智能体正处于加速落地阶段，其中企业 AI 预算释放已经从试点论证逐步步入规模化落地阶段。

相关标的：云计算/算力租赁：阿里巴巴、金山云、粤港湾智算、利通电子、蜂助手、杰创智能；AI 应用：腾讯控股、快手、迈富时、深演智能、迅策、中科闻歌、神州控股、汇量科技、海致科技集团、阜博集团、焦点科技、中文在线、蓝色光标、巨人网络、恺英网络。

风险提示：AI 大模型迭代不及预期；AI 应用商业化不及预期；行业竞争加剧；监管政策影响。

内容目录

1.开源大模型能力持续提升，带动 AI 应用生态繁荣度提升 5
1.1 国产开源模型加速迭代，能力快速追赶海外闭源模型 5
1.2 开源模型推动 AI 从可用升级为易用，有望加速企业端 AI 落地 7
2. AI 驱动云厂商量价齐升，多模型策略带动下游议价权提升 9
2.1 海外 Hyperscaler 及 NeoCloud 业绩高增，AI 驱动云厂商量价齐升 9
2.2 国内互联网云：AI 需求带动收入增速以及利润率同步增长 14
2.3 国内运营商云：云业务快速增长带动估值中枢提升 17
2.3.1 云已经成为运营商第二增长曲线的核心部分 17
2.3.2 运营商云或成为承担十五五规划重要角色 19
3. Token 价格下降带动 AI 应用起量，关注 FDE、AI4S、AI 办公、多模态等领域投资机会 23
3.1 参照移动互联网时代，Token 价格下降有望带动 AI 应用生态繁荣 23
3.2 FDE：加速模型在产业端落地，完成 AI"最后一公里"交付 24
3.2.1 迈富时：FDE 带动公司 AI 应用业务步入爆发周期 25
3.2.2 神州控股：AI First FDE 28
3.3 AI4S：AI 加速"设计-验证"循环，市场空间广阔 30
3.4 AI 办公：白领办公市场空间广阔，通用办公 Agent 加速涌现 35
4.风险提示 38

图表目录

图表 1： Kimi-K3 前端代码竞技胜率高达 76% 5
图表 2： Kimi-K3 前端代码竞技场榜单以 1679 分位居榜首 5
图表 3： Kimi-K3 综合能力居前 6
图表 4： Kimi-K3 综合能力居前 6
图表 5： Qwen 3.8 模型性能位居全球前列 6
图表 6： Qwen 3.8 重点升级内容梳理 6
图表 7： AI 大模型数据飞轮梳理 7
图表 8： 黄仁勋关于开源模型的公开信 7
图表 9： 国内外前沿模型价格对比 8
图表 10： 开源/闭源模型多维度对比 8
图表 11： 主流基础模型任务成本以及综合能力对比 9
图表 12： 21Q1-26Q2 亚马逊 AWS 云业务收入及运营利润 10
图表 13： 21Q1-26Q2 谷歌 Google Cloud 云业务收入及运营利润 10
图表 14： 21Q1-26Q2 微软智能云业务收入及运营利润 10
图表 15： Nebius 在 TCO、采购效率等方面具备显著优势 12
图表 16： CRWV 及 NBIS 季度收入及调整 EBITDA 利润率 12
图表 17： 云厂商有望维持"量价齐升"趋势 13
图表 18： 国内外云厂商 CAPEX 对比（单位：亿美元） 14

图表 19：国内外云厂商研发费用对比（单位：亿美元）.....14
图表 20：中国三大运营商归母净利率.....15
图表 21：阿里云市占率位居国内 AI IaaS+MaaS 市场份额首位.....16
图表 22：中国公有云大模型调用 token 消耗量将维持高速增长.....16
图表 23：2020-2024 年移动云收入及增速.....17
图表 24：2020-2025 年中国电信天翼云收入及增速.....18
图表 25：2020-2024 年联通云收入及增速.....18
图表 26：中国移动上市至今分阶段走势（截止 2026 年 7 月 5 日）.....19
图表 27："东数西算"八大算力枢纽及十大数据中心集群.....20
图表 28：三大运营商 2026 年上半年资本开支情况.....20
图表 29：中国联通粤港澳大湾区枢纽（韶关）数据中心.....21
图表 30：中国联通宁夏中卫云数据中心.....21
图表 31：全球首个"海风直连海底数据中心"中国电信临港智算中心.....22
图表 32：中国移动云 token 中心及苏州分公司生态联盟揭牌.....22
图表 33：通过软件优化可大幅提升 GPU 吞吐量.....23
图表 34：硬件升级可带动 GPU 吞吐量进一步跃升.....23
图表 35：AI Agent 与人工成本对比.....24
图表 36：FDE 主要职责梳理.....24
图表 37：迈富时 AI Agentforce 3.0 产品优势.....26
图表 38：迈富时 Knowforce 知识中台产品优势.....26
图表 39：迈富时 OntologyForce OS 助力消除数据孤岛.....27
图表 40：迈富时 FDE 案例.....28
图表 41：神州控股燕云 AI First FDE 介绍.....29
图表 42：神州控股燕云 AI First FDE 模式.....29
图表 43：科研研究的演进范式.....30
图表 44：AI4S 具备广泛的应用领域.....31
图表 45：AI4S 应用领域具体案例梳理.....31
图表 46：AI4S 具备广泛的应用领域.....32
图表 47：国际主流 AI4S 平台产品矩阵.....33
图表 48：缺乏高质量数据为 AI 商业化落地最重要的卡点.....34
图表 49：Anthropic Mythos 5 进行蛋白质复合物设计.....34
图表 50：自动化实验室工作流程.....35
图表 51：72%工作者使用 codex 进行知识内容制作.....36
图表 52：Claude Cowork 1/3 使用与业务流程相关.....36
图表 53：ChatGPT 与 Codex 合并.....36
图表 54：Claude Cowork 与 Chat 合并为统一入口.....36
图表 55：国内主流办公 Agent 对比.....37

1.开源大模型能力持续提升，带动 AI 应用生态繁荣度提升

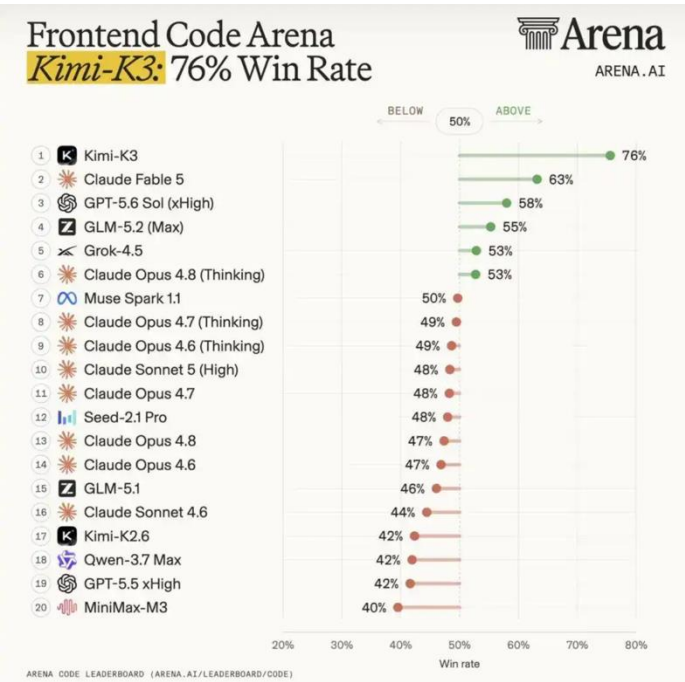
1.1 国产开源模型加速迭代，能力快速追赶海外闭源模型

智谱、Minimax、Kimi、DeepSeek、千问等国内模型厂商陆续发布新一代大模型，国产模型加速迭代。自六月以来，国内多家头部模型厂商密集发布了新一代大模型，呈现出参数规模更大、智能体（Agent）能力更强、开源生态更完善以及多模态应用更深入的趋势。例如，Kimi 于 7 月 17 日正式发布的新一代 AI 大模型，拥有 2.8 万亿参数和 100 万 token 上下文窗口。我们认为国产开源模型厂商已经步入加速迭代期，建议持续关注国内头部模型厂商新一代大模型发布情况。

国产开源模型能力快速提升，与海外头部闭源模型差距持续收窄。随着国内模型厂商加速模型迭代，国产开源模型能力快速提升，与海外头部闭源模型（如 Anthropic、OpenAI 等）差距持续收窄：

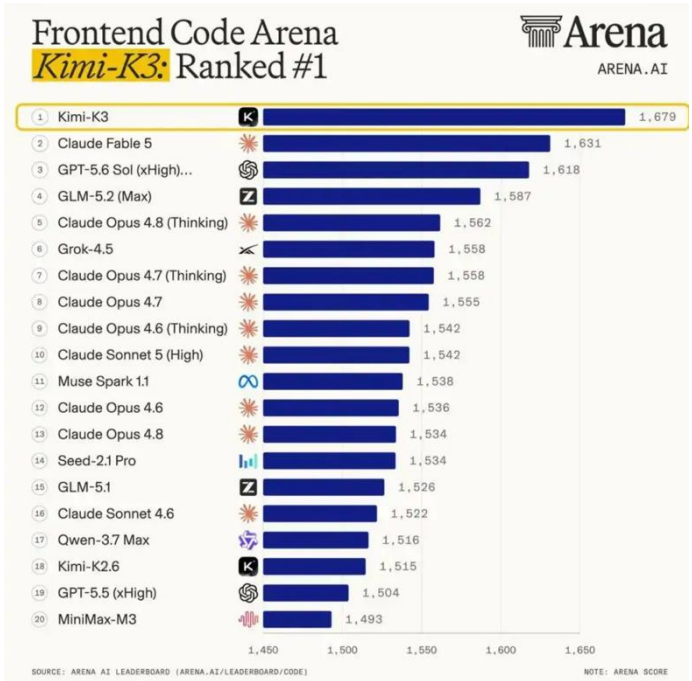
- Kimi K3：Kimi K3 是月之暗面在 2026 年 7 月 17 日发布的最新大模型，拥有 2.8 万亿参数和 100 万 token 上下文。在 Arena AI 前端代码竞技场榜单中，Kimi K3 以 1679 分的成绩排名全球第一，超过了 Claude Fable 5 和 GPT-5.6 Sol 等闭源旗舰模型。该模型基于 KDA 混合线性注意力机制和注意力残差技术构建，有效降低长序列显存压力，更高效的处理超长内容。该模型缓存命中价格为 2 元/百万 token，未命中为 20 元/百万 token（约 3 美元）；输出价格为 100 元/百万 token（约 15 美元）。

图表1：Kimi-K3 前端代码竞技胜率高达 76%



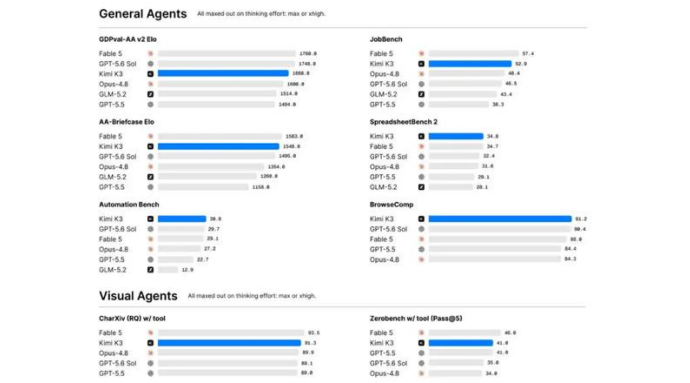
资料来源：Arena，长城证券产业金融研究院

图表2：Kimi-K3 前端代码竞技场榜单以 1679 分位居榜首



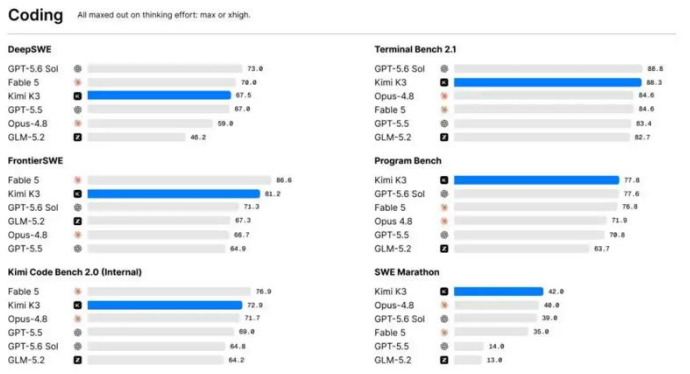
资料来源：Arena，长城证券产业金融研究院

图表3：Kimi-K3 综合能力居前



资料来源：Kimi 微信公众号，长城证券产业金融研究院

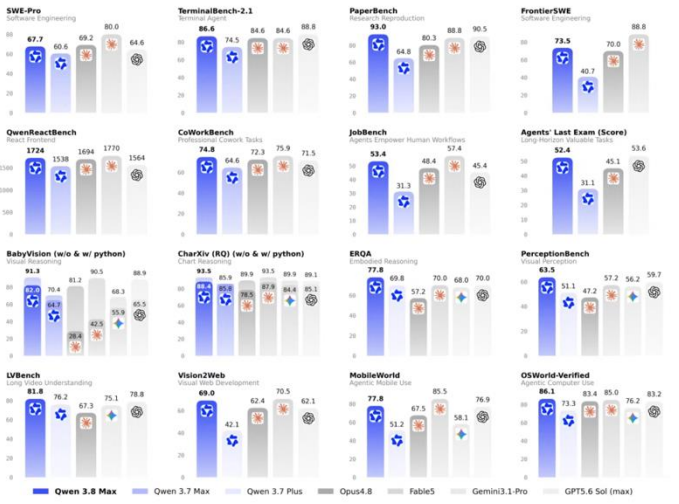
图表4：Kimi-K3 综合能力居前



资料来源：Kimi 微信公众号，长城证券产业金融研究院

➢ Qwen 3.8：阿里巴巴于 8 月 3 日正式发布新一代基座大模型 Qwen3.8，总参数量 2.4 万亿，在编程（Coding）和专业办公（Cowork）方面能力大幅提升。8 月 3 日放榜的权威三方榜单 Arena 中，阿里 Qwen 模型仅次于 Anthropic 的 Claude 系列，整体性能处于全球大模型第一梯队。Qwen3.8 的 API 同时上线千问 AI 平台，并接入阿里同步推出的 Agent 产品“千问办公”。Qwen3.8-Max、Qwen3.8-27B 均已开源。该模型价格为每百万 Tokens 输入 12 元、输出 36 元，隐式缓存命中仅 1.5 元。除了拥有强大的文本和推理能力，Qwen3.8 还提高了 AI 视觉理解能力的极限。8 月 3 日，在权威 Vision Arena 榜单中，Qwen3.8-Max 排名全球第二。

图表5：Qwen 3.8 模型性能位居全球前列



资料来源：千问大模型公众号，长城证券产业金融研究院

图表6：Qwen 3.8 重点升级内容梳理

升级方向	具体情况
编程能力	Qwen3.8 实现了自主编程的新突破。在权威 CodeArena 榜中，Qwen3.8 位居全球第四。2 年前的前沿模型能写好函数、补全代码，成为程序员的有力帮手，而如今，Qwen3.8 可以从一个空文件夹出发，自主完成一个十数天的真实项目交付，全程无需人干预
办公场景	Qwen3.8 可胜任广泛的、真实的专业工作任务（Cowork）。面对完全不同的专业任务、评判标准和计算需求，Qwen3.8 通过真实环境和算力的联合强化学习（RL）扩展，实现了数百种高价值、高频专业任务的真实表现提升，在主流的智能体框架中均可稳定可靠地交付生产级成果
视觉理解	在权威 Vision Arena 榜单中，Qwen3.8-Max 排名全球第二。在千问团队构建的“应用复刻”基准 RecreationBench 长程任务中，模型没有源代码，也无法联网，只通过交互与反馈去理解这个应用，却能从零复刻出整个应用。这表明，Qwen3.8 已经展现出前沿水准的混合多模态智能体能力，通过“迭代编程—交互反馈”的反复循环，将视觉编程（Visual Coding）推向新的高度
长程任务	面对极其复杂的长周期、多约束任务，Qwen3.8 展现出了超越以往的系统级自主规划能力——无论是高精密、强物理约束的芯片设计，还是高度动态、博弈激烈的商业经营，它都能靠“执行—反馈—迭代”的自适应闭环，在数千轮的超长交互中完成算法与策略的深度重构

资料来源：澎湃新闻，长城证券产业金融研究院

我们认为在跨过“可用线”之后，国产大模型能力以及迭代速度均有望持续提升。此前，随着国产大模型智能化水平持续提升，国产模型此前已经跨过了“可用线”，即用户愿意在真实的工作流、工作场景中使用领先的国产大模型。通过用户真实的使用，模型厂商可以逐步去积累高质量的用户数据，包括提示词、代码上下文、输出结果、追问内容、漏洞报告以及用户反馈等，并通过以上数据了解用户具体需求，从而可以针对相应的工作流、工作场景实现针对性的优化，从而实现“AI 时代”的数据飞轮效应。我们认为这种基于真实开发者工作流进行训练反馈的模型能够学习用户如何在真实场景中解决复杂的问题，并在后续迭代中进行针对提升，从而循环往复，持续提升模型能力以及用户体

