浦銀國際 SPDB INTERNATIONAL

2026 年 11 月 11 日



大模型行业：能力兑现，增长加速

超配 **首次覆盖**

我们看好大模型在编程、办公与专业任务中的商业化机遇。随着模型可连续交付更加复杂的长程任务，编程深化与办公智能体普及将共同提升调用强度与单客付费，底层推理成本下探也有望扩大企业级高频调用。国产厂商依托高性价比定价、灵活部署模式及领先的多模态产品实现差异化，并借力云平台生态合作丰富变现渠道。智谱量价齐升验证前沿模型变现力，云托管分成与 Co-work 拓宽增长边界；MiniMax 企业端调用加速增长，多模态矩阵打开全球商业化空间。我们首次覆盖智谱（2513.HK）与 MiniMax（100.HK），均给予“买入”评级，目标价分别为 750 港元与 245 港元。

杨子超，CFA（科技分析师）
charles_yang@spdbi.com
(852) 2808 6409

赵丹（首席科技软件分析师）
dan_zhao@spdbi.com
(852) 2808 6436

欢迎关注
浦银国际研究

本研究报告由浦银国际证券有限公司分析师编制，请仔细阅读本报告最后部分的分析师披露、商业关系披露及免责声明。

目录

大模型行业：能力兑现，增长加速 4

投资要点 4

编程深化与办公智能体普及，驱动调用量增长 8

Scaling 延伸至后训练与推理，推高能力上限 13

跨周期迭代构筑壁垒，成熟降价与前沿溢价并存 16

国产模型构建差异化优势，云与生态合作拓宽变现渠道 19

推理提效与规模效应驱动盈利改善 23

Anthropic：大模型巨头如何率先跑通盈利闭环 24

智谱与 MiniMax：企业端增长提速，商业化路径各有侧重 27

投资风险 30

智谱（2513.HK）：前沿模型加速商业化 31

开放平台成为收入主体，云端服务高速放量 33

编程与长程任务能力支撑产品竞争力 34

客户使用持续加深，调用量与平均售价同步增长 35

云托管分成与专业任务采购丰富变现渠道 36

开放平台毛利改善，经营杠杆加速释放 38

盈利预测 40

估值分析与预测 41

投资风险 43

SPDBI 乐观与悲观情景假设 44

MiniMax（100.HK）：多模态驱动全球扩张 45

能力提升与成本优化带动企业端增长 47

文本与多模态并进，拓展全球商业化机会 48

推理与服务优化带动毛利修复 51

盈利预测 52

估值分析与预测 53

投资风险 55

SPDBI 乐观与悲观情景假设 56

浦银国际研究

首次覆盖 | 互联网行业

浦銀國際 SPDB INTERNATIONAL

大模型行业：能力兑现，增长加速

我们看好大模型作为生产力工具所带来的商业化机遇。当前编程场景已形成较清晰的付费需求，通用办公智能体与多模态产品正加速拓宽应用边界，叠加底层推理成本持续下探，有望扩大企业级高频调用。具备跨周期模型迭代能力、高稳定性工程交付及推理成本优势的厂商更有望受益。我们首次覆盖智谱（2513.HK）与 MiniMax（100.HK），均给予“买入”评级，目标价分别为 750 港元与 245 港元。

- **编程深化与办公智能体驱动企业级需求扩容。**复杂代码库与多步工程任务依赖多轮推理与工具调用，显著拉升单用户 Token 消耗强度。编程沉淀的任务拆解与执行能力正加速向文档处理、数据分析及跨应用协作泛化，推动 AI 渗透至更多非工程职能岗位。我们认为，随着模型可交付的任务范围扩大，新客渗透与存量客户扩容将共同驱动企业付费增长，按用量计费模式有望将任务强度转化为收入弹性。
- **行业呈现“成熟能力降价、前沿任务溢价”分层格局，国产厂商依托差异化定位与生态合作丰富变现渠道。**标准化任务同质化竞争充分，而复杂编程与长程智能体任务更看重首次交付质量以减少人工返工和降低总成本，使持续迭代的前沿模型得以维持定价权。国产厂商凭借开放权重与私有化/混合部署满足政企数据合规及定制需求，以高性价比 API 覆盖成本敏感型客群，并借力领先的视频等多模态产品切入全球内容创作市场。同时，云端 MaaS 平台与办公智能体为独立厂商拓宽了客户入口，商业授权与云托管分成亦丰富了变现渠道。
- **智谱（2513.HK）：量价齐升验证前沿模型变现力，云托管分成与 Co-work 拓宽增长边界；首予“买入”评级，目标价 750 港元。**公司已全面转向云端持续服务模式。受益于复杂编程与长程任务需求放量，Token 实现量价齐升，技术领先优势正转化为收入增长。原厂 API 之外，云托管分成拓宽了第三方云平台部署的变现渠道，Co-work 进一步承接网络安全等高价值专业任务需求。管理层于 9 月将 2026 年底 ARR 指引上调至 30 亿美元。我们看好模型升级与多元渠道带来的增长弹性，国产算力规模化适配与推理优化有望持续提升开放平台毛利。
- **MiniMax（100.HK）：企业端增长提速，多模态矩阵打开全球商业化空间；首予“买入”评级，目标价 245 港元。**企业服务已成为公司核心增长引擎。M 系列文本模型结合高性价比的定价策略持续带动编程与智能体调用放量。行业领先的 H3 视频生成及语音、音乐产品满足全球广告、电商及内容创作等需求。公司依托开放平台 API 及自有 AI 原生应用服务全球客户，管理层于 8 月维持 2026 年底 10 亿美元 ARR 目标。我们看好企业端与全球多模态应用的增长潜力，推理效率提升与套餐优化有望推动今年下半年毛利率环比回升。
- **投资风险：**模型迭代及商业化不及预期；行业竞争与价格压力；毛利率改善不及预期；研发及算力投入较大、融资与股本摊薄；海外监管、数据安全及知识产权风险。

杨子超，CFA
科技分析师
charles_yang@spdbi.com
(852) 2808 6409

赵丹
首席科技软件分析师
dan_zhao@spdbi.com
(852) 2808 6436

2026 年 11 月 11 日

智谱（2513.HK）	**买入**
目标价（港元）	750
潜在升幅/降幅	+76%
目前股价（港元）	426.2
MiniMax（100.HK）	**买入**
目标价（港元）	245
潜在升幅/降幅	+57%
目前股价（港元）	155.8

注：截至 2026 年 11 月 6 日收盘价

资料来源：Bloomberg、浦银国际

浦银国际 | 首次覆盖 | 大模型行业：能力兑现，增长加速

本研究报告由浦银国际证券有限公司分析师编制，请仔细阅读本报告最后部分的分析师披露、商业关系披露及免责声明。

大模型行业：能力兑现，增长加速

投资要点

我们看好大模型作为生产力工具所带来的商业化机遇。随着模型复杂任务处理能力提升与推理成本下探，高频企业级应用更具经济性，正加速驱动工作流向 AI 迁移。贝恩预计 2031 年全球现有 AI 应用年收入将达 1.2 万亿至 1.8 万亿美元，其中企业端贡献约 1.0 万亿至 1.4 万亿美元。我们认为，具备跨周期模型迭代能力、高工程交付稳定性及推理成本优势的厂商，有望率先把算力消耗转化为可持续的收入增长。

编程场景深化与办公智能体普及驱动模型调用需求扩容。编程任务投资回报率清晰，率先跑通商业闭环，复杂代码库的多轮推理与工具调用显著拉升单用户 Token 消耗强度。当前企业级 AI 渗透仍有广阔空间，Ramp 数据显示 2026 年 9 月样本企业 AI 采用率为 55.9%，重度支出仍集中于头部 1%企业。同时，编程沉淀的任务拆解与工具调用能力正加速向非工程岗位泛化，2026 年 9 月国内 20 款主流桌面端 AI 办公智能体月访问量已突破 3 亿次。随着智能体接入企业文档与协同系统，报表分析、跨应用协作等易核验任务有望率先放量，并进一步向个人消费与行程规划等生活服务延伸。

技术演进延伸至后训练与推理侧，持续迭代能力构筑核心壁垒，行业呈现“成熟能力降价、前沿任务溢价”的分层格局。在预训练奠定基座能力的基础上，后训练强化学习与测试时计算成为改善复杂规划的关键路径。由于单次榜单领先易被追平，厂商的核心壁垒在于跨发布周期的研发韧性与规模化工程交付能力。在此背景下，简单问答等标准化场景同质化竞争充分，单价持续下行；而复杂编程与长程智能体任务高度依赖交付成功率以减少返工及降低总成本，使头部前沿模型得以维持较强的定价权与能力溢价。

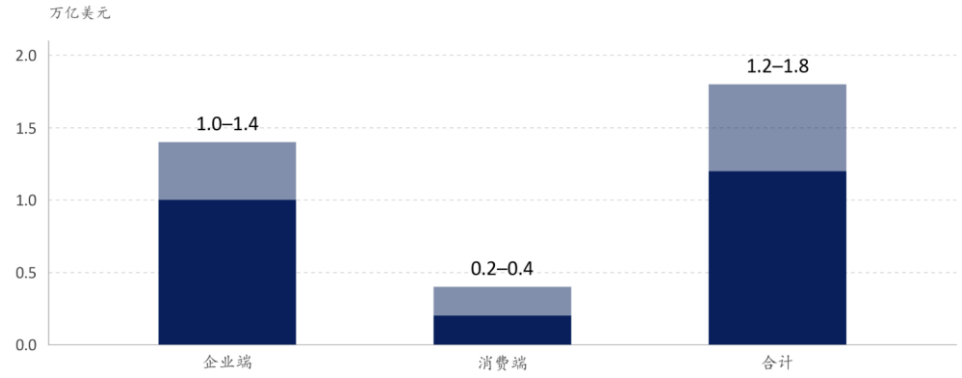
国产大模型厂商依托部署灵活性、多模态优势及高性价比实现差异化，生态合作与推理降本推动盈利改善。国内厂商通过开放权重与私有化/混合部署满足政企数据合规与定制需求，凭借领先的视频生成等多模态产品切入广告、电商及内容创作市场，并以具竞争力的 API 定价覆盖成本敏感型客群。同时，国产大模型厂商借助云厂商平台及办公智能体拓宽分发入口，通过商业授权与云托管分成丰富变现渠道。盈利端，随着调用规模扩张带动集群利用率提升，叠加架构优化与国产算力规模化适配，单位服务成本下行有望持续增加毛利；收入体量快速放大亦将摊薄研发费用率，推动亏损收窄。

智谱（2513.HK）：量价齐升验证前沿模型变现力，云托管分成与 Co-work 拓宽增长边界；首予“买入”评级，目标价 750 港元。公司全面转向持续模型服务模式，1H26 开放平台及 API 收入同比增长 2,735.7%，占比达 86.5%。受益于 GLM 系列在复杂编程与长程任务上的表现提升，截至 8 月底平台 Token 调用量较年初增长超 40 倍，API 平均售价同步翻倍，高价值任务支撑强劲付费意愿。原厂 API 之外，公司与多家国内外头部云厂商签署托管 API 分成协议、以轻资产模式放大分发规模，并以 Co-work 切入网络安全等高价值专业任务，驱动管理层将 2026 年底 ARR 指引上调至 30 亿美元。随着 10 万卡级国产算力集群规模效应释放与全栈推理优化，1H26 公司开放平台毛利率已提升至 24.6%，盈利能力有望持续改善。

MiniMax（100.HK）：企业端增长提速，多模态矩阵打开全球商业化空间；首予“买入”评级，目标价 245 港元。企业级服务已成为公司核心增长引擎，1H26 开放平台及企业服务收入同比激增 703.1%至 7,393 万美元，占总收入 63.4%，7 月 Token 调用量达年初的 20 倍；8 月 ARR 达 8 亿美元，企业端贡献超 80%。管理层维持年底 10 亿美元 ARR 目标。公司 10 月推出的 M3.1 Flash 将智能体任务使用成本大幅降低 70%，高性价比定价策略有望进一步撬动日常开发与高频智能体调用。同时，公司形成文本、视频、语音及音乐全模态布局，通过开放平台与自有 AI 原生应用双轮驱动，承接全球广告、电商及内容创作需求，1H26 海外收入同比增长 224.0%，占比达 60.8%。随着 M3 上线初期补偿影响消退、套餐结构优化及推理效率提升，下半年毛利率有望迎来环比修复。