验。区别于“蒸馏”仅仅可以“获得答案”，上述数据飞轮可以理解为“实况录像”，记录用户的问题、项目背景、问题反馈、迭代循环、人工修整等真实的工作目标及流程。

图表7：AI 大模型数据飞轮梳理

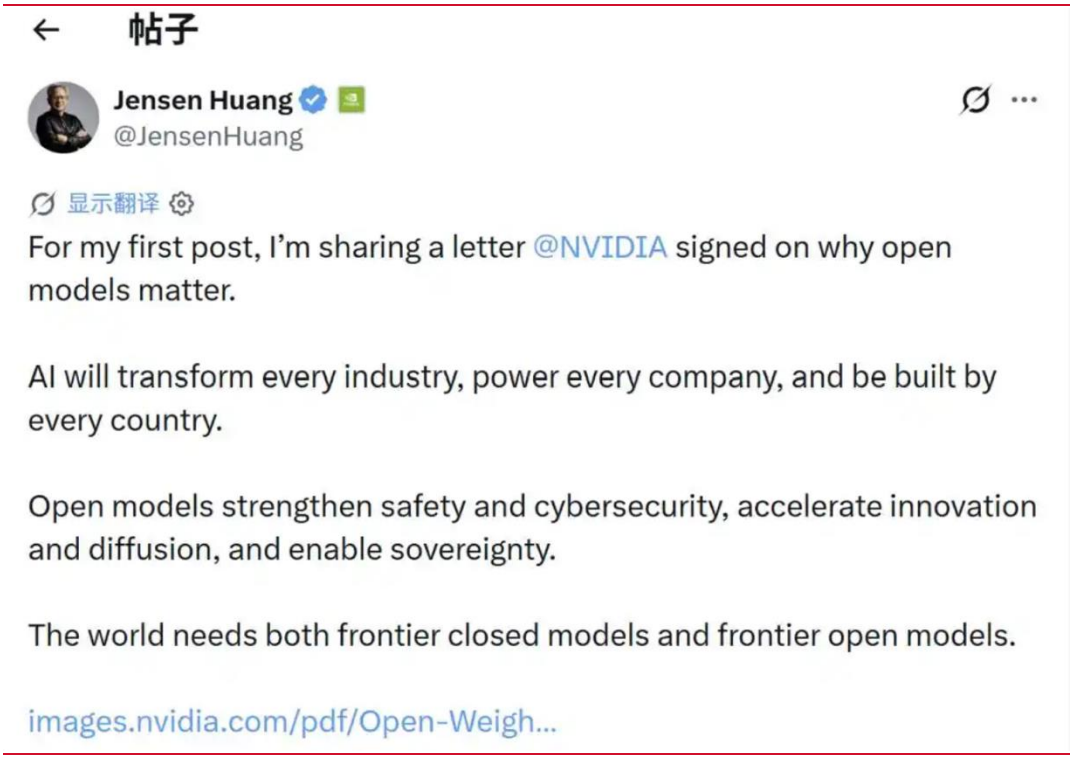


资料来源：长城证券产业金融研究院整理

1.2 开源模型推动 AI 从可用升级为易用，有望加速企业端 AI 落地

开源模型能够强化网络安全，改造每个行业，赋能每家公司。根据 36 氪，英伟达 CEO 黄仁勋此前接受采访时称科学的进步、计算机科学的发展以及网络安全需要依赖开源大模型。因为开源意味着透明，意味着整个技术社区能够共同检查、测试、挑错，并在最大程度上加固系统。这才是对网络安全、国家安全和经济安全最有利的做法。黄仁勋认为 AI 将变革每一个行业，赋能每一家公司，并被每一个国家所构建。开源模型能强化安全性与网络安全，加速创新与普及，并实现主权。世界既需要前沿的闭源模型，也需要前沿的开源模型。开放权重模型能扩大 AI 访问范围、加强竞争与创新、降低成本、提升客户控制力。

图表8：黄仁勋关于开源模型的公开信



资料来源：界面新闻，长城证券产业金融研究院

开源模型在性价比、数据安全、可定制性等方面优于闭源模型：

- **价格：**开源模型通过社区优化等方式，推理性价比优于闭源模型；以国内头部开源模型 Kimi-K3 为例，其缓存命中价格为 2 元/百万 token，未命中为 20 元/百万 token

（约 3 美元）；输出价格为 100 元/百万 token（约 15 美元）；与其模型能力相近的 Opus 4.8 模型标准模式输入价格为 5 美元/百万 token，输出价格为 25 美元/百万 token。根据钛媒体，Kimi K3 单任务成本约为 Opus 4.8 的一半，Token 单价低约 40%；

- **数据层面：**我们认为开源模型可进行全参数微调、LoRA 轻量适配、知识注入，打造企业专属“领域专家”模型，可以更深层次的对模型进行优化，模型行为可以实现自主可控；但是闭源模型通常只能通过提示词工程或有限的外部知识库（RAG）进行浅层优化，如同无法修改“黑箱”内核。
- **模型层面：**闭源模型使用仅仅可以调用固定能力，但是开源模型可以进行微调、蒸馏、对齐等，相同的开源基座模型可以通过企业赋予的行业 Know-How，从而形成针对不同行业、岗位、任务的应用版本，进而赋能更多软件公司。软件公司可以通过自身积累的行业数据、对客户工作流的理解以及模型工程能力，为其客户提供“AI 原生系统”落地的服务。
- **产品层面：**在产品落地上，区别于闭源模型须经外部 API 传输处理，开源模型可以在公有云、私有云、本地、边缘及混合部署，客户具备多种选择。企业级应用在产品层面需求较为广泛，覆盖数据库查询、日志监控、工作流调度等多元化任务，因此对于数据治理、实时监控、权限体系、工作流编排等方面的治理需求进一步增强。软件公司需要提供对应匹配的治理能力，我们认为以上需求有望进一步带来客户价值量提升。

图表9：国内外前沿模型价格对比

模型	输入价格	输出价格
Qwen 3.8 Max	2 美元/Mtok 缓存命中 0.25 美元/Mtok	6 美元/Mtok
Kimi K3	3 美元/Mtok 缓存命中 0.3 美元/Mtok	15 美元/Mtok
DeepSeek V4 flash	平时 1 元/Mtok；峰时 2 元/Mtok； 缓存命中：平时 0.02 元/Mtok，峰时 0.04 元/Mtok	平时 2 元/Mtok；峰时 4 元/Mtok；
GLM 5.2	1.4 美元/Mtok	4.4 美元/Mtok
Claude Fable 5	10 美元/Mtok 缓存命中 1 美元/Mtok	50 美元/Mtok
Claude Opus 4.8	5 美元/Mtok 缓存命中 0.5 美元/Mtok	25 美元/Mtok
GPT 5.6 Sol	5 美元/Mtok 缓存命中 0.5 美元/Mtok	30 美元/Mtok

资料来源：腾讯云，gate.ai，钛媒体，极智 AI，长城证券产业金融研究院

图表10：开源/闭源模型多维度对比

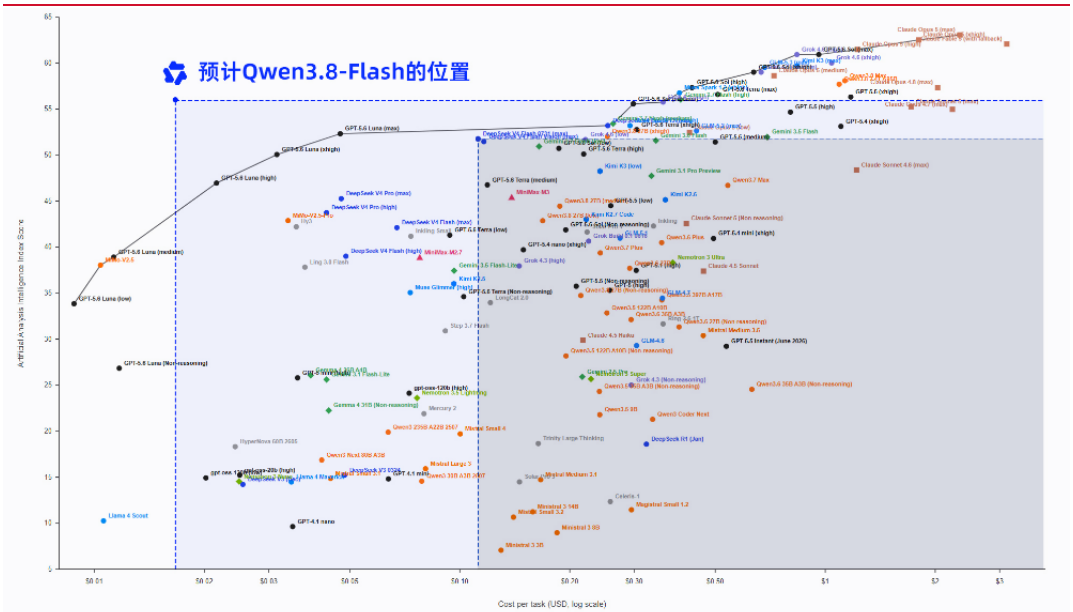
维度	对比
价格	国内开源模型厂商在模型能力快速提升的同时，维持着较高的性价比优势，如 Kimi K3 单任务成本约为 Opus 4.8 的一半，Token 单价低约 40%
数据	可进行全参数微调、LoRA 轻量适配、知识注入，打造企业专属“领域专家”模型，可以更深层次的对模型进行优化，模型行为可以实现自主可控；但是闭源模型通常只能通过提示词工程或有限的外部知识库（RAG）进行浅层优化，如同无法修改“黑箱”内核
模型	闭源模型使用仅仅可以调用固定能力，但是开源模型可以进行微调、蒸馏、对齐等，相同的开源基座模型可以通过企业赋予的行业 Know-How，从而形成针对不同行业、岗位、任务的应用版本，进而赋能更多软件公司
产品	区别于闭源模型须经外部 API 传输处理，开源模型可以在公有云、私有云、本地、边缘及混合部署，客户具备多种选择。企业级应用在产品层面需求较为广泛，覆盖数据库查询、日志监控、工作流调度等多元化任务，因此对于数据治理、实时监控、权限体系、工作流编排等方面的治理需求进一步增强

资料来源：钛媒体，长城证券产业金融研究院整理

开源模型能力大幅提升有望降低企业部署高质量模型门槛。开源模型开放权重后，可以显著降低企业端部署高质量模型的成本。国内 Kimi K3 正式开放模型权重、技术报告及关键训练基础设施；DeepSeek V4-Flash 在 Agent 能力显著提升的同时延续极致性价比；海外模型厂商 OpenAI 亦下调 GPT-5.6 Terra 和 Luna 价格。我们认为随着开源模型能力追赶以及维持较高性价比，顶尖模型 token 价格中枢有望持续下移，进而推动高质量基础模型从可用走向易用，从而加快企业级 AI 部署以及商业化落地。根据 Coreweave 26Q2 业绩会，目前公司需求端已经从从前沿模型实验室扩展到工业、生命科学、金融、公共

部门等多个垂直领域，显示出 AI 模型已经从提升能力逐步走向垂直行业落地的阶段。

图表11：主流基础模型任务成本以及综合能力对比



资料来源：凤凰财经，机器之心，长城证券产业金融研究院

2. AI 驱动云厂商量价齐升，多模型策略带动下游议价权提升

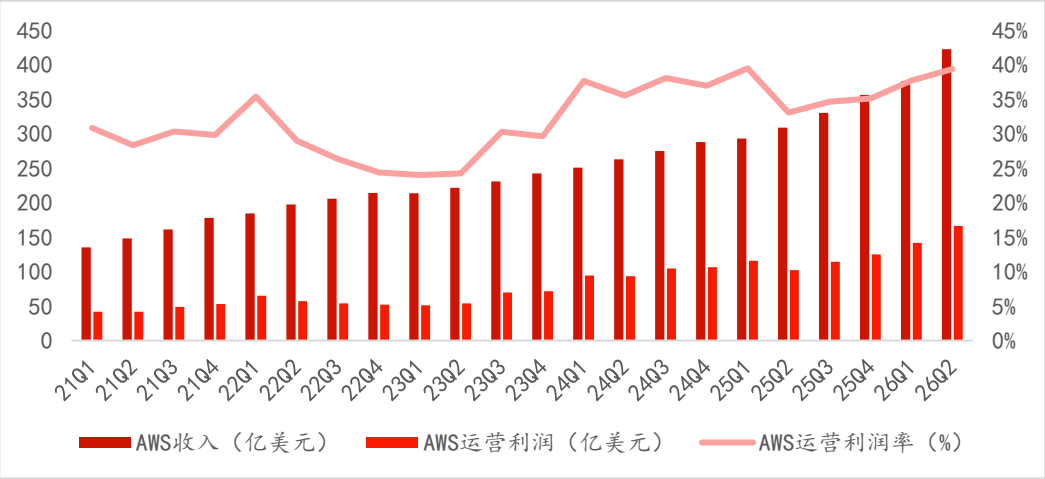
2.1 海外 Hyperscaler 及 NeoCloud 业绩高增，AI 驱动云厂商量价齐升

海外云计算厂商陆续发布业绩，云业务收入均维持高质量增长，我们认为主要受益于企业级 AI 部署需求逐步释放，AI 需求正在由基础设施逐步向平台服务、企业软件等方向传导，云厂商 AI 商业闭环进一步形成：

- **微软：**FY2026Q4 公司 Microsoft Cloud 实现收入 593 亿美元，同比增长 27%；其中 Azure 及其他云服务收入同比增长 43%。Microsoft 365 Copilot 付费席位超过 3000 万，商业剩余履约义务达到 6780 亿美元、同比增长 84%。公司指引 FY2027Q1 Azure 固定汇率增速约 45%，并表示需求仍高于可用供给。微软已经在 AI 基础设施、Azure 多模型平台和 Copilot 软件订阅实现多层变现。随着高质量开源模型供给提升，Azure 作为模型适配、部署和治理平台的价值有望继续提升。
- **亚马逊：**公司 2026Q2 AWS 实现收入 422 亿美元，同比增长 36.7%，为 18 个季度以来最快增速，公司认为长期 AWS 年收入有望达到万亿美元级别；实现营业利润 166 亿美元营业利润率达 39%。公司 AI 业务及自研芯片业务 ARR 均超 250 亿美元，保持三位数增长。AWS 总未履约订单达到 4960 亿美元，同比增长达到三位数，约为 25Q3 的 2.5 倍；2027 年大部分新增容量已经被预订，2028 年也已有相当规模的容量被预订。
- **谷歌：**2026Q2 Google Cloud 实现收入 248 亿美元，同比增长 82%；云业务营业利润 88 亿美元，营业利润率 35.6%，环比持续增长；云业务积压订单 5140 亿美元，环比增加约 500 亿美元。公司称增长由 AI 基础设施和企业 AI 解决方案共同推动，

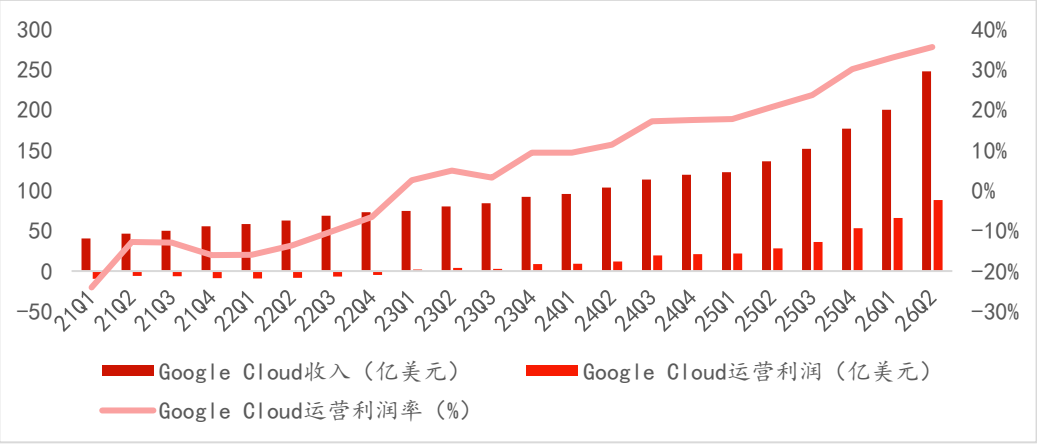
GCP 增速快于 Cloud 整体。Alphabet 2026Q2 资本开支 449 亿美元，约 60%用于服务器、40%用于数据中心和网络设备；公司将 2026 年全年资本开支指引提高至 1950 亿至 2050 亿美元，并表示需求仍快于投资扩张。