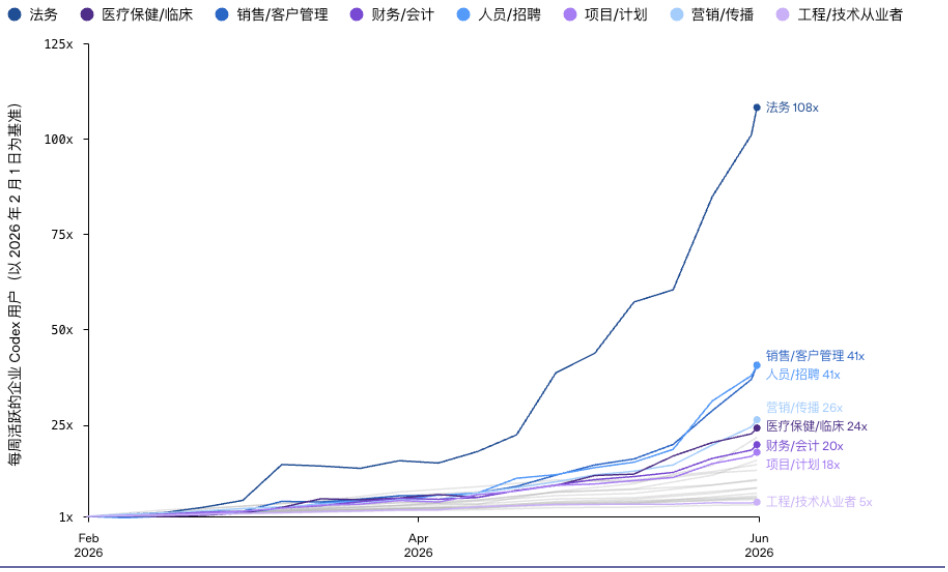
核心图表

图表 1：全球企业端+消费端 AI 市场 2031 年预计达到 1.2-1.8 万亿美元



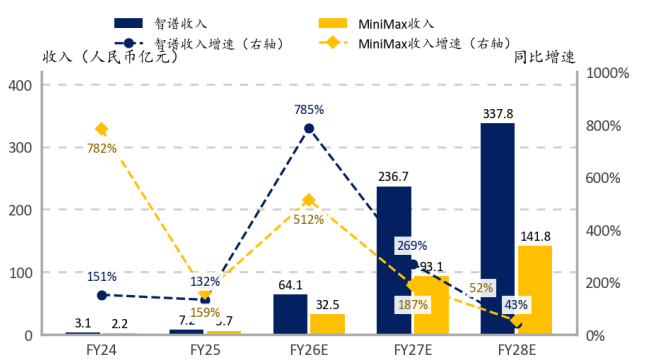
资料来源：贝恩、浦银国际

图表 2：非工程岗位智能体用户快速增长



资料来源：OpenAI、浦银国际

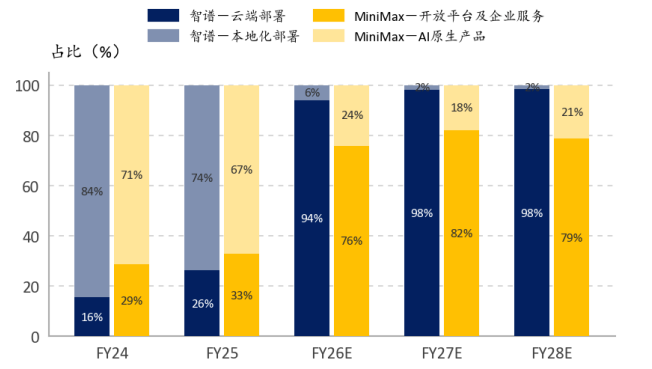
图表 3：智谱及 MiniMax 收入预测



E=浦银国际预测

资料来源：公司资料、浦银国际

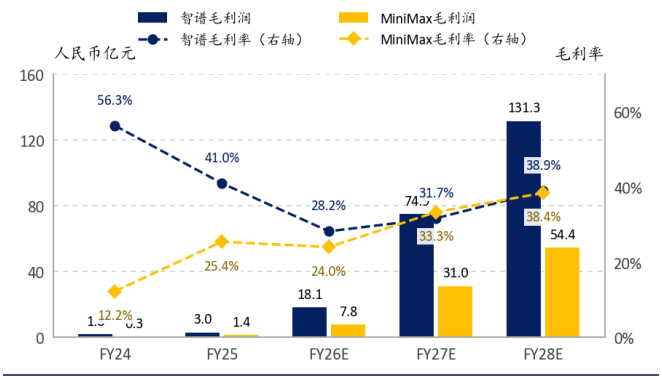
图表 4：智谱及 MiniMax 收入结构预测



E=浦银国际预测

资料来源：公司资料、浦银国际

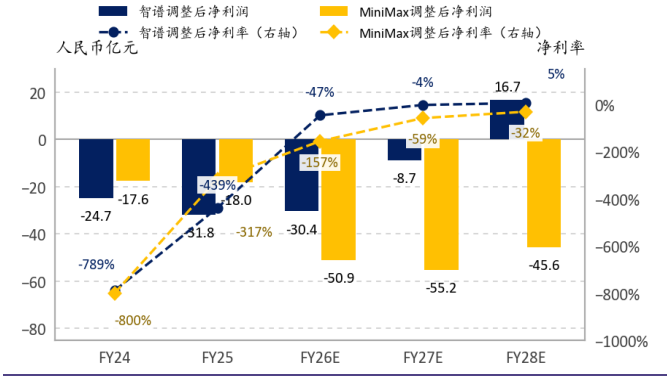
图表 5：智谱及 MiniMax 毛利润预测



E=浦银国际预测

资料来源：公司资料、浦银国际

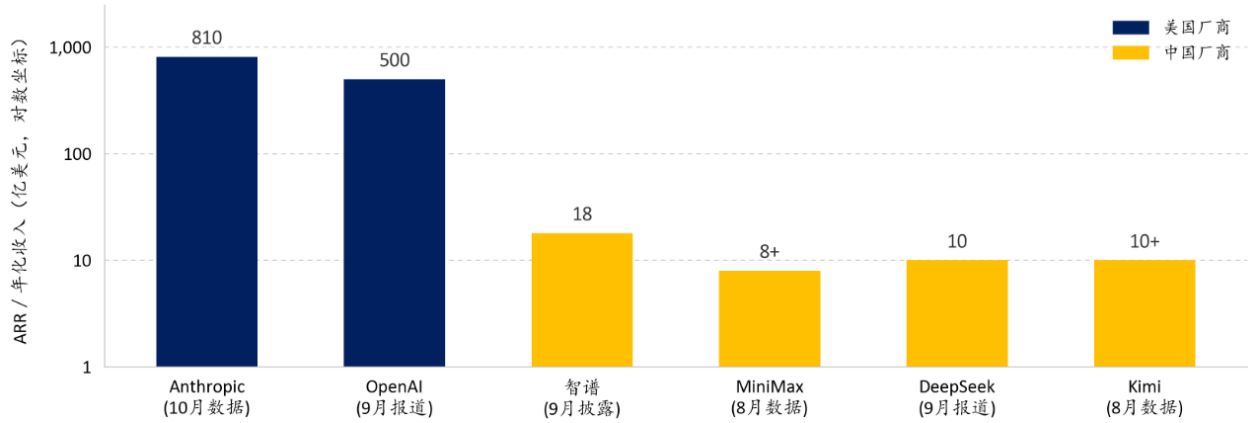
图表 6：智谱 MiniMax 调整后净利润预测



E=浦银国际预测

资料来源：公司资料、浦银国际

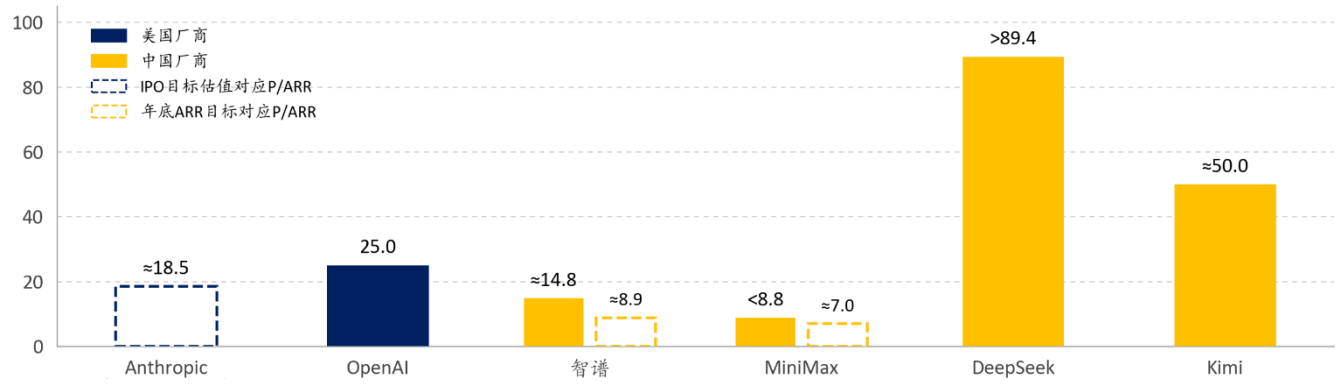
图表 7：中美主要独立模型厂商 ARR



注：ARR 按公司披露或媒体报道的年化收入规模列示，各公司统计时点、年化方式及合作渠道收入确认口径有所不同。

资料来源：The Information、Bloomberg、Financial Times、Reuters、WSJ、公司资料、浦银国际

图表 8：中美主要独立模型厂商 P/ARR



注：市值截至 2026 年 11 月 6 日；

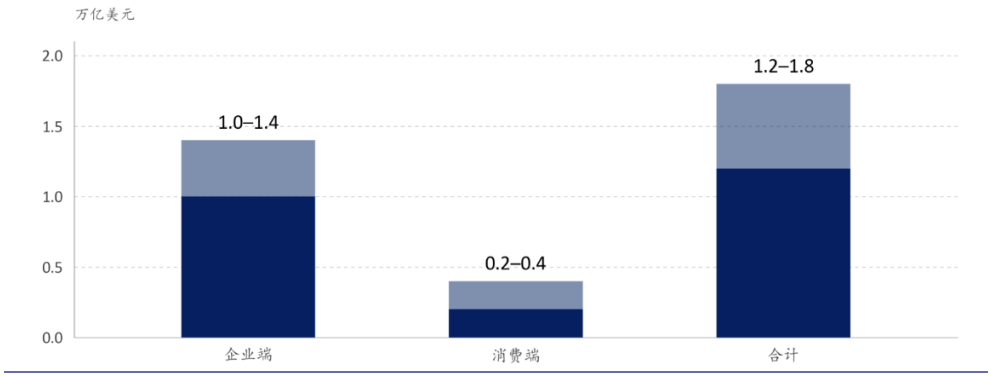
资料来源：The Information、Bloomberg、Financial Times、Reuters、WSJ、Wind、公司资料、浦银国际

编程深化与办公智能体普及，驱动调用量增长

付费客户增加和既有客户使用加深，推动模型服务收入增长。企业将模型推广至更多部门与业务环节，有助于增加使用人数和单客付费。编程已形成较明确的付费需求，办公智能体向更多岗位普及，个人应用进一步拓展了消费端用途。

企业 AI 部署支出快速增长，应用市场空间广阔。根据 Menlo Ventures 估算，截至 2026 年 9 月，企业 AI 部署支出的年化规模为 1,768 亿美元，约为 2025 年的 5.4 倍，其中 AI 应用的年化规模为 773 亿美元。贝恩进一步测算中长期应用空间，预计到 2031 年，企业和消费端现有 AI 应用合计可带来 1.2 万亿至 1.8 万亿美元年收入，其中企业端约为 1 万亿至 1.4 万亿美元，主要来自软件开发、销售、营销、客服和 IT 运维等环节；消费端约为 2,000 亿至 4,000 亿美元，主要通过订阅与广告变现。

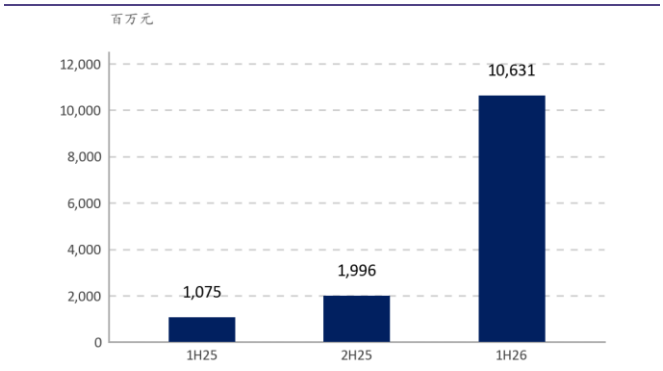
图表 9：全球企业端及消费端 AI 市场规模预测（2031 年）



资料来源：贝恩、浦银国际

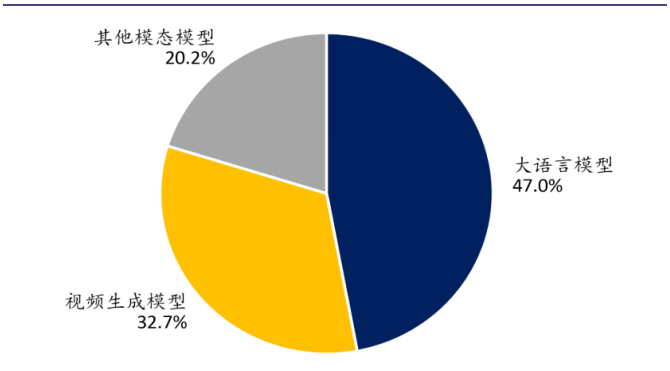
国内公有云模型服务已形成百亿元的半年度收入规模。IDC 统计，1H26 中国公有云 MaaS 收入达人民币 106.3 亿元，同比增长 889.4%，占公有云 PaaS 收入的比重由 2H25 的 5.2%升至 21.0%。同期，大语言模型和视频生成分别贡献 MaaS 收入的 47.0%和 32.7%，视频已成为重要收入来源。