图表12：21Q1-26Q2 亚马逊 AWS 云业务收入及运营利润



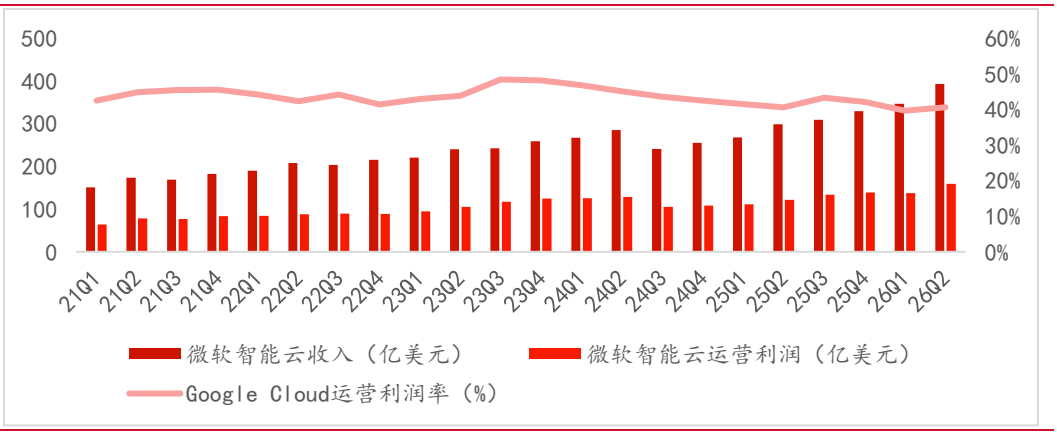
资料来源：亚马逊 21Q1-26Q2 季度业绩报告，长城证券产业金融研究院

图表13：21Q1-26Q2 谷歌 Google Cloud 云业务收入及运营利润



资料来源：谷歌 21Q1-26Q2 季度业绩报告，长城证券产业金融研究院

图表14：21Q1-26Q2 微软智能云业务收入及运营利润

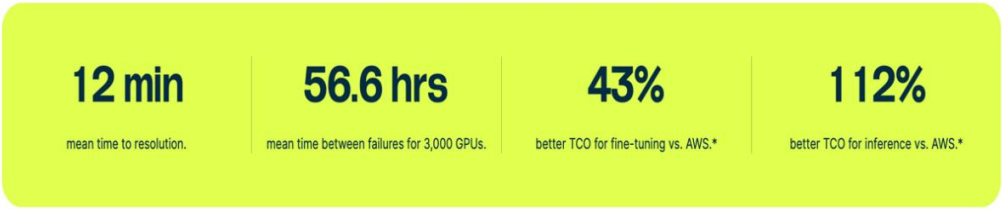


资料来源：微软 21Q1-26Q2 季度业绩报告，长城证券产业金融研究院

NeoCloud 专注提供 GPU 基础设施，具备更高的采购效率以及灵活性。NeoCloud 主打 GPU 算力基础设施服务，在采购落地与使用弹性上具备突出优势。面向 AI 业务时，传统公有云的多项核心能力，例如虚拟化组件、安全隔离机制、自研 SSD 存储等，实际发挥的价值会被削弱，部分机制甚至会对 AI 集群整体性能带来负面影响。以 Nebius 的实践为例，其 GPU 设备与网络通路绕过传统虚拟化转发逻辑，能够输出近乎裸金属级别的硬件性能。基于这套技术思路，NeoCloud 服务商可以做到更高的模型 FLOPs 利用率（MFU），同时降低跨设备通信带来的损耗，整体拉高 GPU 算力底座的运行效率，也给到业务更大的调度弹性。除此之外，各大云服务厂商（CSP）也会把 NeoCloud 当作弹性的短期算力补充资源。借助长租模式，企业能够把风险较高的资本性投入转变为运营成本，微软就是 NeoCloud 服务商的头部客户之一。综合以上内容，可以梳理出 NeoCloud 在 AI 业务场景，对比传统 CSP 厂商的四大核心优势：

- **GPU 集群效率：**大模型训练与推理场景里，CPU-GPU 数据交互、存储 IO 读写、调度开销等问题，会造成 GPU 理论算力无法全部投入模型运算。传统 CSP 需要承载虚拟机、数据库、存储、CDN、Serverless、企业级 IT 等多元化业务负载；而 NeoCloud 的数据中心是专为 GPU、NVLink、InfiniBand/RoCE、高性能存储这套组合架构定制化打造。最终带来更高 MFU、更少通信开销以及更低故障概率。即便 GPU 物理数量相同，也能输出更多可用实际算力，提升集群整体运行效率，帮助客户压缩训练成本。
- **性价比更高：**根据 Nebius 官网，相较 Nebius，AWS 微调任务的 TCO 高出 43%，推理场景 TCO 高出 112%。Nebius 平台平均故障处理时长为 12 分钟，3000 张 GPU 集群可实现接近 57 小时的稳定运行。CoreWeave 也公开表示，其算力底座能够让客户的 MFU 最高提升 20%，吞吐量提升可达 96%。背后核心原因在于传统 CSP 附带整套企业化功能，涵盖安全防护、身份权限管理、数据库、网络组件、监控体系、企业技术支持等模块。而 NeoCloud 只聚焦 GPU、网络、存储以及 AI 相关软件，省去通用云附加组件带来的额外开销，因此租赁成本更低。
- **更高的采购效率以及灵活性：**以 AWS 作为参照，其热门区域的 p5 容量需要提前数周预约，或是签署企业折扣协议才能够拿到。反观 Nebius，H100、H200 以及 Blackwell B200 规格的算力资源支持即时按需取用。采购流程上，用户仅凭企业信用卡，短短数分钟就可以拉起一台 8 卡 GPU 的 Nebius 节点。而 AWS 对等服务，则要走完采购审批、EDP 条款磋商，甚至耗费数月时间完成算力配额谈判。
- **硬件优先分配权：**英伟达主动与 NeoCloud 厂商合作，将其作为核心战略抓手，并承诺（1）优先硬件分配：给头部 NeoCloud 厂商最新 GPU 的优先供货权；（2）融资与债务背书：为 NeoCloud 厂商的债务融资提供担保，降低其融资成本；（3）直接股权投资：对 CoreWeave、Nebius 等头部厂商均有战略持股，深度绑定利益。GPU 性能迭代速度远快于传统服务器，因此 NeoCloud 正将 NVIDIA relationship + GPU allocation + deployment speed 作为自身核心竞争力。

图表15：Nebius 在 TCO、采购效率等方面具备显著优势

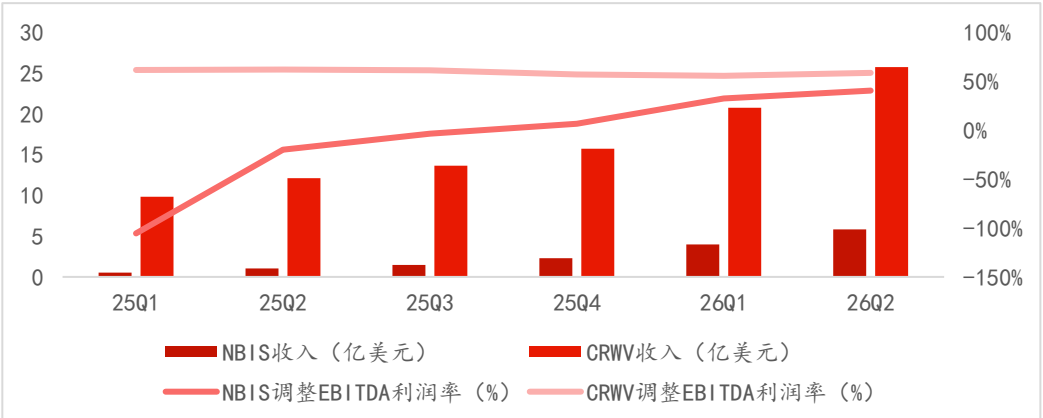


资料来源：Nebius 官网，长城证券产业金融研究院

NeoCloud 公司进一步推出“Token 工厂”业务，进一步提升其全栈服务能力。以 Nebius 为例，公司于 2025 年 11 月推出 Token Factory 托管推理平台，把多个开源模型（DeepSeek、Llama、Qwen 等）进行加工微调，将每个模型的推理成本砍掉最高 70%（通过 ODM 硬件、Eigen CUDA Kernel、推理调度等层面优化），再以 Token 定价对外输出。此前，Nebius 通过收购 Eigen AI 推动模型侧优化，提升模型推理 Token 吞吐量，从而提升 GPU 利用率，降低大规模部署大模型的推理成本，有效提升了 MoE 和推理模型在生产环境中稳定、经济地运行的能力。

NeoCloud 业绩超预期，AI 算力需求旺盛。NeoCloud 公司 26Q2 业绩均显著超市场预期，其中 NBIS 26Q2 实现收入 5.82 亿美元，同比增长 454%，其中 AI 云实现收入 5.75 亿美元，同比增长 514%；公司 26Q2 实现调整后 EBITDA 2.36 亿美元，首次大幅转正；CRWV 26Q2 实现营收 25.75 亿美元，同比增长 112%，净亏损 6.26 亿美元，积压订单达 1040 亿美元，实现调整后 EBITDA 15.10 亿美元，EBITDA margin 达 59%。整体而言，AI 云算力需求仍显著超过近期可交付供给，CoreWeave 26Q2 末已投运功率约 1.5GW、合同功率约 3.7GW，积压订单约 1040 亿美元，而且尚未包括三季度初新增的超过 250 亿美元承诺；Nebius 将 2026 年底签约用电容量目标由超过 4 GW 上调至 5 GW，并计划从 2027 年起每年新增超过 1GW 容量。此外，AI 算力供应方定价权持续增强，例如 Nebius 披露旧代 GPU 价格较一季度上涨超过 30%，二季度新签合同平均年化收入超过每 MW 2000 万美元，部分达到 2000 万—2500 万美元，三季度初的短期紧急容量甚至达到每 MW 4000 万—5000 万美元。向后展望，NeoCloud 厂商正在向“AI 基础设施操作系统”升级。未来竞争不再只是 GPU 租赁，而是提供训练、推理、存储、网络、模型部署、监控、治理、Agent 运行环境以及跨云运营的一体化平台。

图表16：CRWV 及 NBIS 季度收入及调整 EBITDA 利润率



资料来源：NBIS 及 CRWV 25Q1-26Q2 季度业绩报告，长城证券产业金融研究院

向后展望，受益于开源模型能力提升以及 AI 需求爆发，云厂商有望迎来量价齐升：

- **模型分层策略带动下游话语权提升：**此前前沿模型主要由北美闭源厂商垄断，导致下游云厂商以及应用厂商缺乏其他选项，导致中游模型厂商具备较高话语权，头部厂商具备较高毛利率。而现在以中国 DeepSeek、阿里 Qwen、Kimi、智谱为首的开源模型性能逐步接近闭源模型，在大多数任务中早已跨过可用门槛。因此，下游云/应用厂商不需要与模型厂商进行绑定，例如微软已经在其云平台提供超 11000 个模型，可以依据客户对质量、延迟、成本等方面要求为客户提供对应的模型。因此，我们认为模型之间越可替代，云厂商越能掌握路由权、客户入口和定价权；过去闭源模型是云上游的收费站，未来模型更像云平台可以自由调度的一种计算资源。云厂商既可以出租 GPU，也可以在自有基础设施上托管开放模型、销售 Token，从而减少向外部模型厂商让渡的利润。
- **推理成本大幅下行，杰文斯悖论将驱动 AI 云总需求持续扩张：**北美企业正逐步从“Token maxing”转向为“Token optimize”，顶尖模型可以作为“智能中枢”对复杂长程任务进行规划以及拆解，并将子任务根据难度分配给不同模型；此外价格更低的开放权重模型已经开始对整个模型层形成定价压力，美国多家模型实验室正推出更低价产品。以上两方面因素正在推动 AI 实际使用成本（单位 Token 价格）大幅下降，在成本端具备经济效益之后，我们认为 AI 云需求有望得以增长。同时，2026H1 国内外主要云厂商相继上调算力服务价格，价格战成为过去式。北美云厂商、NeoCloud 以及国内云厂商云业务收入增速持续加速，量价齐升逻辑得到报表验证。
- **开源模型带动云厂商 MaaS+IaaS 业务同步增长：**开源模型开放模型权重，用户能够取得模型权重，企业可以下载模型，根据自身数据进行微调或重新训练，每次下载及微调即产生一份独立的算力与内存需求，而闭源模型只需一份权重集中部署。此外，开源模型参数量持续攀升，例如 KIMI 7 月 17 日发布的 K3 模型参数量已经达到 2.8 万亿，单次部署的资源门槛同步抬高，企业自建算力不经济，因此更加倾向于租赁云厂商基础设施，带动云厂 IaaS+MaaS 收入共振。综上所述，我们认为开源模型多端部署有望加深企业对云厂商的依赖。
- **其他增值服务提供更大的变现空间：**企业部署开放权重模型后，通常还需要购买数据库、存储、RAG、安全、身份认证和监控等其他服务，例如将 AI 模型连接内部专有数据等，给云厂商提供进一步变现空间。

图表17：云厂商有望维持“量价齐升”趋势



资料来源：长城证券产业金融研究院整理

对于云厂商而言，token 吞吐量的提升有望进一步提升 ROIC。Token 价格处于下降趋势，但是价格下降对 ROIC 的负面影响有望被吞吐量(即每张 GPU 每秒能产出的 tokens

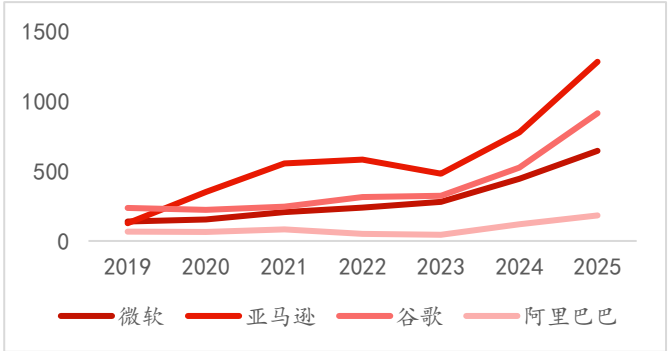
数，Tokens/sec/GPU)的更大幅度提升所抵消并反超，带动 Tokens 生产成本下行。我们认为吞吐量可以通过以下四种方式实现提升（1）采用内存容量更大、互联速度更快的硬件，根据 Semianalysis，GB300 NVL72 在 FP8 精度下的吞吐量,比最优化配置的 H100 高出约 17 倍；如果切换到 FP4，差距进一步拉大到 32 倍；但是同时 GB300 的单 GPU TCO 只比 H100 高出约 70%；（2）进行软件优化，如通过 WideEP（宽专家并行）、Disagg（分离式推理）、MTP（多 token 预测）等方式，实现吞吐量的大幅提升，根据 Semianalysis，在相同硬件端基础上，在三种配置下分别能达到约 1k/8k/14k Tokens/秒/GPU 的吞吐量；（3）利用云巨头自身的软硬件优势，例如谷歌的 TPU 和亚马逊的 Trainium 等定制芯片（ASIC）；（4）通过实施批处理和并发请求优化方案提高 GPU 利用率。随着每颗 GPU 每秒可以生成更多 Token，即使单位 Token 价格下降，每 GW 算力能够创造的收入仍可能保持在较高水平。向后展望，我们认为未来 AI 竞争核心将不再单纯围绕“谁拥有最强模型”，而是越来越取决于“谁能够以更低成本生产更多 Token”。

2.2 国内互联网云：AI 需求带动收入增速以及利润率同步增长

国内互联网云厂商利润率远低于海外云厂商，我们总结了以下三方面原因：

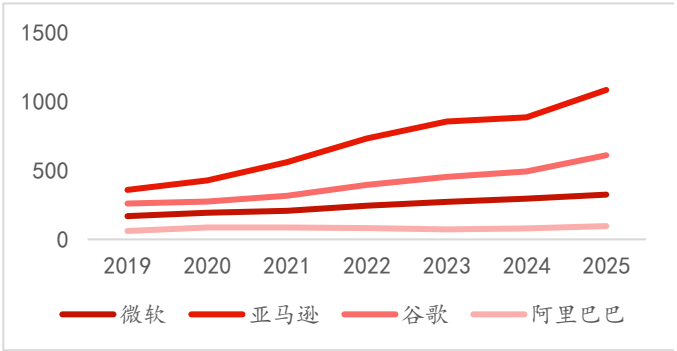
- **规模效应：**以微软、AWS 为代表的美国云厂商起步较早，且进行了全球布局，因此比国内云厂商更早地进入了规模经济收获期，我们认为北美云厂相对国内云厂规模效应主要体现在三方面（1）全球布局：不同地区企业客户的需求负载存在较为明显的时间差异，因此云厂商可以通过跨区域的弹性调度实现计算、存储和网络资源的动态匹配，进而将不同区域的客户需求纳入统一的资源管理池，提高服务器等硬件设备的运行以及使用效率；（2）采购成本：云厂商需要大量采购服务器、网络、数据中心等硬件/软件设备采购，海外云厂商 CAPEX 规模远超国内云厂商，其大体量订单能够获取阶梯折扣/定制硬件，从而压低软硬件单位采购成本；（3）北美头部云厂商维持高强度研发投入，除 IaaS 业务外，持续完善托管数据库、Serverless 函数等 PaaS 服务，全面布局企业数据治理、业务运维、应用研发等流程；IaaS 业务迁移灵活性较高，但是更换 PaaS 服务供应商需要进行大量的数据迁移、改写业务代码、重构分布式架构等工作，因此转换成本较高，有效提升了客户粘性。

图表18：国内外云厂商 CAPEX 对比（单位：亿美元）



资料来源：微软、亚马逊、谷歌、阿里巴巴 2019-2025 年年报，长城证券产业金融研究院

图表19：国内外云厂商研发费用对比（单位：亿美元）

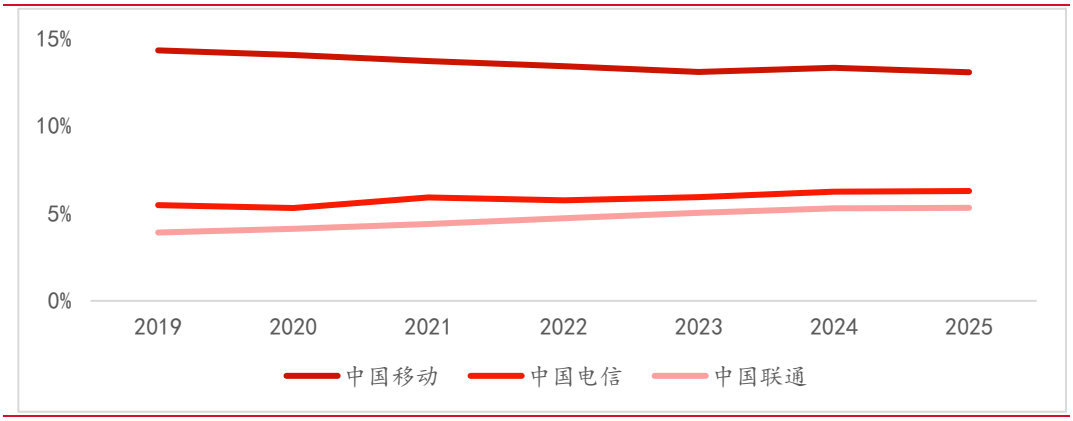


资料来源：微软、亚马逊、谷歌、阿里巴巴 2019-2025 年年报，长城证券产业金融研究院

- **带宽服务：**骨干网为算力调度、IDC 机房互联、数据流通的基础设施支撑，海外允许美国云厂商建立完全自控的全球骨干网络，根据 AWS 官网，AWS 具备近 2000

万公里陆地及海底光纤电缆。但是中国云厂商受合规监管因素，无法在国内自建骨干网络，只能依赖中国电信、中国移动、中国联通等运营商的底层网络基础设施，因此带宽使用费为国内云厂商成本项，因此运营商利润率是中美云厂利润率差异的主要来源之一。

图表20：中国三大运营商归母净利率

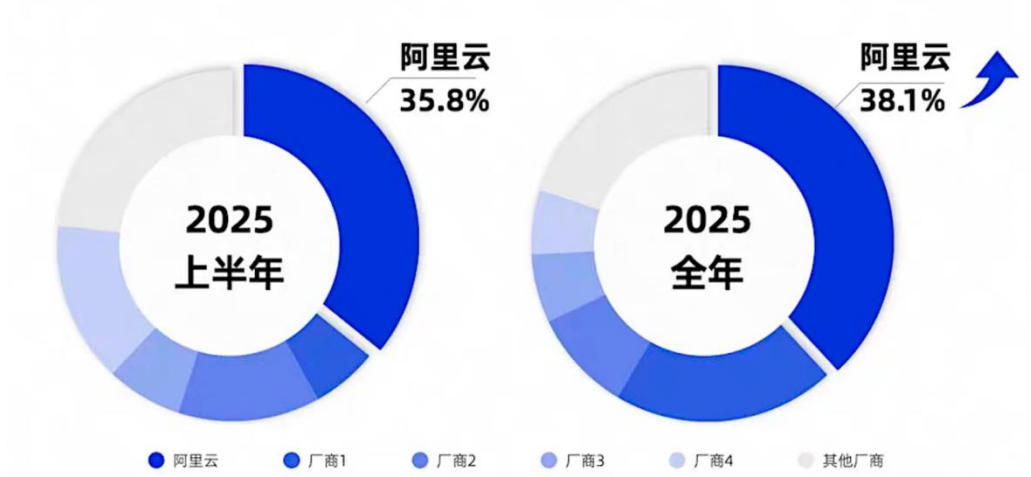


资料来源：中国移动、中国电信、中国联通 2019-2025 年财报，长城证券产业金融研究院

- **客户付费意愿：**相对全球其他地区而言，国内私有云占比偏高，国内国央企均偏好私有云部署方式，根据中国信通院，2022 年中国公有云的占比为 28%，而美国占比达 60%。公有云通过智能调度以及大规模集群管理可以有效提升整体资源利用率，通过全球化调度持续优化资源配置效率、降低边际成本；但是私有云则偏向项目制，其资源搭建、服务器运维、数据存储均服务于特定客户，因此资源无法进行跨区域调度。综上，私有云整体利润率低于公有云。

随业务结构调整完毕以及 AI 时代来临，互联网云厂商市场份额逐步企稳回升。2021 年起，受以下三大因素影响（1）泛互联网行业上云需求下滑；（2）战略聚焦下，主动缩减低利润率的集成项目，同时更为聚焦公有云项目；（3）在获取政企客户上，相对缺乏经验以及优势；互联网云厂商市场收入增速放缓，份额逐步下滑，而运营商云市场份额则实现快速增长。2025 年起，互联网公司云业务结构调整已基本完毕，同时 AI 时代来临，我们认为互联网云有望抓住市场机遇，实现市场份额的企稳回升，主要因互联网公司（1）持续加大针对芯片的投入，包括自研芯片投入以及外部芯片采购；（2）持续加大 AI 大模型层面投入，且基于各个公司生态场景的不同，如阿里巴巴的电商生态、字节跳动的短视频生态、以及腾讯的社交生态，互联网厂商的大模型在细分场景下具备独特优势；（3）于 toB/toC 场景均占据较高份额，具备优质的应用流量入口，因此相关 AI 产品可以由原流量入口导流或以插件形式嵌入原流量入口，因此互联网厂商在 AI 应用推广及迭代上亦具备优势。我们认为互联网云厂商有望通过高效调度异构算力、优化模型推理性能、降低单位 Token 成本等竞争优势在 AI 时代实现云业务收入加速增长。以阿里云为例，根据 IDC 大模型训推公有云报告，凭借在训练和推理工具链的全栈布局，以及领先的 AI 工程优化能力，2025 年阿里大模型训推公有云以 116%的增速，市场份额达到 42.2%；根据 Omdia，2025 年中国 AI 云市场总规模达 567 亿元，阿里云市场份额达 38.1%，其中 AI IaaS 总规模达 392 亿元，阿里云市场份额达 38.6%。

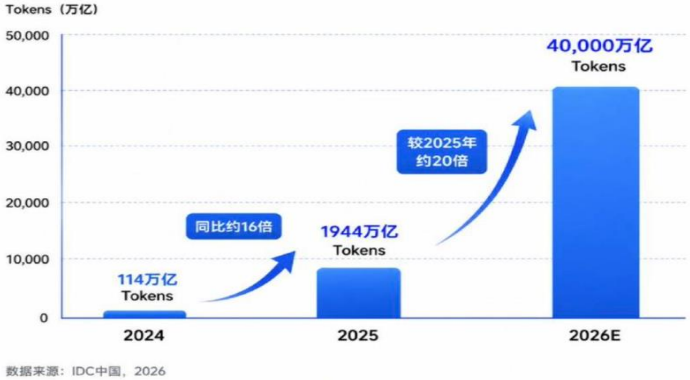
图表21：阿里云市占率位居国内 AI IaaS+MaaS 市场份额首位



资料来源：Omdia，阿里云，长城证券产业金融研究院

利润端：我们预计 GPU 云时代下国内云厂商利润率有望逐步提升。尽管由于无法自建网络的核心差异，国内云厂商与北美云厂商盈利能力仍有差距，但是我们认为国内云厂商利润率在GPU云时代具备明确的上行驱动力(1)国内云厂商加大高毛利MaaS及Agent业务布局，将 AI 应用深度嵌入企业业务流程从而提高客户迁移成本和平台粘性，同时通过 API 调用、模型服务订阅收费，突破单纯资源出租的价格约束，提升单位算力的商业价值；（2）加大海外市场布局，规模效应增强有望带动国内云厂商利润率提升，根据 Gartner，2025 年阿里云国际市场份额同比增长 0.5pct 至 7.7%，同时阿里云在 2026 云栖大会上宣布新一轮全球基础设施扩建计划：未来 12 个月，将于土耳其、芬兰和荷兰首次设立云计算地域节点（Region），并扩建马来西亚、德国、阿联酋、法国和中国香港的数据中心，进一步提升全球算力供给和本地化云与 AI 服务能力；（3）当前大模型正加速在企业端落地，但是大模型训练与推理需要大规模 GPU 集群、数据资源以及模型运维能力，因此普通企业客户自行建设基础模型基础设施面临较高的资本投入和技术门槛，因此大部分企业倾向于通过云厂商提供的MaaS服务获取AI能力，目前国内领先的MaaS平台包括字节火山方舟、阿里云百炼、百度云千帆等，企业用户可以通过 API 调用和模型托管方式向企业开放基础模型能力。根据 IDC，2025 年中国公有云大模型调用量已达 1944 万亿 Tokens，相比 2024 年同比增长约 16 倍，预计 2026 年全年企业级 Token 消耗量将增长约 20 倍，升至约 40000 万亿，MaaS 市场具备较高增长空间。

图表22：中国公有云大模型调用 token 消耗量将维持高速增长



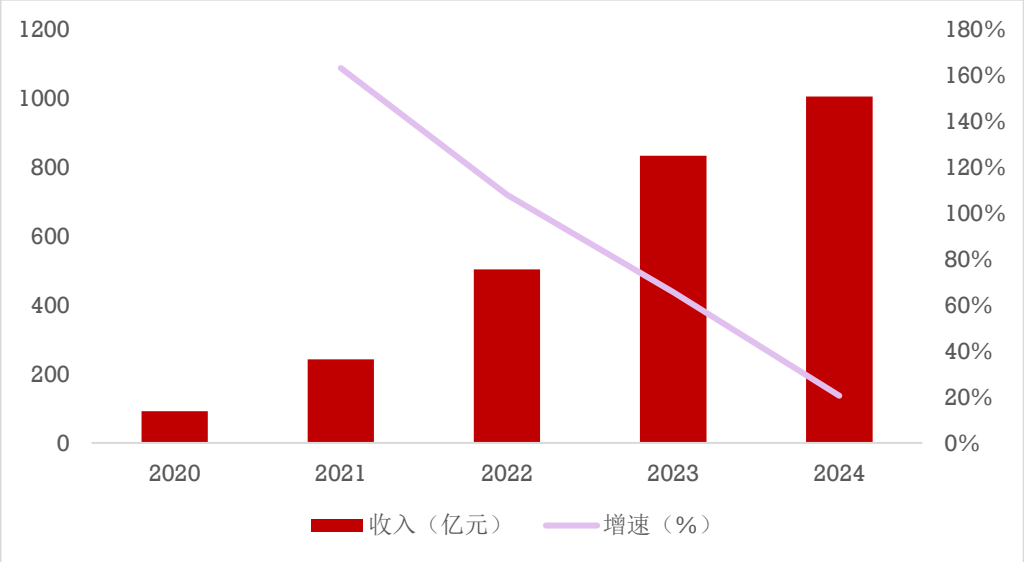
资料来源：IDC，长城证券产业金融研究院

2.3 国内运营商云：云业务快速增长带动估值中枢提升

2.3.1 云已经成为运营商第二增长曲线的核心部分

智算需求拉动移动云增长。2026 年上半年，移动云作为算力供给与智能应用的关键枢纽，着力释放中国移动特色的算力新动能，构筑客户场景驱动的智能新空间，让 AI 真正转化为新质生产力。移动云影响力进一步提升，公有云收入规模首次进入国内业界前三。算力服务是公司高质量发展的重要增长极，公司强化 AIDC 资源供给和储备，“AIDC+网络+算力+运维”服务能力明显提升，移动云品牌更加彰显。中国移动 2025 年数据中心收入同比增长 8.7%，其中 AIDC 收入同比增长 35.4%。得益于智算需求的迅猛增长，智算服务成为第一引擎，增速达 279%，拉动云算服务收入同比增长 13.9%。

图表23：2020-2024 年移动云收入及增速

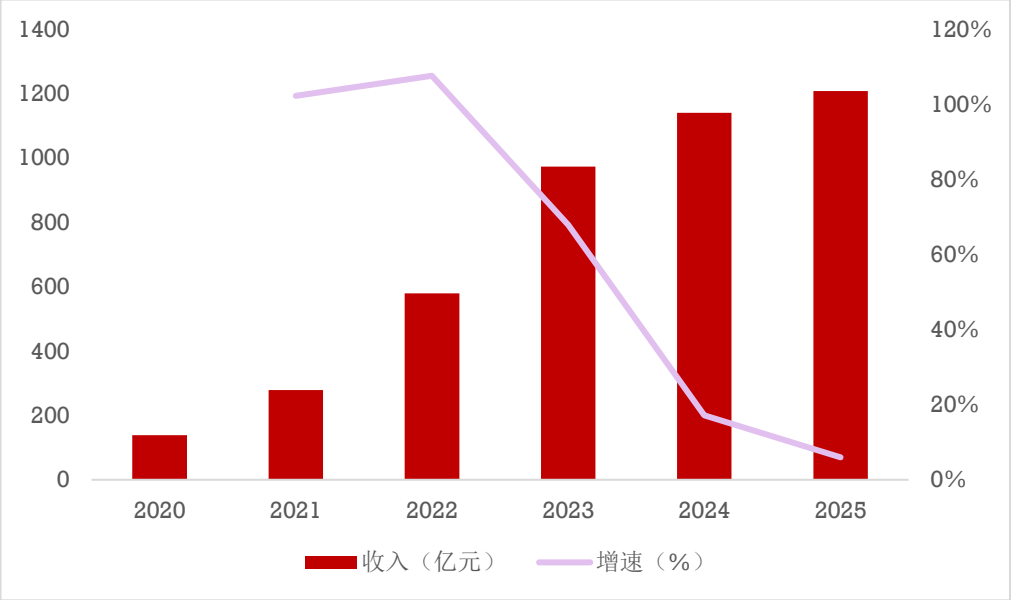


注：移动云 2025 年未披露收入

资料来源：中国移动 2020-2024 年年报，长城证券产业金融研究院

天翼云保持稳健增长并加快向“云智一体”升级。2026 年上半年，天翼云实现收入 618 亿元，同比增长 7.8%。中国电信持续强化自主可控云底座能力，自研操作系统 CTyunOS、数据库 TeleDB 等核心产品取得进展，并基于 TeleCloudOS 4.0 持续演进面向智能时代的 5.0 版本，增强大模型训练、推理及 Agent 承载能力。同时，天翼云深度参与“鹏城云脑”等国家级智算基础设施建设，持续强化云计算、智算调度及 AI 基础设施能力。整体看，天翼云正由传统云资源供给平台加速向“云+算力+大模型+Agent”的智能云平台演进，在运营商云市场中继续保持领先竞争力。