图表 10：中国 MaaS 市场规模



资料来源：IDC、浦银国际

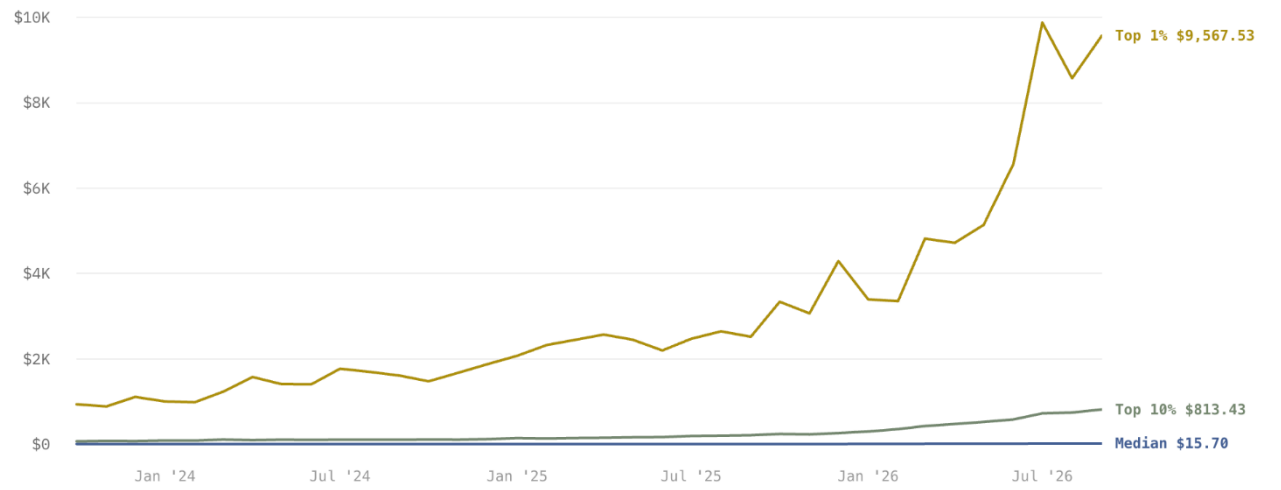
图表 11：中国 MaaS 市场结构（1H26）



资料来源：IDC、浦银国际

企业 AI 投入集中于部分重度用户，使用深度仍有较大差异。Ramp 的 10 月报告显示，2026 年 9 月，样本企业 AI 采用率达到 55.9%；按每员工每月 AI 支出计，高支出前 1%的企业达到 9,568 美元。Ramp 此前分析亦显示，样本中 1%的企业贡献了约 80%的 OpenAI 及 Anthropic 支出。报告认为，新增 AI 付费客户增长放缓后，使用强度提升将成为市场增长的重要来源。

图表 12：美国企业每员工每月 AI 支出分层（美元）



资料来源：Ramp AI Index、浦银国际

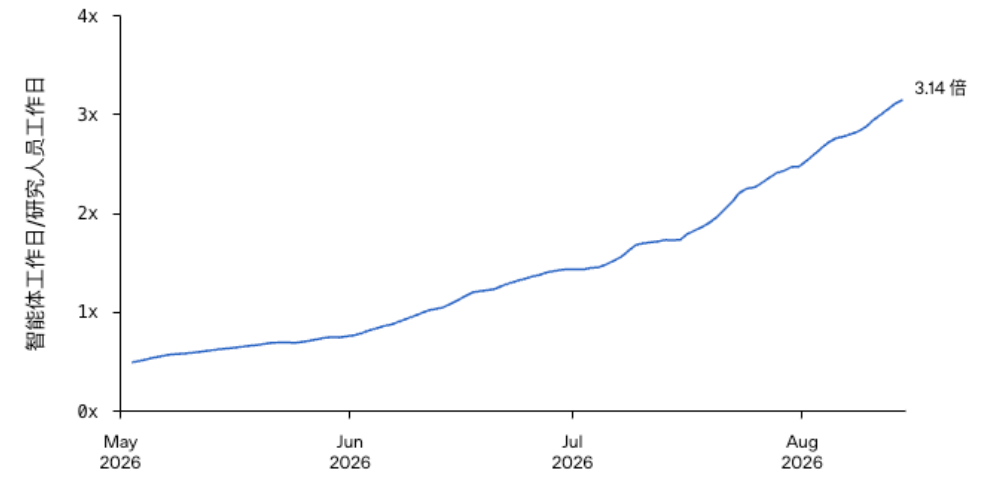
企业客户使用持续加深，大客户扩容成为收入增长的重要来源。Anthropic 招股书披露，截至 2026 年 9 月末，过去 12 个月收入贡献至少 10 万美元的客户超过 11,900 家，上年同期超过 950 家；合资格存量客户的净收入留存率（NRR）约 900%。与此同时，2026 年第三季度，公司来自过去 12 个月收入贡献至少 10 万美元的《财富》500 强客户的总收入，平均不足这些客户总费用的 0.1%，其中收入贡献最大的 10 家对应比例平均不足 1%。我们认为，随着模型能够承担更多工作，企业向更多部门和任务扩大使用，有望持续带动调用与付费增长。

Coding 较早形成付费需求，复杂任务推高调用强度。Coding 是当前大模型应用较深的场景之一，原因在于：1）开发者时间成本较高，代码生成、调试、测试和运维带来的效率提升较易衡量；2）复杂代码库和跨文件任务需要多轮交互、长上下文及工具调用，能够带来较高 Token 消耗；3）编译器、单元测试与持续集成等工具能提供可验证反馈，帮助模型持续改进。

智能体提高任务执行强度，为按量消费提供了增长空间。据 OpenRouter 平台统计，2026 年 2 月初前后，智能体 Token 用量首次超过人类直接交互用量；截至 9 月 28 日，智能体 Token 消耗已约为人类消耗的 7 倍。据 OpenAI 披露，以标准的 8 小时工作日计算，截至 8 月中旬，其内部研究场景中的智能体标准化运行时长约为研究人员工作时长的 3.1 倍。从客户付费看，据

Anthropic 披露，2026 年第三季度，未使用 Claude Code 的 Claude Enterprise 用户月均消费不足 75 美元，而使用 Code 的前 5%高消费用户月均消费约为前者的 20 倍。我们认为，随着智能体承担更长、更复杂的工程任务，单个用户的任务量和执行时长有望持续增加，为按量计费的模型服务带来收入增量。

图表 13：OpenAI 内部研究场景的智能体标准化运行时长超过研究员工时



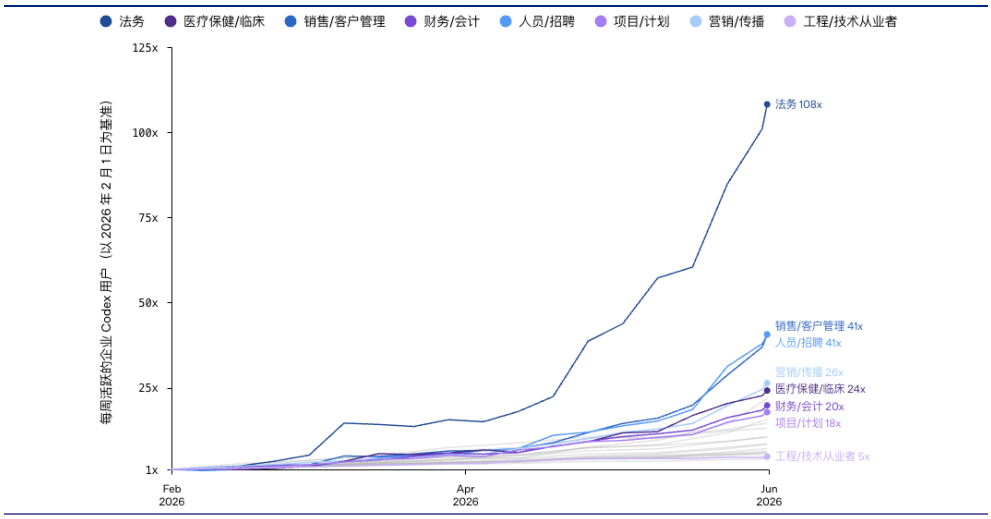
资料来源：OpenAI、浦银国际

Work 将智能体的应用范围扩展至更多知识工作岗位。编程任务中积累的任务拆解、工具调用、沙箱隔离、记忆与权限管理等能力，也可迁移至文档处理、数据分析和跨应用协作等知识工作场景。相比单轮问答，办公智能体能够读取上下文、连续执行并交付文件，扩大面向非工程岗位的应用范围。

Anthropic 围绕 Claude Cowork 持续完善插件、管理员控制、定时任务和计算机操作等能力；OpenAI 通过企业智能体平台、Workspace Agents 及 ChatGPT Work，扩展跨应用协作和多步任务执行。国内厂商也在整合办公智能体与既有协作产品。阿里将 QoderWork、MuleRun 和悟空整合为千问办公；字节推动豆包工作与飞书协同，并整合相关智能体研发团队；腾讯则整合 QClaw 与 WorkBuddy 相关业务。这些产品逐步将模型调用与日常工作流程结合，为企业落地 AI 提供更完善的工具。

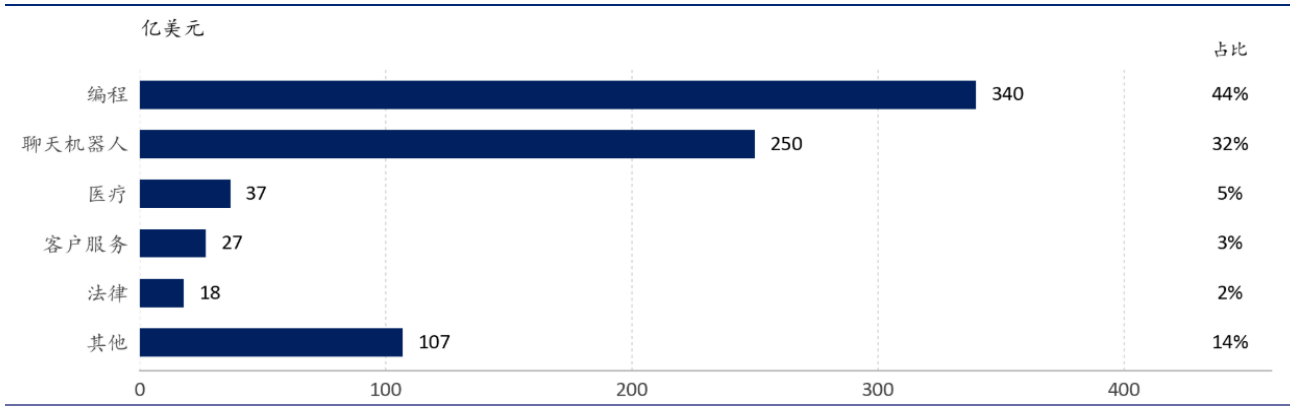
OpenAI 于 2026 年 8 月发布的研究显示，截至 6 月，在其覆盖的企业客户中，Codex 输出 Token 占 Codex 与 ChatGPT 合计输出的 64%；分岗位看，自今年 2 月以来，法务、销售、招聘和营销等职能的 Codex 周活跃用户均显著增长。

图表 14：按职位划分的智能体用户增长情况



资料来源：OpenAI、浦银国际

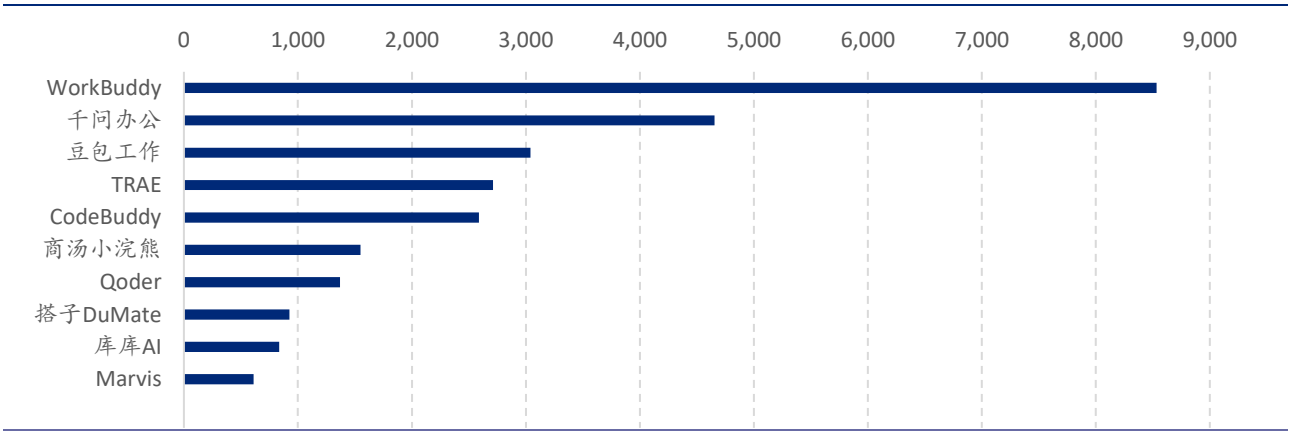
图表 15：企业 AI 应用年化支出分布（2026 年 9 月）



资料来源：Menlo Ventures、浦银国际

国内办公智能体的使用规模也在扩大。易观数据显示，2026 年 9 月，20 款主流桌面端 AI 原生办公智能体合计月访问量超过 3 亿次，其中 WorkBuddy 为 8,534 万次、千问办公为 4,655 万次、豆包工作为 3,039 万次。智能体接入企业的即时通讯、文档、日程及审批系统后，可结合业务信息执行任务，减少用户手动查找资料和切换应用的操作。我们认为，拥有协同办公入口或生态合作基础的厂商，在企业接入方面更有优势。

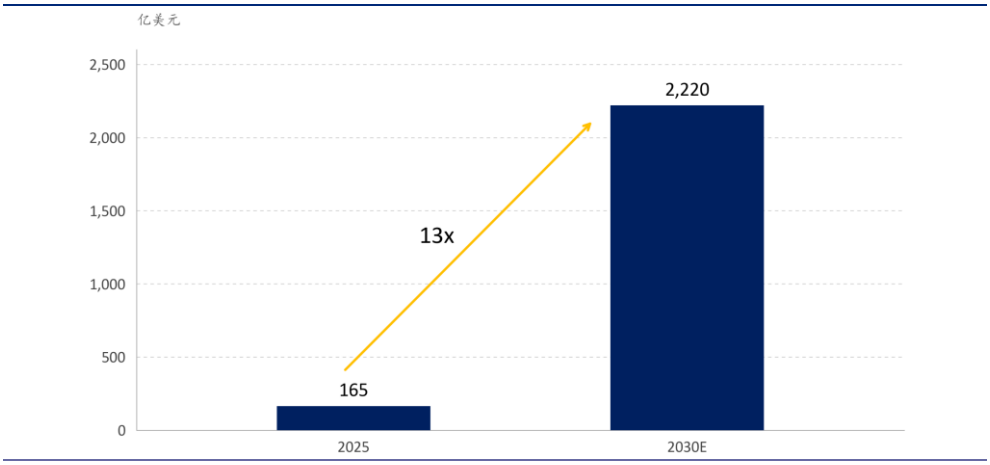
图表 16：2026 年 9 月头部桌面端 AI 办公智能体月访问量（万次）



资料来源：易观、浦银国际

智能体有望拓展企业与消费端的付费需求。Gartner 预计，2030 年 AI 智能体与助手的市场规模将达到 2,220 亿美元，相比于 2025 年的市场规模（165 亿美元）扩大约 13 倍。随着跨应用执行能力普及，智能体的服务范围正由企业工作流程延伸至个人事务。

图表 17：AI 智能体与助手市场规模



注：E=Gartner 预测；资料来源：Gartner、浦银国际

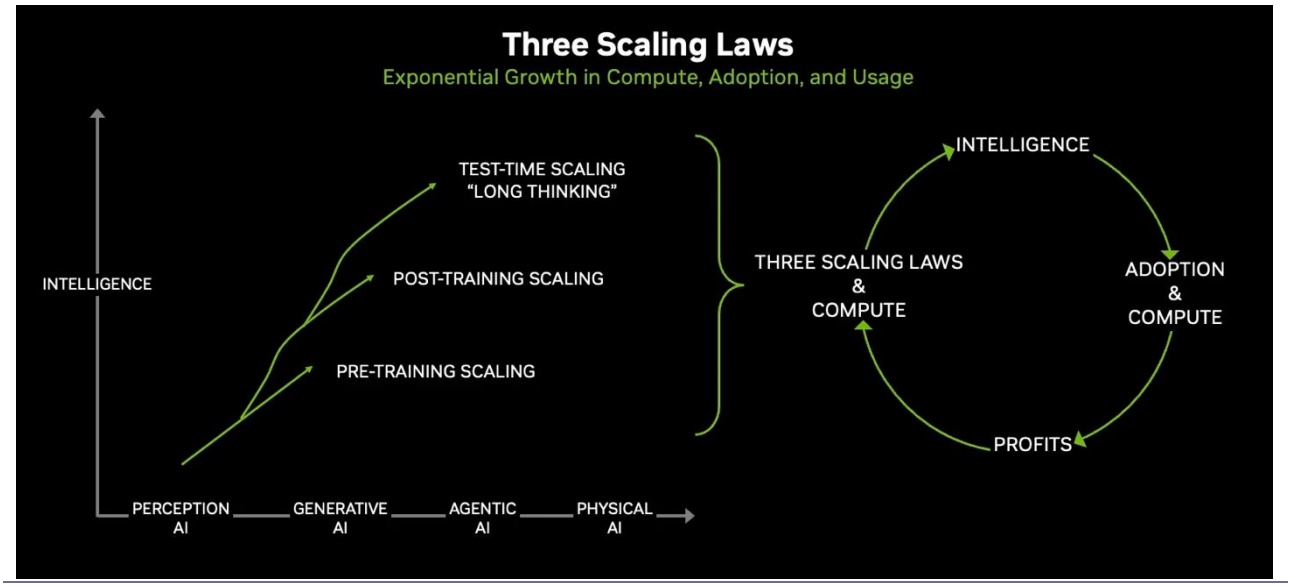
模型厂商正通过个人智能体提供持续执行服务。Meta Muse、Grok Bot 与 OpenAI Dots 均配备云端计算机，可在用户授权后跨应用执行多步骤任务，并在用户离开后继续工作，支持资料研究、文档处理、购物及行程安排。Menlo Ventures 于 2026 年 7 月对 5,067 名美国成年人的调查显示，24%的 AI 用户经常使用智能体，个人用户已开始将部分任务交由智能体完成。

Meta Muse 的早期使用显示，购物是个人智能体的重要应用场景。Meta 在 2026 年 10 月 26 日披露，Muse 上线初期在美国和加拿大吸引超过 1,000 万用户，超过 25%的周活用户使用购物功能；公司另有超过 200 万商家每周使用 Business Agent。我们认为，个人助手与商户服务结合，有助于简化商品查找、比较和购买流程，扩大 AI 在消费决策中的应用。

Scaling 延伸至后训练与推理，推高能力上限

模型能力持续提升，为编程、办公与专业任务的应用扩大提供基础。根据 Epoch AI 追踪，前沿语言模型的训练算力自 2020 年以来每年增长约 4 倍，倍增周期约 5.2 个月；推理模型问世后，其能力指数（ECI）前沿的年化进步幅度进一步加大。由于进一步扩大模型参数量与预训练数据规模受到高质量数据供给、算力和成本的约束，厂商也正加大后训练与测试时计算的投入，拓展模型能力提升的路径。预训练继续扩大模型的知识与能力基础，后训练强化工具调用和长程任务执行，测试时计算则按任务需要增加推理投入。

图表 18：三类 Scaling 路径示意

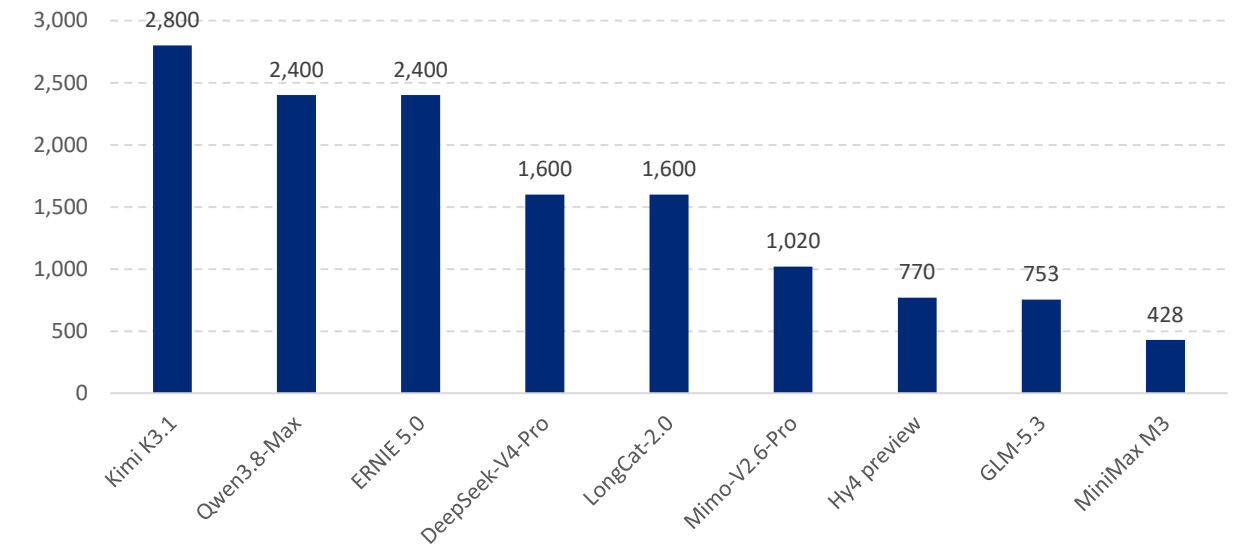


资料来源：NVIDIA、浦银国际

预训练仍是提升模型基础能力的重要途径。以自回归语言模型为例，预训练通过在大规模语料上预测下一个 Token，学习语言规律与知识。经典 Scaling Law 描述的是，在一定训练设定与规模范围内，模型预测误差随参数规模、数据量和训练算力增加而按幂律下降。因此，在给定算力预算下，通过高质量数据筛选、训练配比与架构优化，可以提高算力投入的效率。Epoch AI 数据显示，预训练算力效率平均每年改善约 3 倍，达到给定预训练表现所需的算力持续减少。更强的基座也为后训练提供更好的起点，使厂商能够在同一基础上拓展编程、推理和多模态任务。

当前国产旗舰模型正在持续扩大参数规模，并通过稀疏架构控制单次推理的计算量。Kimi K3.1 总参数达 2.8 万亿，是目前总参数规模最大的开放权重模型；Qwen3.8-Max、DeepSeek V4-Pro、MiMo-V2.6-Pro 总参数亦超过万亿。阿里巴巴、MiniMax 等厂商仍在推进数万亿参数级别模型，有望持续推高模型能力上限。国内主流旗舰模型普遍采用混合专家模型（MoE）架构。稠密模型中，参与单次推理的参数较多，总参数与计算量联系更紧密；而稀疏结构下，模型被拆分为大量子模块，每处理一个 Token 只调用其中很小一部分，激活比例通常可低至个位数百分比，因此总参数与每 Token 计算量不再同比例增长。

图表 19：厂商模型总参数规模对比（十亿）



资料来源：公司资料、浦银国际

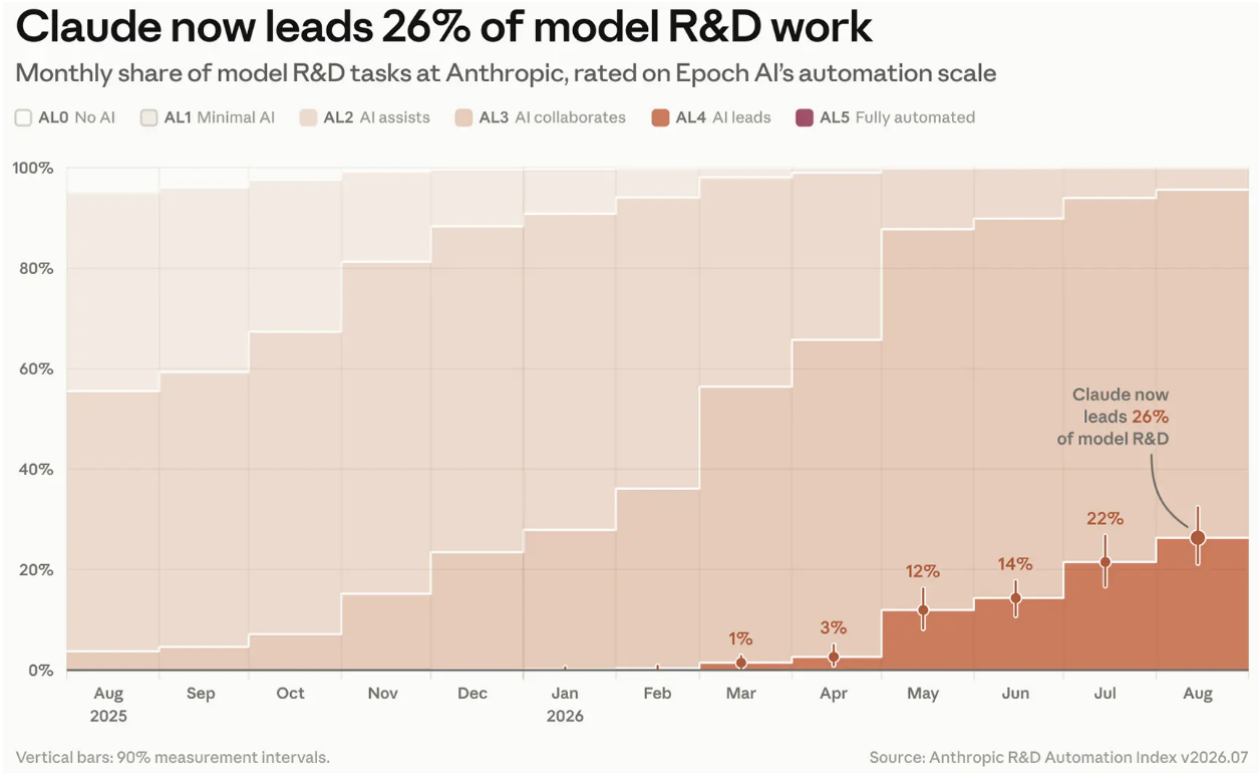
后训练强化学习正在成为模型升级的重要来源。后训练是在预训练模型基础上，通过监督微调、强化学习等方法，进一步提升模型任务能力、可靠性和实用性的训练过程。在代码、数学等结果可验证的场景中，模型可反复尝试任务，并根据编译、测试或答案校验获得反馈。随着环境覆盖更多任务、奖励设计更加准确，厂商可以在不全面重训基座的情况下持续改进实用能力，缩短模型升级迭代周期。