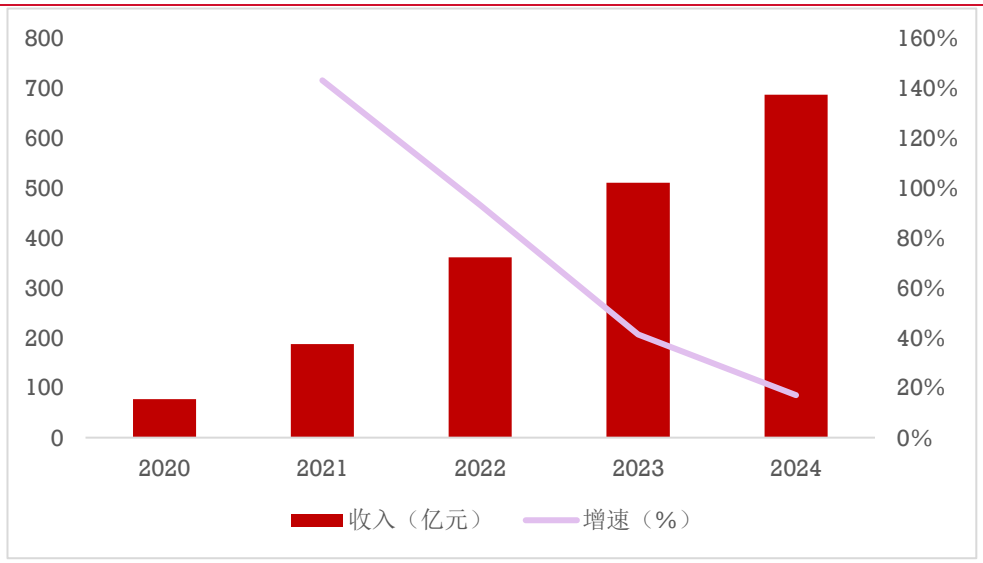
图表24：2020-2025 年中国电信天翼云收入及增速



资料来源：中国电信 2020-2025 年年报，长城证券产业金融研究院

联通云加速由传统云服务转型 AI 云与 Token 服务平台。2026 年上半年，中国联通持续完善"东数西算"算力布局，已形成超 115 万架标准机架、IT 总负载超 2900MW 的资源基础，并依托"联通星罗"算力平台构建全栈 AI 云能力与全域 Token 服务。平台侧，中国联通上线元景 MaaS 一站式 Token 运营平台，强化模型与 Token 分发、应用服务能力，同时以"鹏城云脑"等国家级智算项目提升国产智算技术能力；应用侧，推出 UniClaw 自主执行智能体，并逐步贯穿云智产品体系，推动个人、家庭及政企场景智能化升级。AI 云已成为公司强化算力价值变现、提升云智业务竞争力的重要抓手。

图表25：2020-2024 年联通云收入及增速



注 1：联通云 2025 年未披露收入

注 2：联通云收入口径为融合创新解决方案产生的云 IDC、云资源、云平台、云服务、云集成、云互联、云安全等收入。

资料来源：中国联通 2020-2024 年年报，长城证券产业金融研究院

我们回顾中国移动自 2022 年以来的股价表现，大致可划分为四个阶段：

第一阶段：估值重塑前的定价期，市场主要围绕运营商传统通信业务的稳定现金流与高分红属性定价，股价整体处于低位震荡；

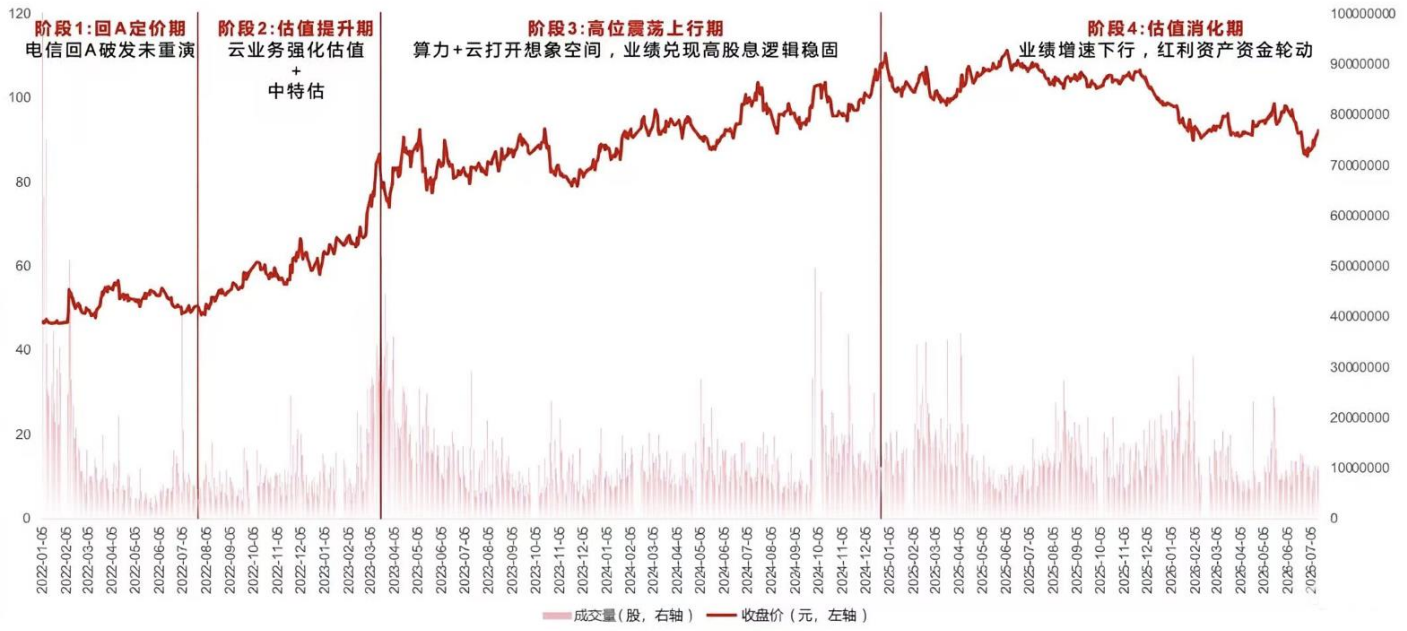
第二阶段：估值提升期，随着"中特估"催化叠加移动云等产业数字化业务快速增长，市场开始重新认识运营商由传统通信服务商向数字基础设施与科技服务商转型的价值，云业务成为估值抬升的重要增量逻辑；

第三阶段：高位震荡上行期，算力网络、云计算及 AI 基础设施持续打开成长空间，同时主营业务稳健、盈利兑现及高股息特征强化估值底部，推动"成长+红利"双重逻辑逐步形成；

第四阶段：估值消化期，在传统业务增速趋缓、产业数字化业务增速边际放缓以及红利资产内部资金轮动背景下，运营商估值进一步扩张对新增成长逻辑提出更高要求。

我们认为云业务是过去一轮以中国移动为首的三大运营商估值中枢由约 10-12x 向 15x 以上跃迁的重要产业变量之一，其意义不仅在于贡献新增收入，更重要的是推动市场对运营商的认知由"低增长、高分红的公用事业型资产"切换为"掌握网络、算力、云及政企客户入口的数字基础设施平台"。随着天翼云、移动云、联通云进一步向智算、MaaS、Token 服务及智能体平台延伸，云业务有望继续支撑三大运营商科技属性与成长性溢价。

图表26：中国移动上市至今分阶段走势（截止 2026 年 7 月 5 日）



资料来源：wind，长城证券产业金融研究院

2.3.2 运营商云或成为承担十五五规划重要角色

十五五期间政策持续加码算力建设。2026 年 3 月发布的《"十五五"规划纲要》明确提出，强化算力算法数据高效供给，适度超前建设新型基础设施，深入推进"东数西算"工程，构建多层次算力设施体系和全国一体化算力网。全国一体化算力网已被纳入国家"六张网规划"建设体系，并进入"十五五"109 项重大工程清单；同时政策明确要求推动云边端协同、智算云服务，降低全社会用算成本。2026 年政策进一步加码。2026 年政府工作报告提出实施超大规模智算集群、算电协同等新基建工程，加强全国一体化算力监测调

度，支持公共云发展。此外，工信部 2025 年 5 月印发《算力互联互通行动计划》，要求加快高性能路由器、高速无损网络技术研究，支撑高效入算、算力无损互联。

图表27："东数西算"八大算力枢纽及十大数据中心集群



资料来源：财经新商业公众号，长城证券产业金融研究院

三大运营商资本开支结构正加速由传统通信网络向算力网络、智算中心及 AI 基础设施倾斜。2026 年上半年，中国移动资本开支 610 亿元，同比增长 4.5%，其中算力网投入 156 亿元、同比增长 71.3%；中国电信资本开支 324 亿元，其中算力网投入 154 亿元、同比增长 97%，占资本开支比重提升至 48%，同比提升 24.6pct；中国联通上半年资本开支 241 亿元，算力投资占比提升至 37%，投资额同比增长超过 80%。其中，**中国移动和中国电信均提出将适度超前布局算力基础设施建设。**可以看到，在整体资本开支趋于理性、传统网络投资逐步进入成熟期的背景下，算力已成为三大运营商最明确的增量投资方向。我们认为，随着"十五五"期间全国一体化算力网、人工智能基础设施及"六张网"等新型基础设施建设进一步推进，运营商兼具全国性网络资源、云平台、智算中心、政企客户入口及安全可信优势，有望成为国家算力基础设施建设和 AI 产业落地的重要承担主体。移动云、天翼云、联通云亦有望由传统云资源提供商进一步升级为承接国家算力资源、模型服务、Token 分发及行业 AI 应用的综合数字基础设施平台，运营商云在"十五五"期间的战略地位及收入贡献有望进一步提升。

图表28：三大运营商 2026 年上半年资本开支情况



资料来源：中国移动投资者推介材料，中国电信微信公众号，中国联通 2026 中报，长城证券产业金融研究院

运营商作为算力建设国家队，依托网络优势，全域布局构建“东数西算+全球协同”的新型算力基础设施矩阵。中国联通围绕“东数西算”战略部署的京津冀、长三角、粤港澳大湾区、成渝、内蒙古、贵州、甘肃、宁夏八大国家算力枢纽节点，完成了资源布局的全覆盖，深度参与国家规划的 10 个数据中心集群建设。同时，依托网络优势，中国联通构建了低时延、广联接、智能化的一体化新型算力网络，通过“新八纵八横”骨干网连接算力资源，形成超大规模智算集群，实现了算力资源在云边端之间的有效连接和灵活调度。

图表29：中国联通粤港澳大湾区枢纽（韶关）数据中心



资料来源：中国联通公众号，长城证券产业金融研究院

图表30：中国联通宁夏中卫云数据中心



资料来源：中国联通公众号，长城证券产业金融研究院

国云国芯国模协同筑牢安全底座。天翼云持续推动国产软硬件从“可用”迈向“好用”，实现自主可控和安全可信。在国产化适配方面，天翼云一云多芯兼容多种国产芯片，是国内支持国产芯片种类最多的云平台。天翼云实现 DeepSeek-V4 发布后 0day 首日适配，智谱 GLM 系列模型完成全栈国产化闭环，验证了“国云+国芯+国模”战略布局可快速匹配市场需求。在安全保障方面，天翼云构建了覆盖接入、平台、内容、数据的全链路主动安全防御能力，实现 Token 全生命周期安全防护，系统性部署“增强算力系统整体韧性”的安全架构。

面向不同地区、行业、场景的算力需求，天翼云助力构建一体化算力网，实现算力资源的标准化供给与高效协同。打通公共、区域、行业三级算力互联网，天翼云衔接“东数西算”国家枢纽节点，赋能中小企业数智化转型与算力券政策落地，横向支撑重点行业端到端解决方案，形成算力协同格局。其中，息壤平台作为算力调度与监测能力的核心支撑，依托天翼云“4+4+31+X”云网资源布局，打通中心、省、边缘、端四级算力体系，实现通算、智算、超算等多元异构算力统一接入、调度和供给，助力北京、上海、长三角全域算力资源协同共享。

图表31：全球首个“海风直连海底数据中心”中国电信临港智算中心



资料来源：天翼云公众号，长城证券产业金融研究院

中国移动通过建立 Token 中心将 AI 商业模式由“卖算力”向“经营 AI 服务流量”延伸。中国移动云 Token 中心由中国移动云公司、中国移动苏州分公司联合苏州高新区政府共同建设，立足苏州、辐射全国，围绕 Token 生成、传输及应用构建全链条服务体系，重点布局算力互联、模型汇聚、应用赋能三方面能力，推动企业 AI 应用由分散对接算力、模型和工具，向统一供给、统一调度和一体化服务演进。

Token 中心是中国移动围绕 Token，构建连接算力—模型—智能体—产业应用的综合服务枢纽。算力侧，依托算网大脑统一纳管和调度异构算力资源，提升算力利用率及 Token 稳定供给能力；模型侧，依托 MoMA 模型服务网打造多模型统一调用入口，目前已汇聚超 340 款主流模型，覆盖文本、语音、图像、视频及多模态场景，并与火山引擎、阿里千问、智谱、MiniMax、Kimi 等模型厂商开展合作；应用侧，通过“算力+平台+工具+服务”一站式能力，开放智能体框架及 300 余款模型工具链，推动 AI 从实验验证向工业等实际生产场景落地。

图表32：中国移动云 token 中心及苏州分公司生态联盟揭牌



资料来源：移动云公众号，长城证券产业金融研究院

3. Token 价格下降带动 AI 应用起量，关注 FDE、AI4S、AI 办公、多模态等领域投资机会

3.1 参照移动互联网时代，Token 价格下降有望带动 AI 应用生态繁荣

以史为鉴，上网资费大幅下降为移动互联网时代应用爆发以及用户快速增长基础。复盘移动互联网时代，上网资费大幅下降带动 DOU 快速上涨以及移动互联网应用（如游戏、直播、短视频等）的快速发展。具体而言，根据工信部历年通信业统计公报、中国信通院《宽带资费水平报告》等报告，2014-2017 年，三大运营商流量资费均价分别为 134.7/78.1/47.9/28.0 元/GB；2018 年，根据中央网信办公开数据，手机上网流量价格平均为 8.5 元/GB，较 2014 年下降 94%。目前，以中国移动为例，动感地带芒果卡 139 元档国内通用流量为 55GB，价格约为 2.5 元/GB，较 2014 年下降约 98%。上网资费的大幅下降带动了移动互联网应用以及内容生态的蓬勃发展，游戏、短视频、直播等相关行业规模均实现快速增长，移动互联网 DOU 亦快速攀升。

Tokens 价格类比上网资费，人均调用量类比 DOU，Tokens 生产成本已经处于快速下降阶段。尽管 GPU/TPU 等芯片仍然有较为明显的代际涨价，但是被吞吐量(即每张 GPU 每秒能产出的 Tokens 数，Tokens/sec/GPU)的更大幅度提升所抵消并反超，带动 Tokens 生产成本下行。

图表33：通过软件优化可大幅提升 GPU 吞吐量



资料来源：semianalysis，长城证券产业金融研究院

图表34：硬件升级可带动 GPU 吞吐量进一步跃升



资料来源：semianalysis，长城证券产业金融研究院

智能体应用渗透率快速提升，带动 Tokens 混合均价持续下行。2025 年 12 月为 AI 智能体拐点，Agentic AI 不再只局限于编程场景。例如在金融场景下，分析师可以用 agent 完成各种任务，把 Excel 模型转换为 dashboard、自动生成图表、撰写报告、搭建财务模型、分析上市公司财报等，LLM 已经完成了由对话式机器人向生产工具的转变。目前智能体使用率处于快速提升阶段，Agent 可以连续工作 4-8 小时而不需要人干预。而智能体的快速普及也迅速带动了 Tokens 混合均价的持续下行，主要因：

- 极高的输入输出比：以 Claude Code 为例，其智能体工作 IO 比例为 300:1，而传统聊天场景该比例为 1:1-1:3；Tokens 输入价格远低于 Tokens 输出价格；
- 极高的缓存命中率：拥有 90%以上的缓存命中率，缓存命中的输入 Tokens 成本仅为未命中成本的 1/10 或更低；

以上两大原因带动 Tokens 的混合均价持续下行，以 Opus 4.7 为例，在执行智能体任务

时，运行 Opus 4.7 的实际混合价格仅为 0.99 美元/百万 token；远低于其显示的输入价格（$5/百万 token）、输出价格（$25/百万 token）。

图表35：AI Agent 与人工成本对比

Token Spend vs Labor Cost on Real SemiAnalysis Workflows

Task	Token Cost	Human Cost	Human Time Taken	ROI (Human Cost / Token Cost)
Chart Cursor vs Anthropic ARR Growth (past 4 months)	$4.66	$50	1 hour	10.7x
Build tokenomics disclosure Slack bot	$58.02	$1,000	20 hours	17.2x
Set up DigitalOcean droplet for Terminal Bench	$35.97	$500	10 hours	13.9x
Initiate on Hewlett Packard Enterprise: roadmap, balance sheet, and capex sustainability	$21.33	$1,000	20 hours	46.9x
Perfomed keyword analysis on, and thematically organized full OCP 2024 conference	$16.69	$450	9 hours	27.0x
Search past conferences for CPU trends, demand, pricing, and attach ratios	$7.61	$500	10 hours	65.7x
Marvell earnings recap with CXL and CPO updates, matched to prior TEL recap format	$2.82	$200	4 hours	70.9x
Search past conferences for optical networking and optical circuit switching trends	$8.20	$500	10 hours	61.0x
Pull 5 years + YTD financials, email results with Excel tables attached	$1.87	$150	3 hours	80.2x