据智谱披露，今年 8 月发布的 GLM-5.3 基座与 GLM-5.2 相同，全部能力增益来自后训练规模化。公司表示，能力提升主要来自长程任务环境扩展、任务类型增加及后训练算力投入。相较于 GLM-5.2，GLM-5.3 在 Z.ai Code Bench 提升约 50%，Terminal-Bench 3.0 由 4.6 升至 28.3。这说明后训练可以在基座不变的情况下，显著改善特定任务的表现，为模型能力提升提供了另一条路径。

测试时计算使任务表现与使用成本可以按需权衡。模型训练完成后，调用方仍可通过延长推理过程、并行生成多个候选答案并择优，或增加智能体工具调用与执行步数，改善复杂任务的完成效果。计算投入由此从一次性训练延伸至每次任务执行。当前主流模型普遍提供可选的推理强度档位，简单任务使用较低档位控制成本，复杂任务增加推理投入以提高完成质量，使模型产品能够覆盖更广的任务和价格区间。

AI 辅助研发有望提高迭代效率。现阶段，模型已开始参与实验工具开发、评估器编写、训练方案搜索和任务环境构建，承担大量重复性研发工作。Anthropic 内部评估显示，AI 参与模型研发的程度正在提高。截至 2026 年 8 月，Claude 可在人工监督下主导约 26%的内部研发工作，较同年 2 月不足 1%实现显著增长；AI 协作及以上程度的工作占比超过 90%。我们认为，AI 参与实验和评测有望缩短研发周期，使团队能够更快验证训练方案并改进后续模型。

图表 20：在 Anthropic 内部，Claude 主导的模型研发工作占比升至 26%



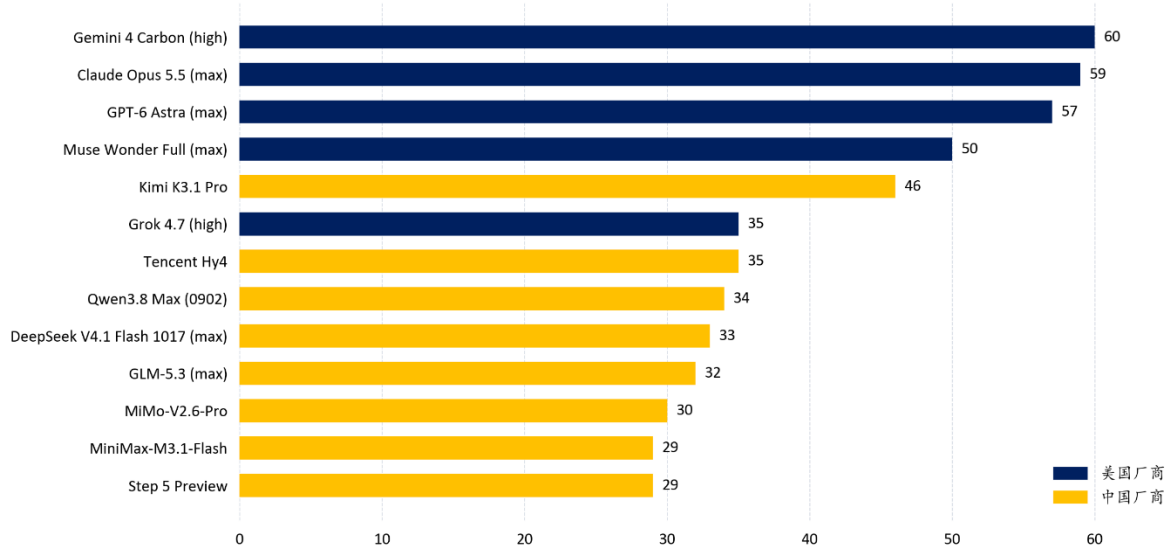
资料来源：Anthropic、浦银国际

跨周期迭代构筑壁垒，成熟降价与前沿溢价并存

我们认为，大模型行业的竞争壁垒主要在于持续研发与迭代能力。榜单排名或随新品发布交替变化，但能持续迭代前沿模型的厂商，通常需要长期积累人才、训练经验和算力工程能力。我们看好能够连续推出有竞争力的模型、并将新能力稳定用于客户任务的厂商。

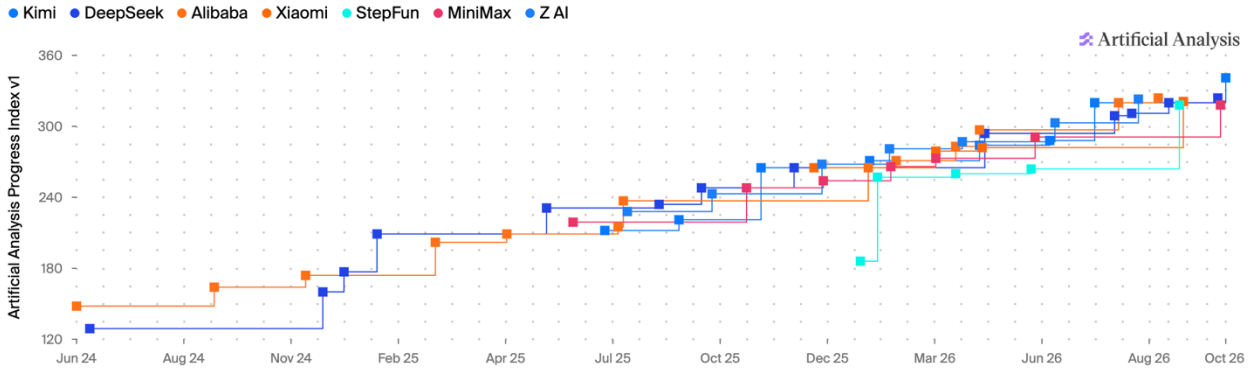
国产旗舰模型持续迭代，领先位置随新品发布变化。2026 年以来，智谱 GLM、月之暗面 Kimi、DeepSeek、阿里千问等系列先后推出新版本，均在编程、推理与综合能力评测中取得进展。根据截至 2026 年 11 月 6 日的 Artificial Analysis 智能指数，Kimi 3.1 Pro 处于国内领先地位，随后是腾讯 Hy4、Qwen 3.8 Max、DeepSeek-V4.1-Flash-1017 及 GLM 5.3 等模型。

图表 21：全球主要厂商模型能力比较（Artificial Analysis 智能指数，截至 2026 年 11 月 6 日）



资料来源：Artificial Analysis、浦银国际

图表 22：国内主要厂商模型能力持续提升，领先位置交替变化



资料来源：Artificial Analysis、浦银国际

研发能力也需要与工程交付相配合。模型开放权重后，其他服务商也可以部署同一模型；原厂的差异化更多体现在更新速度、推理成本、服务稳定性及配套工具上。推理优化依赖模型研发与部署团队配合，并在真实流量中反复调整。调用规模扩大，有助于团队更快发现性能瓶颈并检验优化效果。

任务环境与训练经验的积累有助于形成迭代优势。后训练强化学习依赖覆盖真实需求的任务环境、能够识别有效结果的奖励设计，以及稳定的评测方法。相近规模的基座模型，经过不同后训练也可能呈现明显的任务表现差异。研发团队在长程任务、工具调用和错误恢复中积累的经验，可用于下一轮训练和产品升级。

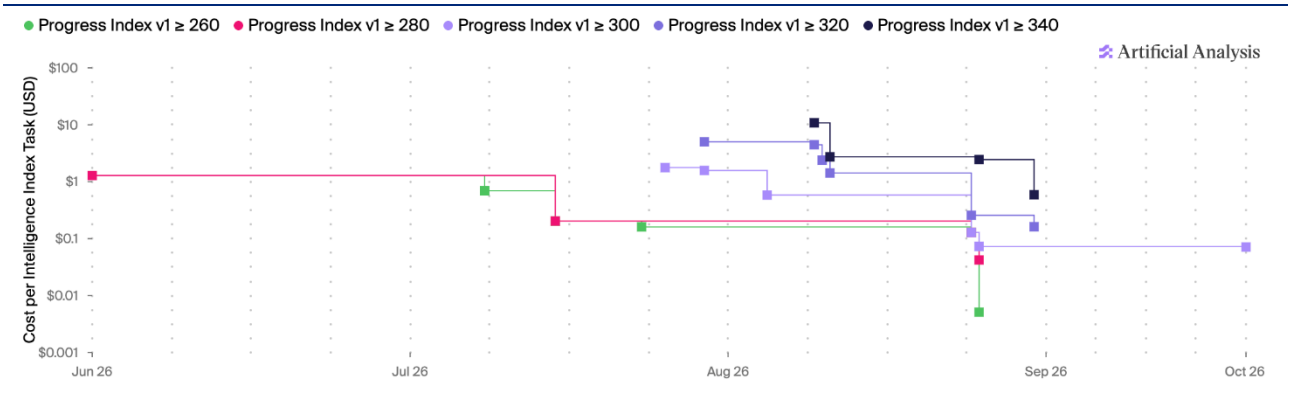
深入业务流程的集成有助于增强客户黏性。模型接入客户的代码库、内部数据及业务流程后，围绕特定任务形成的提示词、权限管理、评测和工具接口，通常需要持续适配和维护。对于集成较深的工作流程，更换模型涉及重新测试任务完成率与服务稳定性，因而采购决策会同时考虑效果、价格和迁移成本。

成熟能力降价与前沿任务溢价并存

我们认为模型层将呈现分层竞争，复杂任务仍能支持前沿模型的能力溢价。通用问答、文本改写等成熟任务已有较多可替代模型，价格竞争相对充分。复杂编程与多步智能体任务则对可靠性、长程执行和交付质量提出更高的要求，从而打开能力溢价空间。

成本改善为成熟能力降价提供支撑。稀疏混合专家架构、注意力机制优化和推理调度，可以降低给定任务质量下的算力消耗。我们认为，单位服务成本的降幅高于实际售价降幅，有助于提高毛利率。Epoch AI 数据显示，AI 芯片的单位美元性能自 2023 年以来每年提升约 49%。硬件性价比与推理优化共同降低服务成本，为降价提供空间。

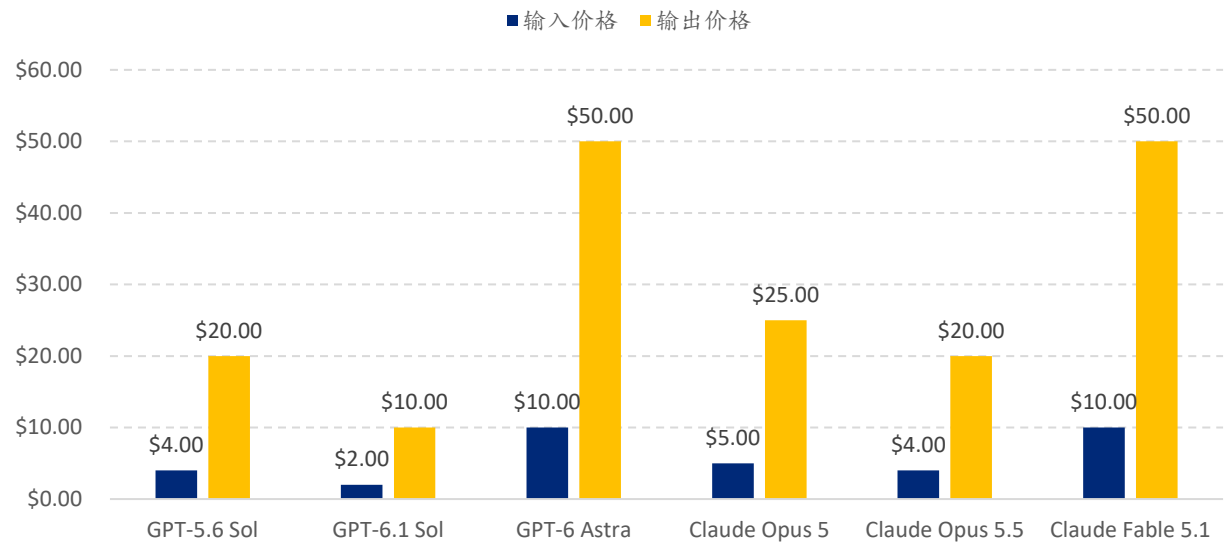
图表 23：同能力档位 Token 价格随时间下降



资料来源：Artificial Analysis、浦银国际

价格呈现分层结构，成熟能力降价与前沿任务溢价并存。同一能力档位的 Token 价格随时间下降，部分新旗舰模型则以更高的刊例价进入市场。按标准刊例价，Claude Opus 5.5 每百万 Token 输入与输出价格分别为 4 美元和 20 美元，GPT-6.1 Sol 分别为 2 美元和 10 美元，整体刊例价相较前代有所下降，而旗舰产品 GPT-6 Astra 与 Claude Fable 5.1 的标准输出价均为 50 美元。

图表 24：旗舰模型价格上移，次旗舰模型价格下行



注：数据截至 2026 年 11 月

资料来源：公司资料、浦银国际

复杂任务更看重完成质量与总成本，因此前沿模型仍有定价空间。尽管 Token 单价容易比较，但不同模型的分词方式、推理消耗和成功率并不相同。低价模型若需要更多重试和人工修改，最终成本可能更高；较强模型若能一次完成任务，则可能减少调用、工具使用、人工复核和失败损失。我们认为，随着智能体进入生产流程，在确保交付质量的前提下，客户将更看重单次任务成功交付的综合成本。