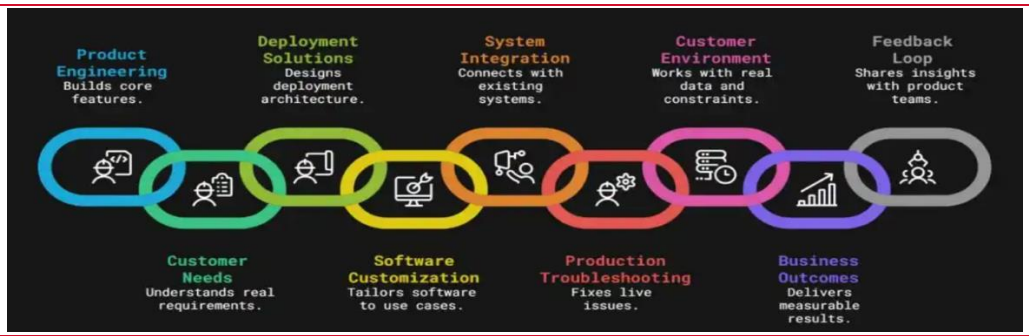
资料来源：semianalysis，长城证券产业金融研究院

3.2 FDE：加速模型在产业端落地，完成 AI“最后一公里”交付

大模型能力持续提升，但大部分企业没有进行业务端 AI 进行实际落地。根据腾讯研究院公众号，尽管大模型的能力每季度都在跃升，但超过 60%的企业 AI 试点仍然停在实验阶段。并非是模型不够强，而是 Demo 到生产之间有一条很宽的 gap，包括系统对接、权限梳理、流程改造、合规审计、组织适应，同时最重要的是需要让客户明确到底需要 AI 做什么。

FDE：Forward Deployed Engineer，即前线部署工程师，正助力模型厂商加速企业端落地。FDE 中文名称为前线部署工程师，FDE 并不等同于“更懂 AI 的实施工程师”。FDE 需要真正关注的是团队在项目结束后是否有通用知识及能力的沉淀，项目完成后如果只留给客户一个系统，其实和传统外包没有本质区别；如果带回项目经验但无法复用，则是项目制交付；FDE 的本质是在严格去掉客户信息后，把行业通用的经验沉淀为 Skill、模板和产品能力，沉淀后显著降低下一个同类客户的交付成本，即为可规模化的 FDE。其实 FDE 可以简化为“双向沉淀”机制。面向客户，FDE 把散落在员工脑中的业务经验和流程规则转化为 AI 可调用的知识库、工作流和应用能力；面向平台，FDE 把行业场景、系统接口、测试方法和失败案例沉淀为可复用的产品组件，持续迭代产品。两层沉淀的共同载体，是本体层：结构化的业务知识图谱，让同一个行业的不同客户能够共享对业务对象的理解方式。

图表36：FDE 主要职责梳理



资料来源：新浪财经，数字化企业，长城证券产业金融研究院

FDE 项目成本随客户数量增长而持续下降。根据腾讯研究院公众号，FDE 首次项目成本高于传统项目制，因为除了交付客户系统，还要额外完成知识抽象和验证。但是随着同类客户数量持续增长以及本体层稳定，后续交付成本将显著下降。FDE 模式通常由 Echo 和 Delta 双角色构成（1）Echo-分析师/策略师：侧重业务深度，负责需求挖掘、场景定义、价值评估及客户关系维护，背景多为行业专家或产品经理；（2）Delta-工程师：侧重技术速度，负责基于 Echo 的定义快速编写代码、搭建原型、集成系统及现场调试。目前，AI 编程、低代码平台和智能体工具正在大幅压缩 Delta 成本：原型搭建、系统对接、测试验证不再需要大型工程团队。但 Echo 不仅没有变便宜，反而更加稀缺。客户最大的瓶颈正在从“能不能搭”转向“知不知道搭什么”：业务部门有痛点却难以转化为 AI 任务，IT 部门懂技术却离业务太远，决策层想推 AI 却不知道组织应该如何调整。因此我们认为“自然语言构建应用”门槛越低，FDE 的价值反而越上升：从搭建者转为场景发现者、需求翻译者和方法教练。我们认为 FDE 未来会明显分化：低端实施和重复搭建被 Agent 和平台持续压缩，场景定义、客户信任、组织推动和行业知识沉淀会持续升值。

3.2.1 迈富时：FDE 带动公司 AI 应用业务步入爆发周期

陆续推出 AI-Agentforce 3.0 智能体中台、Knowforce 知识中台、OntologyForce OS AI 操作系统等产品，由 SaaS 企业向原生 AI 应用平台转型：

- **AI-Agentforce 3.0 智能体中台：**根据公司微信公众号，AI-Agentforce 3.0 主要有以下三大亮点（1）**技术升级**：创新性引入了智能体元编程与统一标准能力层技术，用户只需用自然语言描述业务需求，系统即可自动识别意图、推荐任务所需的统一标准能力工具、自主拆解任务、生成执行指令、自动创建智能体。进一步提升了开发企业级 AI 智能体应用的效率，人机交互从“低代码、图形化”向“无代码、自然语言”升级，实现了 AI 智能体应用开发“言出智行”的新范式；（2）**协同能力升级**：AI-Agentforce 3.0 搭载了先进的多智能体协作系统（MAS），该系统突破了单一智能体的能力局限，支持并行、层级、监管等多智能体协作模式。通过灵活可控的思维流编排与调度多智能体，实现对长链条、复杂多任务流程的高效处理，显著提升 AI 智能体解决企业实际问题的能力；（3）**安全架构升级**：通过多租户资源隔离、精细化权限管控及全链路风险监测，构建起“软硬一体、安全可控”的数据防线。支持云上、本地及混合部署模式，确保数据主权与业务合规，为大型企业的智能化转型筑牢根基。

图表37：迈富时 AI Agentforce 3.0 产品优势



资料来源：迈富时公司官网，长城证券产业金融研究院

- **Knowforce 知识中台**：传统知识管理系统往往停留在“电子仓库”的定位，只能用作存储功能。根据公司微信公众号，迈富时 KnowForce AI 知识中台的突破在于它不仅能存储，更能理解、连接和主动响应。平台全面支持文本、图片、音频、视频等多模态知识处理，用户只需自然语言提问，系统即可精准理解意图并快速定位答案。在此基础上，系统通过自选大模型自动提取文档中的人物、项目、技术点，生成可视化知识图谱，让隐藏的业务关联一目了然，帮助员工快速把握复杂项目的全貌。同时，平台新增结构化领域知识能力，支持将行业标准、技术规范、业务流程等高价值内容构建为可机读、可推理的结构化知识库，进一步支撑复杂场景下的精准决策与智能推理，帮助员工快速把握复杂项目的全貌。

图表38：迈富时 Knowforce 知识中台产品优势



资料来源：迈富时公司官网，长城证券产业金融研究院

OntologyForce OS：根据公司微信公众号，该系统定位为本体驱动 AI 操作系统。它不是一个单一的 AI 应用，而是一个承载企业所有 AI 能力的语义底座，打破“孤岛式 AI”，实现从“人充当集成层”到“AI 原生驱动”的代际跨越。尽管目前模型能力持续提升，但是无法理解公司业务。例如公司销售提问“华东区 Q1 战败线索中哪些可以通过跨品

牌转派盘活？"，大模型无法直接对问题进行回答，因为它不知道什么是"战败线索"，不理解"跨品牌转派"的业务规则，更无法穿透公司 CRM、DMS、CDP 三套系统拉取数据做交叉分析。

OntologyForce OS 对标 Palantir AIP，打造企业级 AI 平台。传统的“+AI”模式下，AI 能力被局限在各个独立的 SaaS 系统中。根据公司微信公众号，OntologyForce OS 作为一个操作系统 (OS)，在底层实现了模型中立架构。它向下兼容 GPT、Claude、Qwen、DeepSeek、豆包等国内外主流模型，向上则通过统一的语义层为企业提供一致的智能化支持。OntologyForce OS 可以有效消除 AI 孤岛，将 ERP、CRM、DMS 等异构系统的数据，通过本体论映射为互联的“数字有机体”。同时可以对其语义统一性，确保 AI 在不同业务环节对客户、订单、资产等核心概念拥有完全一致的上下文理解。同时，公司配套推出 OAG（Ontology Augmented Generation）推理引擎，作为 OntologyForce OS 的大脑。不同于简单的 RPA 或规则引擎，OAG 引擎具备多跳推理能力，是对传统 RAG 的全面超越。它能基于 Auto-Ontology 技术自动从历史数据中提取知识，并在业务逻辑边界内进行自主规划。这意味着 AI 不再是根据固定脚本运行，而是根据实时业务上下文，自主决定下一步的“行动路径”。

图表39：迈富时 OntologyForce OS 助力消除数据孤岛



资料来源：迈富时微信公众号，长城证券产业金融研究院

迈富时已经完成 FDE 团队搭建，搭配完整的产品矩阵，推进面向客户的 AI 能力交付。公司已经完成 FDE 团队搭建，并完成 AI 能力交付。根据公司微信公众号，公司向便携消费电子制造龙头交付 AI+CRM 产品，客户希望把海外销售的聊天记录沉淀下来，但是客户信息在 WhatsApp 里，报价规则在销售脑子里，审批在 OA 里，订单和库存又在 ERP 里。公司工程师通过观察销售真实跟单过程，从线索、询盘、报价、审批一路画到订单和库存，再把三套系统里的字段逐一对齐。比如哪些对话代表客户意向，哪些报价要触发审批，哪些历史订单能反推客户分层。FDE 直接把上述流程做进原型，借助 AI Coding，团队可以按天迭代 Demo，而传统外包模式下则需要重新评审、重新排期、重新报价，流程较长；而 FDE 在现场，知道上下文，也能马上改代码，有效节省沟通成本，同时帮助客户进行销售数据、客户资产、报价逻辑和经营规则的沉淀。

图表40：迈富时 FDE 案例



资料来源：迈富时微信公众号，长城证券产业金融研究院

3.2.2 神州控股：AI First FDE

神州控股将 AI First FDE 作为“Data × AI”与“AI for Process”战略的工程化落地抓手，推动 AI 项目由传统“先建平台、再找场景”转向以业务问题为起点、以结果验证为导向。FDE 深入客户真实业务流程，通过大模型、数据工程与智能体技术快速构建可运行原型，并将项目中形成的行业 Know-how、业务规则及流程语义沉淀为可复用数字资产。

公司 FDE 能力主要体现在三方面：

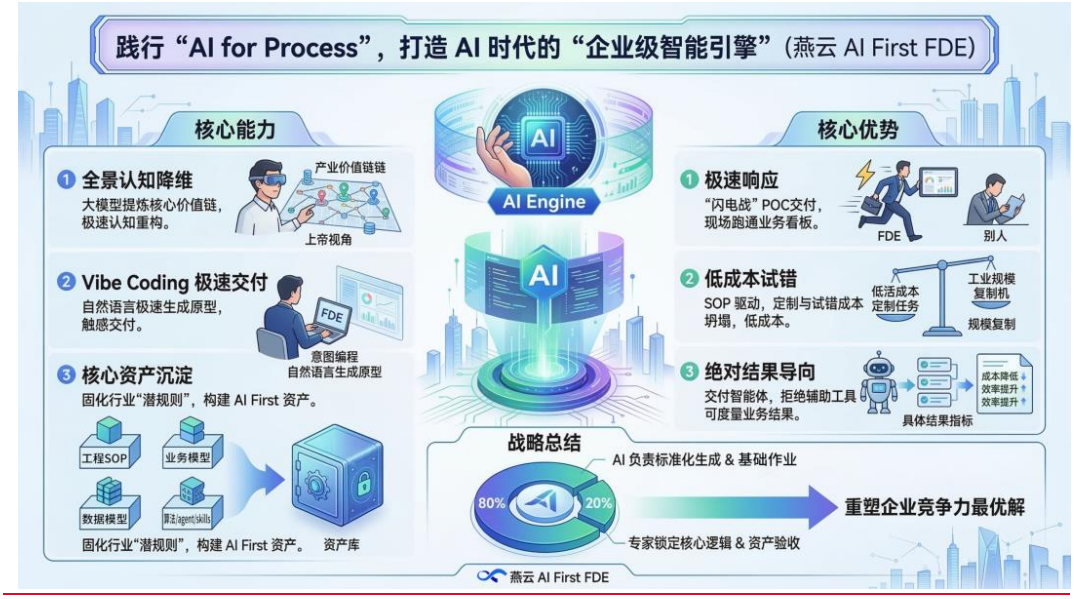
1.全景认知降维，借助大模型推理快速识别行业核心价值链，显著压缩前期调研周期；

2.Vibe Coding 快速交付，通过自然语言编程快速形成可交互原型，提升项目验证效率；

3.核心资产沉淀，将非标准化行业经验固化为模型及数字资产，提升后续场景复制能力。

相较传统企业软件主要交付系统及工具，燕云 AI First FDE 更强调交付可完成端到端业务任务的数字智能体及可量化的最终业务结果，包括降本、提效及决策自动化等指标。项目分工上，AI 承担约 80%的标准化工作，专家聚焦 20%的核心业务逻辑，由此形成响应速度快、试错成本低、结果导向强三方面优势，并有望提升 AI 项目交付效率与规模化复制能力。

图表41：神州控股燕云 AI First FDE 介绍



资料来源：神州控股公众号，长城证券产业金融研究院

传统 IT 交付是一个漫长且昂贵的线性交付过程。神州控股的 AI First FDE 范式依托 Infinity 底座，通过“认知前置”和“Vibe Coding”“敏捷验证，将交付周期提速了 5-7 倍，整体运营效率跃升 30-50 倍。公司的 AI First FDE 先通过 AI 能力快速获取行业知识，将行业认知前置掌握 know-how。第二阶段，用 Vibe Coding 进行快速开发。用客户的真实数据，做出可演示、可交互的系统原型。原型验证通过后，将原型沉淀在燕云 Infinity 平台上，进行工程化的交付。在过程中持续沉淀指标、数据，形成一个个可被管理的标准产品，反哺持续迭代升级。AI First FDE 最核心的价值在于，FDE 每完成一次交付，都在平台上沉淀下了可复用的行业数据资产。公司在服务同类客户时的边际研发成本大幅降低，在行业内实现规模化扩张和利润不断提升。

图表42：神州控股燕云 AI First FDE 模式



资料来源：神州控股业绩推介材料，长城证券产业金融研究院

标杆案例已开始验证"快速诊断+MVP+业务闭环"模式。

园区场景：AI 重构仓储履约流程。公司旗下子公司科捷与海蒂的花园开展「xAI·供应链」生态共创，针对大型园区拣货距离长、人工凑单依赖经验及复杂路况等痛点，搭建仓储监控、物流及 AI 调度三大看板，并打通波次生成、任务分配、AGV 执行及复核监控全流程。项目推动备货决策由人工经验向"AI 建议+人工确认"转变，库存管理由事后盘点向实时预警升级，跨环节协同进一步实现系统自动触发。

校园场景：FDE 显著压缩 AI 应用验证周期。针对校园网络运营客户 C 端精细化运营需求，FDE 团队 2 天内完成需求确认及可演示系统搭建，首版 POC 投入约 5-6 人天，构建 AI 用户防流失运营平台。系统可对全量用户自动打标及风险评分，全量处理耗时约 13.6 分钟，并通过干预前后快照量化挽回率及 AI 增量率。项目体现 FDE 由传统售前实施向业务咨询、产品设计与智能体开发一体化角色延伸，有助于提升 AI 项目交付效率及场景拓展能力。

3.3 AI4S：AI 加速"设计-验证"循环，市场空间广阔

AI for Science（科学智能）是指利用人工智能技术和方法来加速科学研究和发现的过程。科学智能（AI for Science，AI4S），就是借助 AI 技术加快科研探索与新发现的产出。和侧重信息处理、内容生成的通用大模型不一样，AI4S 的核心定位是挖掘科学规律、提升研发效率。一方面，它可以从海量高维数据里挖掘潜在机理，强化对分子、材料、生命、自然体系的理解与预测；另一方面，把模型预测融入"假设提出-计算筛选-实验验证-数据回流"的完整科研闭环，以更少时间与成本产出更优质科研成果。过去科研大多依靠 AI 处理已有的实验历史数据，而 AI4S 还能够融合物理、化学等底层科学原理，完成仿真、运算与数据生成。即便是数据匮乏、未知范围广的研究场景，也可以提供新的分析思路与验证手段。在 AI4S 推动下，全球科研进入第五范式，科研工作由数据辅助模式升级为智能系统工程。AI 不再只是简单辅助工具，而是完成角色跃迁，全链路重塑科研流程，真正实现从预测到落地实践的转变。

图表43：科研研究的演进范式

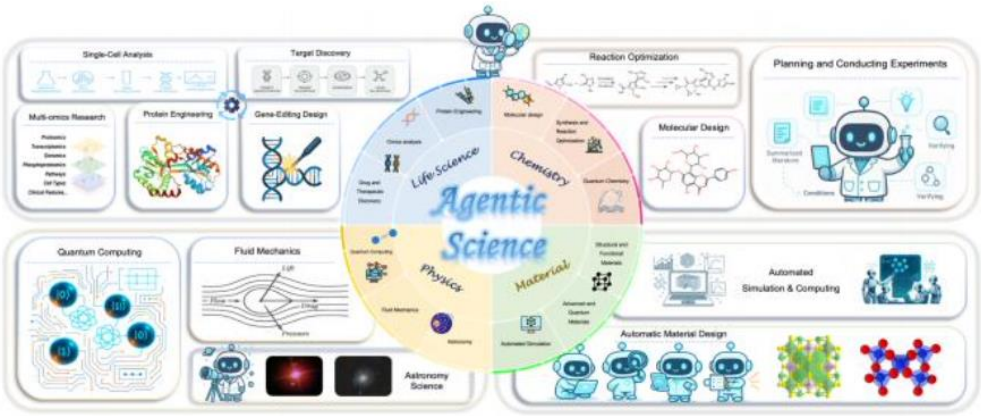


资料来源：头豹研究院，长城证券产业金融研究院

AI4S 应用领域广泛，已经在化学、生命科学、物理学等方向承担相应科研任务。AI4S 拥有十分广阔的落地场景，现已在化学、生命科学、物理等诸多学科落地科研工作。具

体来看：生命科学领域可开展组学解析、靶点挖掘、蛋白质改造以及基因编辑方案设计；化学方向能够实现分子生成、量子化学运算和化学反应调优；材料行业可加快材料性能预判、仿真模拟与新材料开发；物理领域则赋能量子计算、流体力学、天文观测的数据解析。伴随着模型能力、数据集以及实验设施不断迭代，AI4S 的应用上限还会继续拓宽。它的内核是把传统以实验为主导的科研问题，转变为计算与数据共同驱动的求解模式，理论上能够覆盖一切可建模、可数字化的科研场景。这种底层方法论的革新，说明 AI4S 不只是某个细分赛道的机遇，更是带有平台属性的技术变革。

图表44：AI4S 具备广泛的应用领域



资料来源：《From AI for Science to Agentic Science:A Survey on Autonomous Scientific Discovery》，长城证券产业金融研究院

在持续深耕 AI for Science 的科技企业引领下，以生物学为代表的科学智能正在进入一个高产出、快迭代的应用落地期。行业有望步入涌现阶段，形成规模化格局。随着资本注入、技术成本下降、AI Infra 的能力提升，多个项目正从单点原型走向平台化应用落地。当前 AI4S 细分赛道中生命科学较为成熟，材料学、物理学、气象学、计算机和数学紧随其后。其中在生物医药行业，2024 年 Google DeepMind 参与开发 AlphaFold 的两位科学家凭借在蛋白质折叠预测上的突破获得诺贝尔奖，标志 AI 在生命科学关键问题求解中的科研价值获得顶级学术认可。

图表45：AI4S 应用领域具体案例梳理

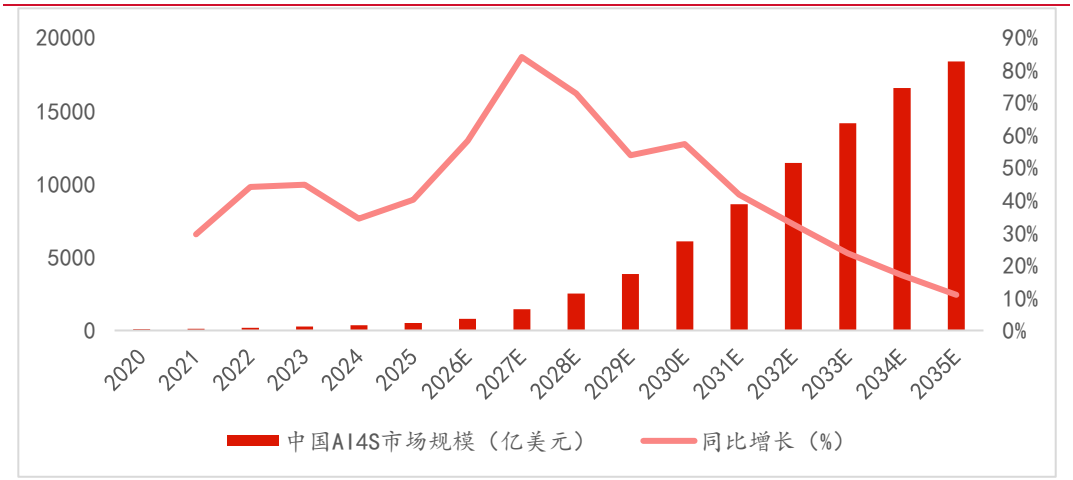
领域	案例以及成就
AI+生命科学	谷歌和耶鲁大学联合发布 270 亿参数单细胞分析基础模型 C2S-Scale，生成了关于癌细胞行为的全新假设，并在多次体外实验中得到验证。这展示了利用 AI 提出原创科学假设的潜力，有望据此探索出开发抗癌方法的新途径
	微软 BioEmu 模型在蛋白质动力学模拟方面填补了空白，并实现了高达 10 万倍的模拟速度提升；中科院团队提出整合结构和进化约束的反向折叠蛋白质预测模型，为蛋白质工程开辟新路径，相关成果发表于《Cell》期刊
	谷歌通过 10 年持续研发和探索，逐步构建了从基因测序、读取和变异识别，到基因表达预测和致病潜力评估，再到疾病基因检测和诊断的 AI 基因组学研究和应用体系。
AI+医疗	腾讯生命科学实验室与广州医科大学第一附属医院、广州呼吸健康研究院联合研发的 DeepGEM 病理大模型，已在肺癌基因突变预测中完成大规模验证，只需常规病例切片图像，1 分钟内完成肺癌基因突变预测，精准度达 78%~99%
	谷歌发布 DeepSomatic 工具集用于肿瘤细胞中基因变异的识别，适用于白血病、乳腺癌、肺癌等癌症类型，识别准确率均优于现有方案

	北京大学第三医院等多家医院与剂泰科技联合进行 AI 优化候选药物 MTS-004 成为国内首款完成Ⅲ期临床的 AI 赋能制剂新药，药物预计面向渐冻症、脑卒中等神经系统疾病
AI+材料学	前 OpenAI 和 Deepmind 成员创立 Periodicabs 开展新型超导材料等 AI 自动化发现。孵化自北大深研院深港河套科创中心的鼎犀智创正构建 RhinoWise 材料创新平台，开展新能源、半导体等领域关键材料创新
AI+物理学	DeepMind 的飓风 AI 模型已成功预测“梅利莎”等超强飓风的路径和强度变化帮助美国及周边国家提前预警。黑洞理论物理学家 Alex Lupsasca 利用 GPT-5 在半小时内推导出黑洞理论新特性。核聚变初创公司 CFS 利用谷歌开源 TORAX 工具来辅助研发 SPARC 核聚变装置
AI+数学	数学研究人员利用 GPT5 探索解决历史数学难题-埃尔德什难题，谷歌基于 AlphaEvolve 推进数学和理论计算机科学研究，英伟达开源模型系统 GenCluster 获得 IOI 2025 竞赛金奖

资料来源：腾讯云开发者平台，长城证券产业金融研究院

AI4S 市场空间广阔，有望成为继 Coding 后下一重要增量方向。2026-2035 年，国内外 AI4S 市场有望维持高速扩张，逐步成长为 AI 产业里举足轻重的应用分支。全球科研相关的智能工具、平台及服务的投入不断提升，推动行业规模持续走高。根据弗若斯特沙利文的数据，全球科学智能市场由 2020 年 378 亿美元，有望增长到 2035 年接近 5 万亿美元，15 年增幅超 130 倍；2025-2030 年复合增速达到 60.5%，属于行业高速发展窗口。生命科学生物产业、高端制造将是市场的两大核心驱动力：前者得益于 AI 赋能新药和基因技术研发，后者则紧跟全球工业智能化的发展大势。国内市场同样迎来高速发展，2020 到 2035 年规模预计从 98 亿美元增长至 1.8 万亿美元。整体而言，全球 AI4S 已经告别早期零散试点阶段，进入以生命科学、高端制造为核心的集中爆发期，也是未来科技产业竞争的重要底层支撑。

图表46：AI4S 具备广泛的应用领域



资料来源：《科学计算与 AI4S 白皮书》上海人工智能协会（2026），长城证券产业金融研究院

从全球头部云厂商 AI4S 布局来看，云厂商正从“算力底座”向上层“AI4S 应用平台”全面渗透。与早期 DeepMind、Meta 以开源模型为主、聚焦单一结构预测任务不同，AWS、Oracle、IBM 等云厂商不再满足于提供 GPU 算力或通用 AI 服务，而是直接推出面向生物医药等领域的专用平台，这些平台内嵌了分子生成、药物筛选、临床数据分析、实验设计自动化等科学任务。例如，IBM WatsonX 同时支持模型训练与 Agent 自动化科研流程，Microsoft Discovery 则整合多智能体协作完成从假设生成到实验模拟的完整研发闭环。Agent/Agentic 工具将促使云厂商用“托管式科研自动化”替代传统单点软件工具，行业竞争焦点从模型精度本身转移到能否真正取代手工操作流程。AWS、Oracle、

Microsoft、IBM 均采用企业级 PaaS、私有化部署或云原生行业应用的收费模式，目标客户明确为药企、材料公司或高校科研机构。云厂商正在生命科学、材料设计等 AI4S 高价值场景中构建原生平台，部分任务将对传统仿真与工业软件形成替代压力。但由于 CAE/EDA 仍深度嵌入工程验证与生产流程，短期内将呈现“局部替代+平台竞合”的关系。

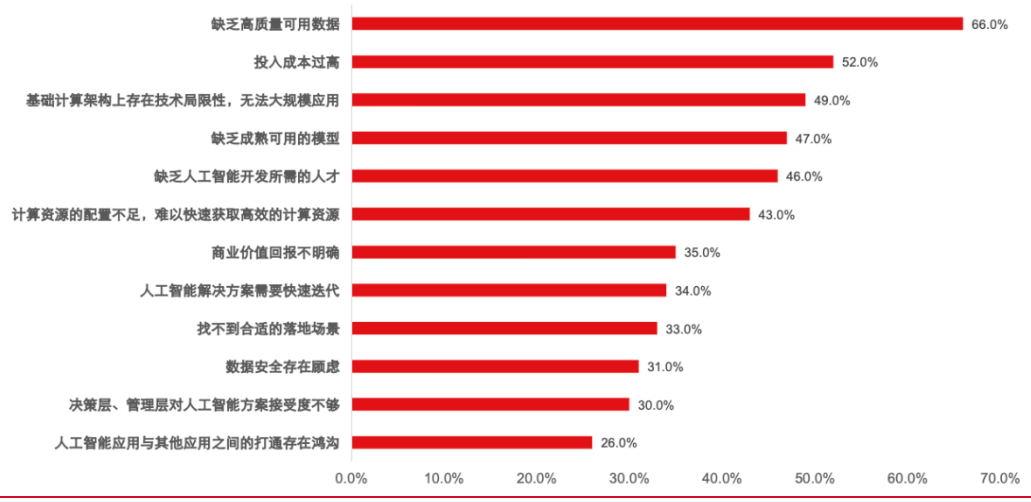
图表47：国际主流 AI4S 平台产品矩阵

牵头主体	平台/模型	应用领域	产品类别	主要任务	产品形态	商业化路径
Google DeepMind	AlphaFold	生命科学	科研大模型	蛋白质结构预测	论文+开源模型/代码+数据库	云端科研平台
	GNoME	材料科学	科研大模型	稳定晶体发现/DFT 数据飞轮	论文+开源模型/代码+数据库	云端科研平台
Microsoft	MatterGen	材料科学	科研大模型	稳定无机材料生成、材料结构设计	数据框架（训练/微调）+推理微服务	云端科研平台
	Microsoft Discovery	通用	Agent/Agentic 工具	整合图知识库+多智能体协作，支持假设生成、文献分析、实验模拟与迭代学习的完整研发闭环	企业级 PaaS 平台/支持私有化部署	云端科研平台
Meta AI	ESM 系列	生命科学	科研大模型	蛋白语言模型、单序列结构预测、宏基因组结构库	开源模型/代码	AI 基础设施+企业软件许可
AWS	Amazon Bio Discovery	生物医药	科研大模型+Agent+科研平台	分子生成、药物筛选、实验设计自动化、科学计算加速	云原生 AI 平台+行业应用平台	云端科研平台
Oracle	OCI Data Science / Life Sciences AI Data Platform / AI Data Platform	生物医药、医疗、工业仿真、数据科学	AI4S 科研平台+算法平台	数据整合、建模训练、临床数据分析、分子模拟	数据+算力一体化平台	云端科研平台
IBM	Watsonx	医疗、材料、金融、工业优化	科研大模型+Agent/Agentic 工具+AI 开发平台	模型训练、科学数据分析、Agent 自动化科研流程	AI 开发平台+Agent 平台+数据框架	AI 基础设施+企业软件许可

资料来源：头豹研究院，长城证券产业金融研究院

此前，AI4S 因数据匮乏以及模型能力限制导致落地困难。过去 AI4S 商业化落地存在阻碍，主要来源于数据供给不足与模型本身能力的局限。其中数据短缺是最主要痛点，AI4S 任务需要针对特定科研场景搭建高质量、结构化、可核验的数据集，但适配 AI 训练的科学数据供给存在明显短板：（1）正负样本分布不均衡：科研场景中，被记录、公开的大多是实验成功的结果，大量失败案例、无效尝试以及实验中间环节的数据没有留存，造成模型无法充分学习完整的科学机理；（2）对数据标准化与可验证性门槛较高：用于 AI 训练的科学数据需要附带实验环境、测试手段、各类参数等完整背景信息，还要完成标注校验。但现有科研数据格式杂乱、质量差异大，无法直接拿来训练模型；（3）数据孤岛现象显著：高价值实验资料分散在各大高校、科研院所与企业手中，受数据权属、知识产权以及共享机制约束，很难形成大规模的公开数据集。除数据之外，早期 AI 模型性能同样存在短板。虽然自 2023 年起大模型在知识吸收、信息处理上进步明显，但处理复杂科学推理任务依旧存在短板，仅在代码领域实现较为突出的效果，各类评测基准也反映出科研推理能力仍有待加强。以博士级别难度的 GPQA 科学测试集为例，该基准 2023 年 11 月推出之时，GPT-4 得分仅 39%，远低于 70%的人类专家基准水平。

图表48：缺乏高质量数据为 AI 商业化落地最重要的卡点



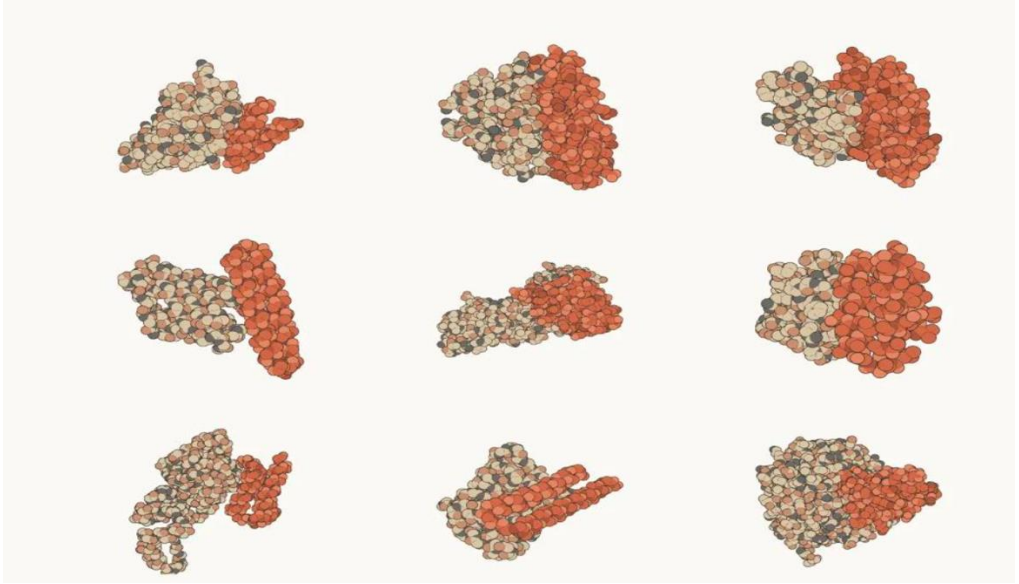
资料来源：IDC，长城证券产业金融研究院

目前，以上两大主要难题正在逐渐被攻克：

（1）模型能力持续突破，前沿模型在生命科学领域继续摸高

以 Anthropic 为例，Claude Opus 4.5 在科学图表解析、计算生物学、蛋白相关分析等生命科学任务上实现明显提升。该公司 2026 年 6 月发布 Mythos5/Fable 5，进一步增强生命科学相关能力，药物设计工作效率也得到改善（1）药物设计部分环节效率有约 10 倍提升。模型可以承接多项科研人员常规工作，包括定位蛋白结合位点、调用对应的蛋白设计工具，同时处理各类异常问题。相关测试中，由模型生成的 14 个蛋白靶点里，有 9 个得到具备开发潜力的药物候选分子。(2) 可以产出新颖、逻辑自洽的科学假设。在 1001 组对比测试中，Anthropic 内部科研人员 80%的场景更认可 Mythos 输出的假说。（3）在基因组学方向完成创新性探索。Mythos 5 在一周的自主运行周期内，收集 138 种动物数百万细胞的单细胞数据，并自主搭建训练专用机器学习模型。该模型体量仅为《科学》刊载同类模型的 1%，却取得更优效果，整个过程仅需要少量人工介入。