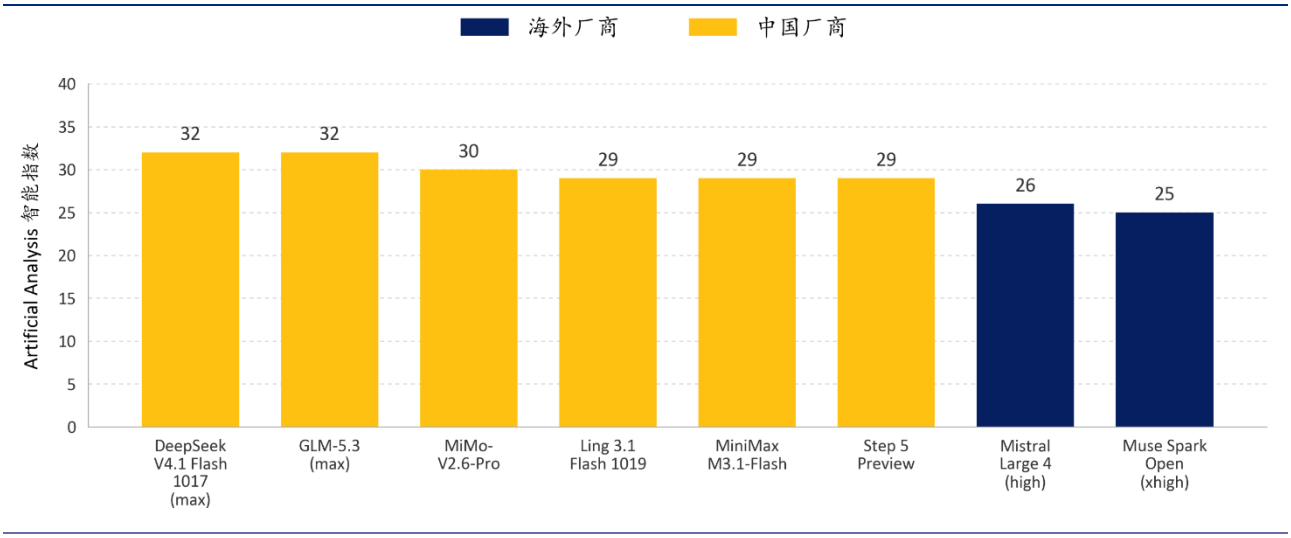
模型产品将按任务需求形成分工。开放权重和模型路由降低了接口层切换成本，标准化任务更依赖服务规模、成本效率和分发能力。我们认为，轻量模型承接高频、成本敏感的调用，旗舰模型服务于对完成质量与可靠性要求更高的任务，两者将共同支持模型服务市场扩张。

国产模型构建差异化优势，云与生态合作拓宽变现渠道

我们认为中国模型厂商在开放部署、多模态产品及部分产品定价上具备相对优势：1）开放权重及本地部署为企业提供了更多数据管理和定制选择；2）视频生成等多模态产品承接广告、电商及内容制作需求；3）部分产品较低的 API 定价有助于降低使用门槛。

开放权重与本地部署有助于满足企业的数据管理和定制需求。国内厂商提供了较丰富的开放权重模型，企业可结合自身算力、业务数据和任务要求选择部署环境，并进行模型适配。对具备开发及运维能力的客户，本地部署还支持自主安排版本更新、优化推理服务和管理调用资源，以适配自身的业务系统和工作流程。

图表 25：中国厂商在领先开放权重模型中占据重要位置



资料来源：Artificial Analysis、浦银国际

国产模型在视频生成领域已具备较强竞争力。根据截至 2026 年 11 月 6 日的 Artificial Analysis 带音频视频生成盲测榜，阿里 Wan 3.0 位列文生视频第一，MiniMax H3（768p）与字节 Dreamina Seedance 2.5 分别位列第三和第五；图生视频榜由 Dreamina Seedance 2.5 位列第一，Dreamina Seedance 2.0 和 MiniMax H3（768p）分别位列第三和第四。相关产品可用于广告素材、电商展示与内容创作，通过 API 及自有应用承接内容制作需求。

图表 26：Artificial Analysis 视频生成盲测排名

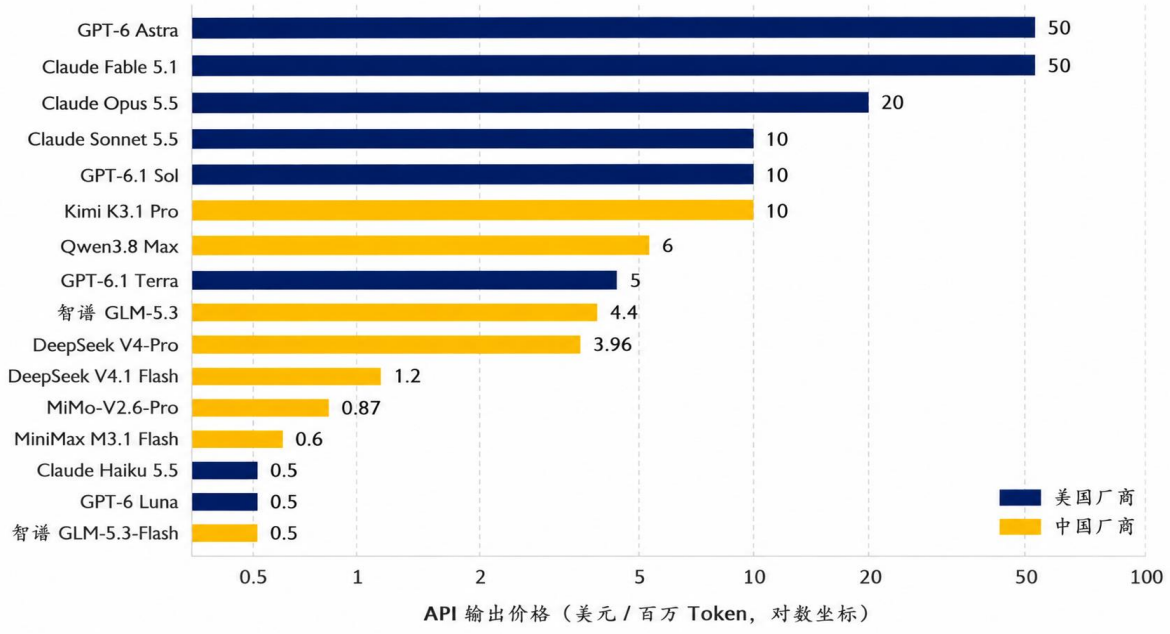
厂商	前沿模型	Elo 等级分	发布时间	API 价格 (USD/每分钟)
文生视频（带音频）				
阿里巴巴	Wan 3.0	1172	2026 年 8 月	$12.00
Utopai Studios	Utopai X (Based on MiniMax H3)	1169	2026 年 9 月	未提供 API
MiniMax	MiniMax H3 (768p)	1160	2026 年 7 月	$4.80
fal	FLUX 3 Max	1156	2026 年 10 月	$4.35
字节跳动	Dreamina Seedance 2.5	1155	2026 年 7 月	$34.12
图生视频（带音频）				
字节跳动	Dreamina Seedance 2.5	1159	2026 年 7 月	$34.12
Utopai Studios	Utopai X (Based on MiniMax H3)	1147	2026 年 9 月	未提供 API
字节跳动	Dreamina Seedance 2.0	1135	2026 年 3 月	$22.45
MiniMax	MiniMax H3 (768p)	1131	2026 年 7 月	$4.80
HeyGen	HeyGen Video 1 (Based on MiniMax H3)	1124	2026 年 10 月	$2.40

注：数据截至 2026 年 11 月 6 日；

资料来源：Artificial Analysis、浦银国际

部分国产模型的 API 定价低于海外旗舰，有助于覆盖成本敏感型需求。按截至 2026 年 11 月 6 日的公开标准输出刊例价，MiMo-V2.6-Pro、GLM-5.3 与 Kimi K3.1 Pro 分别为每百万 Token 约 0.87、4.40 与 10 美元，GPT-6 Astra 与 Claude Fable 5.1 均为 50 美元。

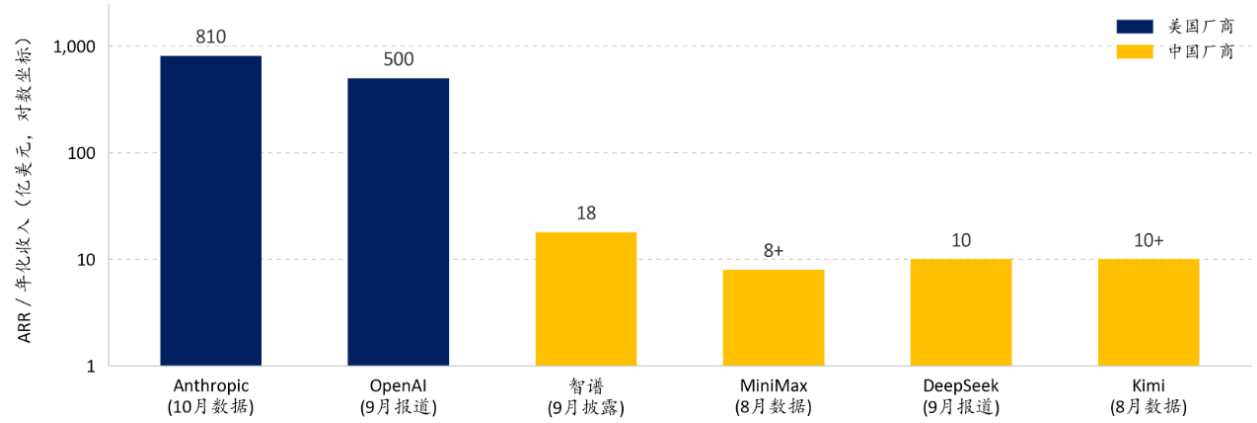
图表 27：中美部分模型 API 输出价格对比



资料来源：公司官网、浦银国际

国产独立厂商的商业化增长较快，但规模与海外龙头仍有差距。中国独立模型厂商披露或被报道的 ARR 仍在数亿至十数亿美元区间。据公司披露，智谱今年 9 月中旬 ARR 约 18 亿美元；MiniMax 8 月 ARR 达到 8 亿美元；据 The Information 和 Bloomberg 9 月报道，DeepSeek 的 ARR 在 API 提价后数月内翻倍至约 10 亿美元，月之暗面 8 月 ARR 突破 10 亿美元。而海外头部厂商已形成较大的商业化规模，据 Financial Times、Reuters、WSJ 等媒体报道，OpenAI 截至 2026 年 9 月末的 ARR 约 500 亿美元；Anthropic 截至 10 月末的 ARR 超过 800 亿美元。

图表 28：中美主要独立模型厂商 ARR

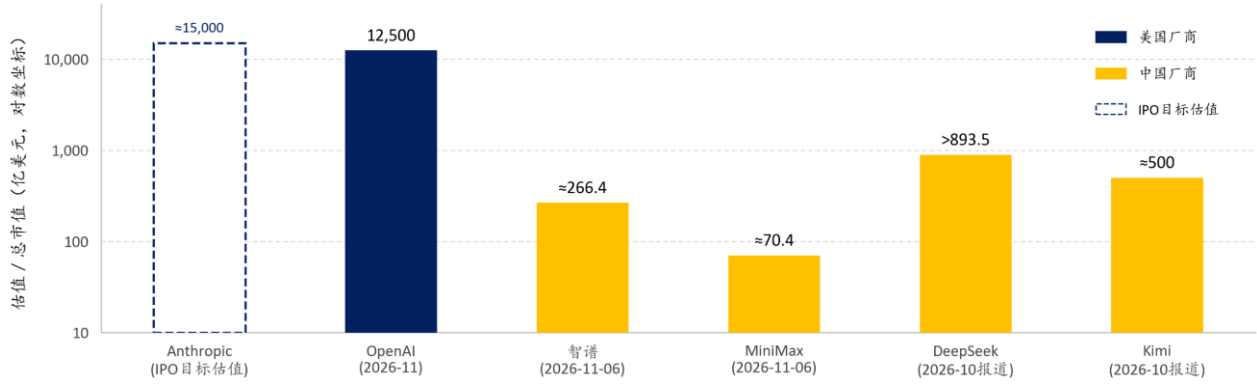


注：ARR 按公司披露或媒体报道的年化收入规模列示，各公司统计时点、年化方式及合作渠道收入确认口径有所不同。

资料来源：The Information、Bloomberg、Financial Times、Reuters、WSJ、公司资料、浦银国际

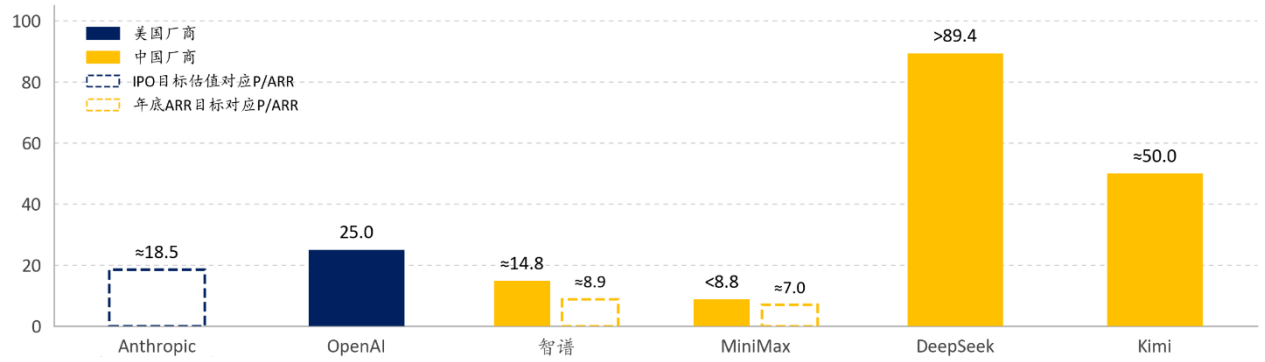
估值方面，根据公司资料和 Reuters、Financial Times、The Information、Bloomberg 等媒体报道，Anthropic 于 2026 年 5 月的投后估值为 9,650 亿美元，IPO 目标估值约 1.5 万亿；OpenAI 估值为 1.25 万亿美元（2026 年 11 月）；DeepSeek 估值超过人民币 6,000 亿元（2026 年 10 月报道）；月之暗面估值约 500 亿美元（2026 年 10 月报道）。

图表 29：中美主要独立模型厂商估值/市值



资料来源：The Information、Bloomberg、Financial Times、Reuters、Wind、公司资料、浦银国际

图表 30：中美主要独立模型厂商 P/ARR



注：市值截至 2026 年 11 月 6 日

资料来源：The Information、Bloomberg、Financial Times、Reuters、WSJ、Wind、公司资料、浦银国际

平台与独立模型厂商竞合并存

平台型厂商的优势在资源与分发，独立厂商则可依靠研发聚焦和差异化产品参与竞争。平台拥有云基础设施、应用入口、企业客户和主业现金流，在算力采购、获客及产品整合上具备综合优势；独立厂商则聚焦模型研发，可与多家平台合作，并通过开放权重吸引开发者。

独立模型厂商在产品布局与应用方向上各有侧重。智谱围绕 GLM 系列拓展编程、长程及企业专业任务，通过 API 和云托管合作扩大模型服务收入。MiniMax 同时布局文本、视频、语音等模型，通过开放平台和自有应用服务全球客户。DeepSeek 持续提升推理、编程及智能体能力，并通过开放权重和 API 提供服务。月之暗面则以 Kimi 拓展编程与知识工作，强化长程任务执行能力。

平台厂商主要结合云服务、协同办公、内容、终端入口推进应用。阿里巴巴将千问接入淘宝、支付宝、高德等产品形成交易闭环，并以千问办公、百炼 MaaS 和开源生态扩大影响力。字节跳动以豆包与 Seedance 覆盖文本和视频需求，依托抖音等超级 App 入口扩大模型使用，并以“豆包工作”联动飞书切入办公场景。腾讯以 WorkBuddy 服务办公场景，推动混元模型与协同办公产品整合，并依托微信内的原生智能体触达个人用户。百度将文心大模型嵌入搜索改造与企业级云服务。

竞合并存，平台大厂与独立厂商形成多层次合作。平台在自研模型之外，也接入第三方模型以覆盖不同任务的能力、成本和部署需求；独立厂商则可借助平台的企业客户与应用入口扩大使用。

云端模型服务平台和智能体产品为独立厂商拓展了分发与商业化渠道。以腾讯为例，腾讯云 TokenHub 接入 DeepSeek、智谱、Kimi 和 MiniMax 等厂商的模型，企业客户可通过统一接口调用；WorkBuddy 等办公智能体也支持使用多家第三方模型，将模型能力引入具体工作场景。在海外，云渠道已形成较大收入贡献，Anthropic 披露，2026 年上半年通过云服务商市场销售的收入占 48%。平台的基础设施、客户关系与应用入口，有助于独立厂商扩大服务覆盖，降低部署和获客投入。

商业授权与云托管分成，也使开放权重模型能够通过第三方平台获得收入。Kimi K3 和 GLM-5.3 等开放权重模型的许可对部分商业使用设有单独授权要求，可与云厂商签署托管 API 收入分成协议获得相关收入。我们认为，这类合作将帮助模型能力较强的独立厂商更快扩大商业规模。

推理提效与规模效应驱动盈利改善

我们认为，行业盈利改善主要由两方面推动：一是调用规模增长、推理优化与国产算力应用带动毛利率提升；二是模型商业化收入增长逐步摊薄训练等研发开支，推动研发费用率下降。

推理成本是影响毛利的主要变量。训练投入支撑模型能力提升与持续迭代，通常体现为研发费用；而直接用于客户服务的推理成本与调用量相关，通常计入收入成本，对毛利率的影响更直接。API、Agent 和多模态应用的收入增长，也会同步增加推理负载。单次任务包含的工具调用、上下文处理和重试次数，都会影响服务成本。厂商通过更高效的模型和部署方法承接这些需求，是实现盈利改善的主要途径。

国产芯片为推理侧的成本优化增加了选择。智谱披露，2026 年初将国产芯片纳入主力推理算力后，已实现十万级规模化部署。MiniMax 也表示，公司计划于 2026 年第四季度提高国产芯片使用占比。国产芯片的规模化使用扩大了推理算力选择，也为厂商优化服务成本提供了空间。

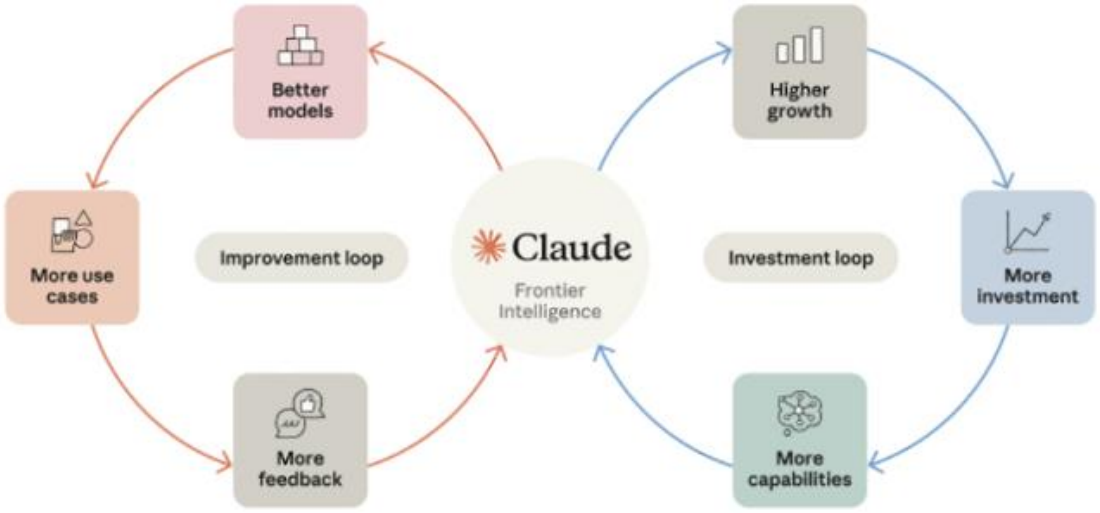
单位推理成本下降有望逐步转化为毛利率提升。智谱在 1H26 中期业绩中表示，其单位 Token 推理成本较年初下降约 80%，成本优化的效果仍在逐步释放。部分厂商的 API 业务已实现较高毛利率。据 The Information 报道，DeepSeek 2026 年 1 至 7 月 API 业务毛利率为 82.9%。我们认为，调用增长带动集群利用率提升后，推理降本有望更充分地体现在毛利率中。

收入放量有助于逐步降低研发费用率。模型厂商仍处于较大规模研发投入阶段，2026 年上半年智谱与 MiniMax 经调整净亏损均约 2.9 亿美元，各自研发投入约 3 亿美元，持续研发投入对当期利润形成较大影响。我们认为，收入增速持续高于研发开支增速，有助于降低研发费用率；新增毛利则可覆盖更多研发费用，推动亏损收窄。

Anthropic：大模型巨头如何率先跑通盈利闭环

Anthropic 的经营数据展示了企业用量增长对盈利的带动。公司聚焦企业市场，通过模型升级推动客户扩大使用，增加的收入用于支持后续研发与算力投入。2026 年二季度，公司已实现 GAAP 经营盈利；三季度收入与调整后经营利润继续增长。我们认为，这些经营数据为前沿模型厂商通过客户持续付费支持研发、逐步改善盈利提供了参照。

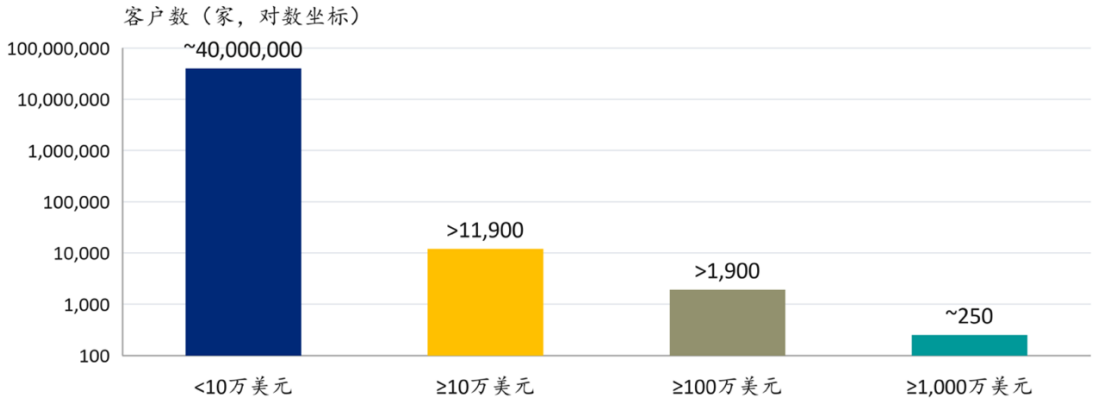
图表 31：Anthropic 自我增强循环



资料来源：公司招股书、浦银国际

用量计费使客户使用加深转化为收入增长，大客户已形成较高的付费规模。2026 年上半年，Anthropic 按用量计费收入占比 80.2%，订阅收入占 19.8%。企业通过 Claude Platform 及 Code、Cowork 等产品使用模型，任务量与调用强度增加可直接带动按量收入增长。截至 2026 年 9 月末，过去 12 个月收入贡献至少 100 万美元的客户超过 1,900 家，收入贡献至少 1,000 万美元的客户约 250 家。

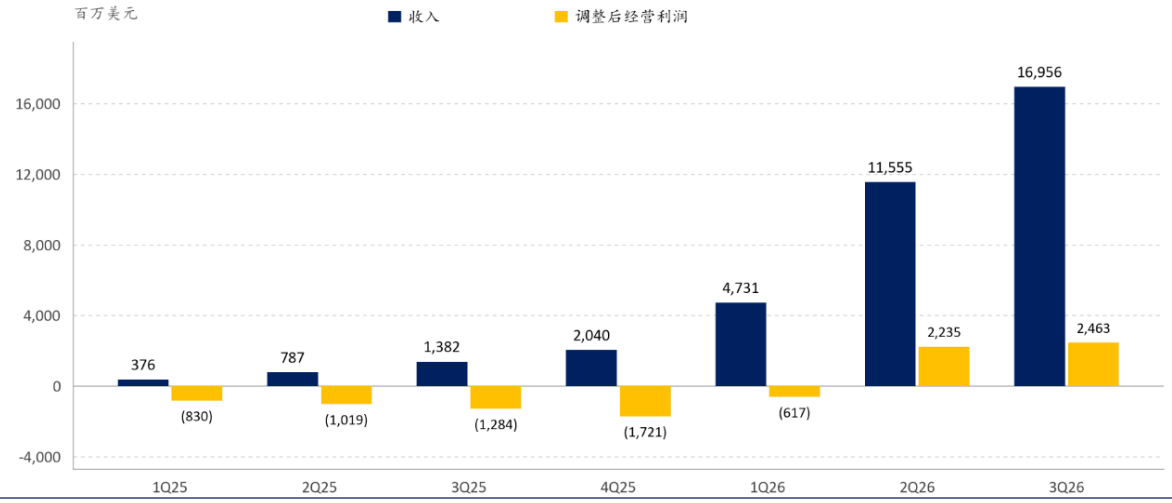
图表 32：过去 12 个月收入贡献达百万美元的客户超过 1,900 家（2026 年 9 月）



资料来源：公司招股书、浦银国际

收入扩张带动经营杠杆释放，算力采购影响季度利润。2026 年第二季度，Anthropic 收入环比增长 144.2%至 115.55 亿美元，快于算力及基础设施费用 64.1%的增幅，相关费用率由一季度的 71.6%降至 48.1%，推动 GAAP 经营利润由亏损 16.67 亿美元转为盈利 9.74 亿美元。三季度收入进一步环比增长 46.7%至 169.56 亿美元，算力及基础设施费用率则回升至 58.2%，其中约 50 亿美元来自单位成本较高的短期算力采购；叠加约 19.6 亿美元非现金捐赠相关费用，三季度 GAAP 经营亏损 5.67 亿美元，调整后经营利润升至 24.63 亿美元，对应利润率约 14.5%。