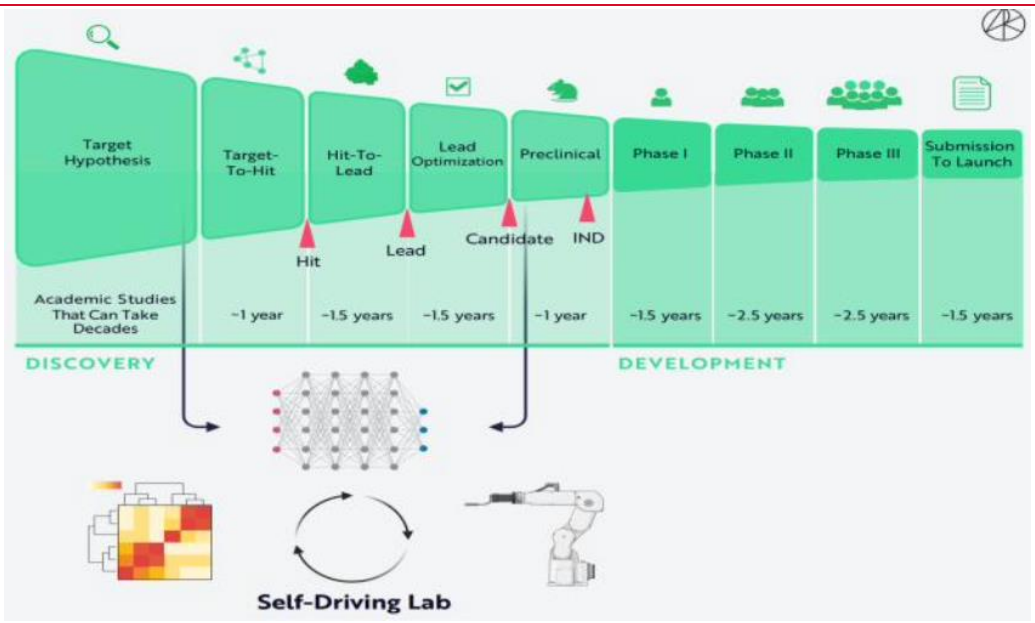
图表49：Anthropic Mythos 5 进行蛋白质复合物设计



资料来源：Anthropic，长城证券产业金融研究院

自主实验室正在形成全新科研范式，帮助 AI4S 搭建数据飞轮闭环。自动化实验室不再只是科研辅助手段，已经逐步成为 AI4S 体系中至关重要的物理执行载体。传统依靠人工开展的科研工作存在不少局限，包括人力成本压力、有毒试剂带来的安全隐患、难以实现不间断持续实验，同时还容易出现人为操作误差，对研发效率形成约束。2025 年初，《Nature》把自驱动实验室评选为全球七大前沿技术之一，它的核心价值是打通“AI 预测-自动化实验验证-数据回传”的完整循环。对于药物研发、材料合成这类周期漫长的长流程研究，自动化设备可以有效降低人为因素带来的干扰，显著提升物理实验处理规模，把 AI 给出的科学推演转化为实际产出，是推动科学智能从仿真计算走向实际落地的重要支撑。而模型、数据、算力共同组成 AI4S 的数字大脑，为系统提供科学理解、逻辑推理以及生成科研假设的能力。

图表50：自动化实验室工作流程

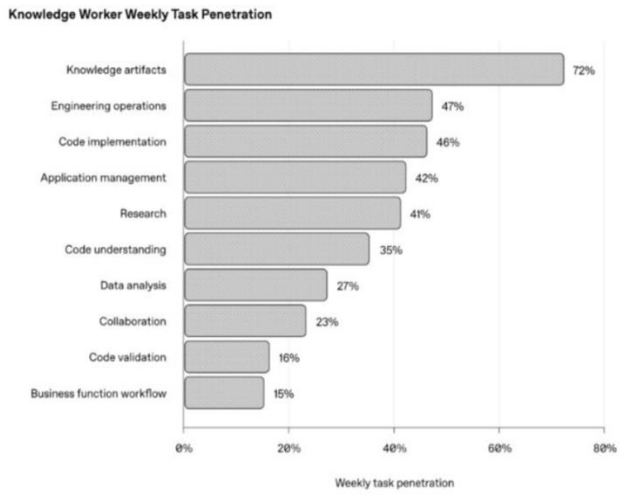


资料来源：Anthropic，长城证券产业金融研究院

3.4 AI 办公：白领办公市场空间广阔，通用办公 Agent 加速涌现

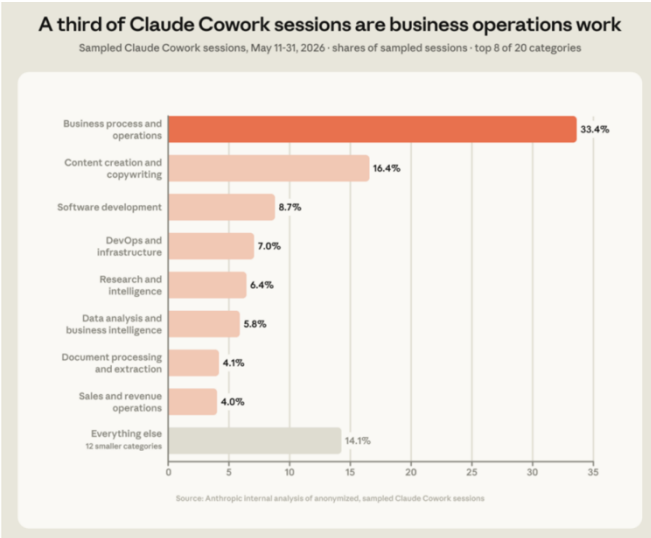
Anthropic 与 OpenAI 是 Agent 领域的代表性企业，二者产品已经体现出通用办公 Agent 的商业发展空间。OpenAI 在 2026 年 6 月 2 日发布的《Codex is becoming a productivity tool for everyone》报告提到，Codex 周活用户突破 500 万，对比 2 月份桌面端上线时提升 6 倍以上。知识工作者占全部用户的 20%，该群体增长速度高于开发者群体 3 倍。这类用户主要借助 Codex 完成报告、表格、PPT、合同等业务产出，数据分析、调研以及知识性内容创作是增长最快的任务类型；不少用户还会同时发起多项 Codex 任务，同步开展资料调研、文稿撰写以及工作流自动化操作。Anthropic 则统计了 2026 年 5 月下旬两周内来自 60 余万家机构合计 120 万次 Claude Cowork 匿名会话数据。结果显示，业务流程运营类应用占比达到三分之一，典型用途包含整合零散信息输出报告、拟定入职清单、校验表格数据等；内容文案类场景占 16%，例如生成初稿、制作演示材料、撰写推文与商业提案。

图表51：72%工作者使用 codex 进行知识内容制作



资料来源：OpenAI，长城证券产业金融研究院

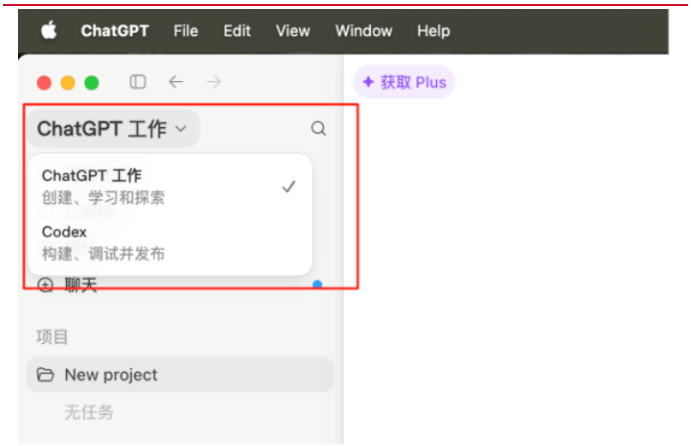
图表52：Claude Cowork 1/3 使用与业务流程相关



资料来源：anthropic，长城证券产业金融研究院

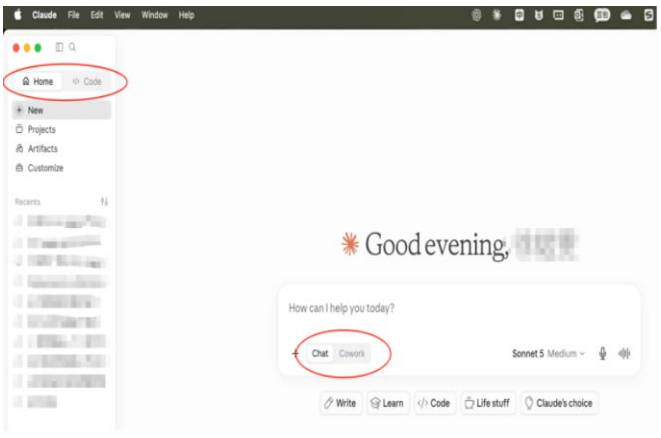
两家头部企业近期都对产品架构做出调整，加大布局通用办公 Agent 方向。2026 年 7 月 7 日，Anthropic 把原本仅面向桌面端的 Claude Cowork 更新为全场景办公 Agent，可适配桌面、网页、移动端，三端会话数据互通，支持云端后台持续执行任务，不受设备离线状态影响。时隔两天，OpenAI 将独立 Agent 产品 Codex 整合进 ChatGPT 桌面客户端，划分出 Chat、Work、Codex 三类功能模式，把对话交互、办公处理、开发能力整合到一处，后续还打算把 Work 模式逐步开放给网页及移动端付费用户使用。

图表53：ChatGPT 与 Codex 合并



资料来源：OpenAI，长城证券产业金融研究院

图表54：Claude Cowork 与 Chat 合并为统一入口



资料来源：Anthropic，长城证券产业金融研究院

我们认为办公场景是 AI 技术商业化变现路径清晰、战略价值高的"超级入口"。WorkBuddy 今年 3 月上线，当月 PC 端访问量达 885 万，有效承接了 OpenClaw 流量；后续三个月期间，PC 端访问量逐步攀升至 2097 万，反映 AI 办公赛道增长斜率陡峭。企业以及员工为效率付费意愿动力强，AI 提高工作效率，用户愿意买单。因此，AI 办公应用能形成完整的收入链条，包括个人订阅、企业订阅，以及任务执行带来的 Token 和云资源消耗。此外，我们认为 AI 在办公场景的高频调用，能持续获取企业私域数据（如审批流、文档），用以微调模型，形成用户使用越多→模型越懂企业→用户越离不开的正向数据飞轮循环。

腾讯、阿里、字节都在加大投入 AI 办公相关产品，其中腾讯 WorkBuddy 依靠较早布局，现阶段处于国内领先水平。国内办公 Agent 当前以桌面端产品为主，代表产品有阿里 QoderWork、腾讯 WorkBuddy、字节 TRAE Work。QoderWork 侧重安全合规，面向政企办公场景，采用完全本地强隔离方案，数据安全能力突出。WorkBuddy 面向大众用户，重点实现全桌面自动化，权限配置灵活，能够执行无人值守的自动任务。TRAE Work 定位全员 AI 原生工作台，同时覆盖办公和开发两类使用场景。个人产品层面，TRAE Work 提供免费版本；QoderWork 与 WorkBuddy 主要采用积分模式。三家厂商都针对企业客户设置分级定价，匹配不同规模企业的使用要求。从日活数据来看，依托先发优势，腾讯 WorkBuddy 目前在国内办公 Agent 市场保持领先。

图表55：国内主流办公 Agent 对比

	WorkBuddy	QoderWork	TareWork
产品定位	AI 原生工作台	AI 原生工作台	AI 原生工作台
产品形态	桌面应用	桌面应用	网页/桌面/移动应用
产品价格	个人： 体验版：免费-500 积分 标准：99 元/月-4000 积分 高级：199 元/月-9000 积分 旗舰：999 元/月-50000 积分 超额购买加量包 企业： 198/316 元/席/月，超额可购买加量包	个人： 体验版：免费-300 积分 专业：59 元/月-2000 积分 高级：169 元/月-6000 积分 旗舰：559 元/月-20000 积分 企业： 99/149/199 元/席/月（3000 积分），超额可商务签约定制包	个人： 免费（无 Token/次数上限），但可选择 99/239/699/1999 元/月速通套餐（回答过程无需排队、多个云端任务同时执行、优先解锁 SOTA 模型） 企业： 49/99/199 元/席/月，超额按实际 Token 消耗×模型单价计费
执行位置	本地优先、云端可切换的混合执行架构，默认基于终端本地环境完成全流程自动化任务，同时支持云端沙盒兜底执行	纯本地执行架构，全量任务运算与文件处理均执行位置在终端本地完成，原始业务数据不上传云端	支持本地离线执行与云端沙盒执行双向切换，桌面版可自主选择运行环境，网页版仅支持纯云端执行

资料来源：各公司官网，长城证券产业金融研究院

4.风险提示

AI 大模型迭代不及预期：若 AI 大模型技术迭代不及预期，则下游 AI 应用渗透率以及云厂商收入增长或不及预期。

AI 应用商业化不及预期：目前仍处于 AI 产业发展初期，MoE 模型的发展一定程度降低了大模型的使用门槛，但 AI 应用商业化落地仍需较大的资本开支投入，在短期 AI 应用的商业化落地有可能不及市场预期。

行业竞争加剧：科技巨头争相投入资本开支抢市场份额，如果出现某个云厂商发起价格战，则可能在某个阶段全行业增收不增利。

监管政策影响：云厂商需遵守本国及全球多区域监管规则，叠加地缘局势持续扰动，数据安全、跨境数据流动、算力出口管制、反垄断等监管政策处于动态调整阶段，监管政策影响加剧。

免责声明

长城证券股份有限公司（以下简称长城证券）具备中国证监会批准的证券投资咨询业务资格。

本报告由长城证券向专业投资者客户及风险承受能力为稳健型、积极型、激进型的普通投资者客户（以下统称客户）提供，除非另有说明，所有本报告的版权属于长城证券。未经长城证券事先书面授权许可，任何机构和个人不得以任何形式翻版、复制和发布，亦不得作为诉讼、仲裁、传媒及任何单位或个人引用的证明或依据，不得用于未经允许的其它任何用途。如引用、刊发，需注明出处为长城证券研究院，且不得对本报告进行有悖原意的引用、删节和修改。

本报告是基于本公司认为可靠的已公开信息，但本公司不保证信息的准确性或完整性。本报告所载的资料、工具、意见及推测只提供给客户作参考之用，并非作为或被视为出售或购买证券或其他投资标的的邀请或向他人作出邀请。在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议。在任何情况下，本公司不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任。

长城证券在法律允许的情况下可参与、投资或持有本报告涉及的证券或进行证券交易，或向本报告涉及的公司提供或争取提供包括投资银行业务在内的服务或业务支持。长城证券可能与本报告涉及的公司之间存在业务关系，并无需事先或在获得业务关系后通知客户。

长城证券版权所有并保留一切权利。

特别声明

《证券期货投资者适当性管理办法》、《证券经营机构投资者适当性管理实施指引（试行）》已于 2017 年 7 月 1 日 起正式实施。因本研究报告涉及股票相关内容，仅面向长城证券客户中的专业投资者及风险承受能力为稳健型、积极型、激进型的普通投资者。若您并非上述类型的投资者，请取消阅读，请勿收藏、接收或使用本研究报告中的任何信息。

因此受限于访问权限的设置，若给您造成不便，烦请见谅！感谢您给予的理解与配合。

分析师声明

本报告署名分析师在此声明：本人具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，在执业过程中恪守独立诚信、勤勉尽职、谨慎客观、公平公正的原则，独立、客观地出具本报告。本报告反映了本人的研究观点，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接收到任何形式的报酬。

投资评级说明

公司评级		行业评级	
买入	预期未来 6 个月内股价相对行业指数涨幅 15%以上	强于大市	预期未来 6 个月内行业整体表现战胜市场
增持	预期未来 6 个月内股价相对行业指数涨幅介于 5%~15%之间	中性	预期未来 6 个月内行业整体表现与市场同步
持有	预期未来 6 个月内股价相对行业指数涨幅介于-5%~5%之间	弱于大市	预期未来 6 个月内行业整体表现弱于市场
卖出	预期未来 6 个月内股价相对行业指数跌幅 5%以上		
	行业指中信一级行业，市场指沪深 300 指数		

长城证券产业金融研究院

深圳
地址：深圳市福田区福田街道金田路 2026 号能源大厦南塔楼 16 层
邮编：518033
传真：86-755-83516207

北京
地址：北京市宣武门西大街 129 号金隅大厦 B 座 27 层
邮编：100031
传真：86-10-88366686

上海
地址：上海市浦东新区世博馆路 200 号 A 座 8 层
邮编：200126
传真：021-31829681

网址：http://www.cgws.com

长城证券 GREAT WALL SECURITIES