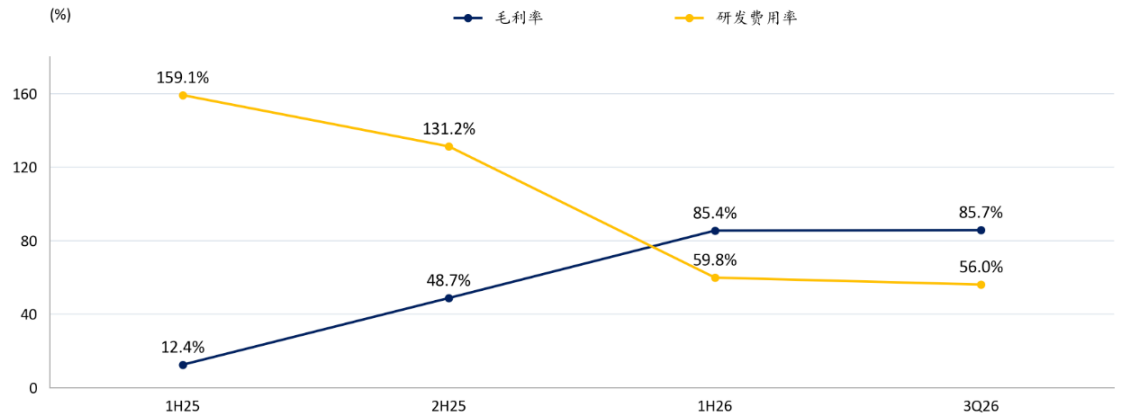
图表 33：收入快速增长，调整后经营利润连续两个季度为正



资料来源：公司招股书、浦银国际

新模型持续贡献收入，收入扩张降低研发费用率。按费用功能分类测算，公司 2026 年第三季度毛利率约 85.7%，与上半年的 85.4%基本持平；同期研发费用 94.99 亿美元，已超过 2025 年全年的 63.38 亿美元，研发费用率则由 2025 年的 138.2%降至 56.0%。公司披露，三季度超过 80%的收入来自此前六个月发布的模型版本。我们认为，新模型能够较快转化为客户消费，收入增长也使公司能够在增加研发投入的同时降低研发费用率。

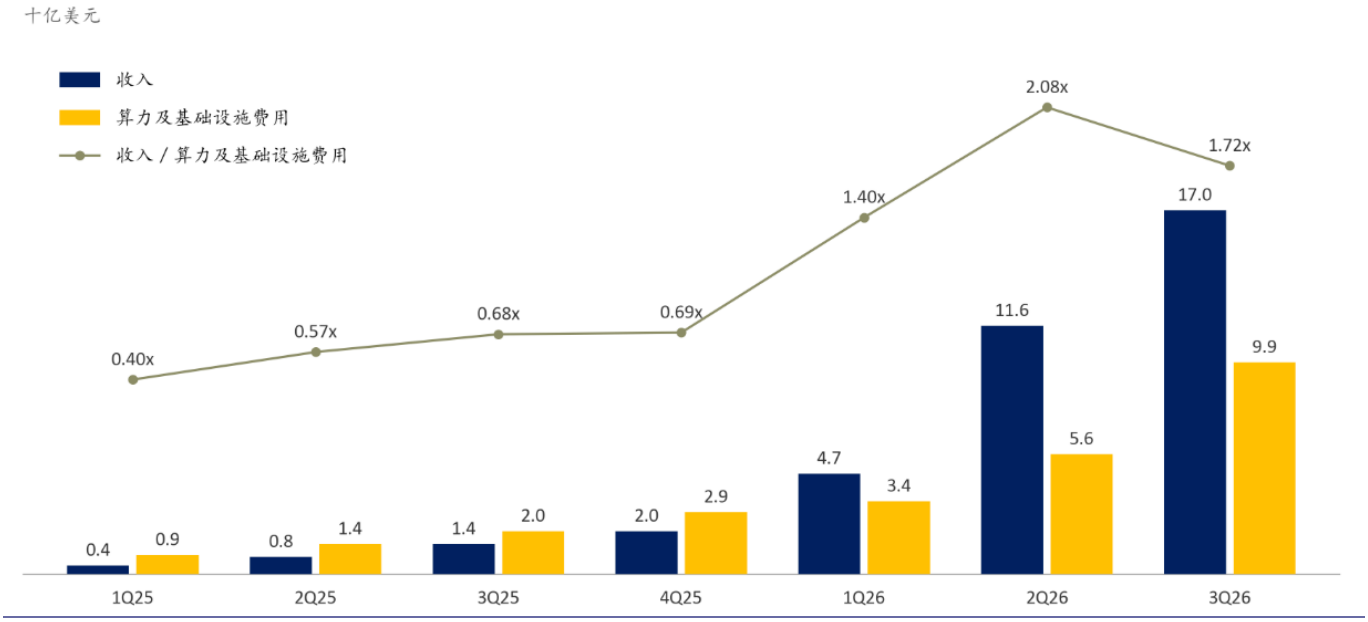
图表 34：毛利率维持高位，研发费用率下降



资料来源：公司招股书、浦银国际

共享算力池与工程优化支持公司提高算力利用效率。公司表示，训练、客户服务与内部使用共享算力资源，研究团队与推理基础设施团队协同推进算法及工程优化，提高单位算力可承接的工作量。从收入对投入的覆盖看，每 1 美元算力及基础设施费用对应的收入，由 2025 年一季度约 0.40 美元升至 2026 年二季度约 2.08 美元，三季度受短期算力采购增加等因素影响回落至约 1.72 美元。

图表 35：每美元算力及基础设施支出对应的收入趋势



资料来源：公司招股书、浦银国际

智谱与 MiniMax：企业端增长提速，商业化路径各有侧重

我们看好智谱在复杂任务中的变现能力，以及 MiniMax 企业端与多模态业务的增长空间。智谱调用量与 API 平均售价同步提升，体现了复杂任务的付费潜力，云托管及 Co-work 进一步拓宽收入来源；MiniMax 开放平台与自有应用共同增长，文本模型升级与降本有助于扩大企业端调用，视频、语音和音乐产品则面向全球内容制作需求。

智谱加速转向持续模型服务，MiniMax 由开放平台与自有应用共同推动增长。智谱 1H26 开放平台及 API 收入占比由上年同期的 15.2%升至 86.5%，本地化部署收入同比下降 20.5%，部分原采用私有化部署的客户逐步转向平台调用。同期，MiniMax 开放平台及其他基于 AI 的企业服务占收入 63.4%，AI 原生产品收入同比增长 100.9%。AI 原生产品收入占比降至 36.6%，主要由于平台业务增长更快。MiniMax 同期海外收入占比 60.8%，海外市场是其重要收入来源。

图表 36：智谱与 MiniMax 商业化路径比较

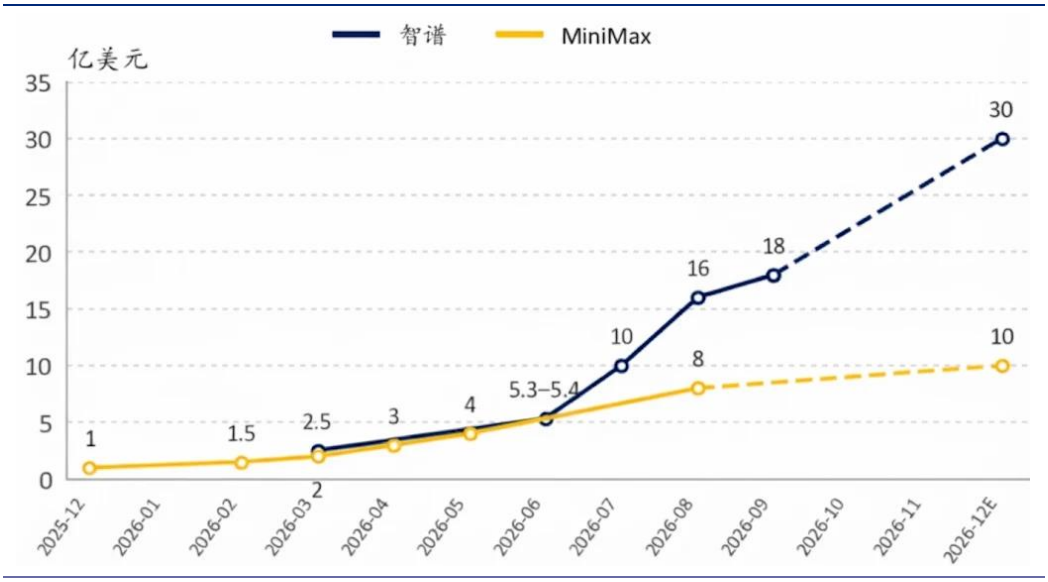
维度	智谱	MiniMax
收入结构	开放平台及 API 占 1H26 收入 86.5%，本地化部署占 13.5%	开放平台及其他企业服务占 1H26 收入 63.4%，AI 原生产品占 36.6%
产品侧重	GLM 旗舰与 Flash，重点拓展编程、长程及专业任务	M 系列文本模型、H 系列视频模型，以及语音和音乐产品
客户与分发	企业及开发者；原厂 API 与合作云平台托管	全球企业、开发者及个人用户；开放平台与自有应用
收入来源及变现方式	API 按量收费及订阅、云托管分成、Co-work 专业任务采购	API 按量收费、Token 订阅计划及自有应用付费
主要观察点	客户调用及实际售价、托管与订单兑现、平台毛利	企业端客户采用、多模态付费、全球业务增长及服务毛利

资料来源：公司公告、浦银国际整理

智谱量价同步提升，MiniMax 以较低使用成本拓展需求。根据智谱 1H26 业绩会披露，平台 Token 调用量较年初增长超过 40 倍，同期 API 平均售价上升约 101%。我们认为，量价齐升体现了公司在扩大调用规模的同时提升商业变现的能力。MiniMax 于 2026 年 10 月 19 日发布 M3.1 Flash，并披露其智能体任务使用成本较前代下降 70%。任务使用成本下降，有助于客户增加调用。

两家公司 ARR 持续扩大，管理层对年底增长保持积极预期。智谱管理层在今年 9 月中旬披露，9 月上旬 ARR 约 18 亿美元，并将年底指引上调至 30 亿美元；MiniMax 管理层在 8 月交流中披露，当月 ARR 约 8 亿美元，按周度收入年化，企业端占比超过 80%，并表示非常有信心实现 2026 年底 10 亿美元的 ARR 目标。

图表 37：智谱及 MiniMax ARR 趋势



E=公司目标

资料来源：公司资料、浦银国际

智谱平台毛利率持续改善，MiniMax 服务优化有望带动毛利修复。智谱收入向毛利率较低的云端业务切换，集团毛利率回落；开放平台自身毛利率则持续提升，收入增长也在增加毛利贡献。MiniMax 毛利率受新品上线初期补偿、Token 计划及文本与多模态收入组合影响，管理层预计补偿影响退出、套餐优化及推理效率提升将推动下半年毛利率回升。

收入增速高于研发开支，两家公司的研发费用率均已下降。1H26 智谱和 MiniMax 的研发费用率分别为 223.4%及约 254.7%，研发仍是主要费用项。智谱销售及行政费用、MiniMax 销售费用同比下降，也有助于减轻持续研发投入对当期利润的压力。

图表 38：智谱与 MiniMax 主要财务指标比较（1H26）

指标	智谱	MiniMax
收入	9.54 亿元	1.166 亿美元
收入同比增速	399.7%	283.1%
开放平台相关分部收入	8.25 亿元	7,393 万美元
开放平台相关分部收入占比	86.5%	63.4%
集团毛利率	26.4%	17.9%
开放平台分部毛利率	24.6%	未单列披露
研发开支	21.31 亿元	2.969 亿美元
研发开支同比增速	33.6%	138.8%
研发开支率	223.4%	约 254.7%
经调整净亏损	19.64 亿元	2.930 亿美元

资料来源：公司公告、浦银国际整理

资金储备支持模型研发与算力扩张。为满足持续研发与算力建设的投入需求，两家公司通过 IPO 及后续配售、可转债融资补充资金。2026 年 1 月至 9 月，智谱上述融资净额合计约 755 亿港元；MiniMax 自 1 月 IPO 起，连同 7 月配售及可转债发行，融资净额合计约 211.7 亿港元。

图表 39：智谱与 MiniMax 资金储备及资本安排

事项	智谱	MiniMax
现金及现金等价物	人民币 39.94 亿元	9.309 亿美元
IPO 及上市后融资净额	2026 年 1 月至 9 月合计约 755 亿港元	IPO 及 7 月融资合计约 212 亿港元
--IPO	全球发售 H 股，价格 116.2 港元/股，募资净额约 49 亿港元	全球发售 A 类股，价格 165 港元/股，募资净额约 53 亿港元
--2026 年 7 月融资	配售新股：价格 1,588 港元/股，募资净额约 314 亿港元	配售新股：价格 268 港元/股，募资净额约 94 亿港元；零息可转债：本金 65 亿港元，2027 年 7 月到期，募资净额约 64 亿港元；初始转股价为 335 港元/股
--2026 年 9 月融资	配售新股：价格 714 港元/股，募资净额约 157 亿港元；零息可转债：本金人民币 201.4 亿元，2027 年 9 月到期，募资净额 30.11 亿美元（约 236 亿港元）；初始转股价为 892.5 港元/股	/
后续资本安排	A 股发行方案已获股东大会批准，相关辅导工作已完成；首发前已上市 H 股的一年期禁售安排将于 2027 年 1 月届满，涉及股份占总股本约 36.6%	已启动科创板上市辅导，10 月披露的第二期辅导进展提及后续验收准备；创始人、部分管理层及紧密联系人所涉股份禁售期于 2027 年 1 月 8 日届满，涉及股份占总股本约 26.0%

注：现金及现金等价物截至 2026 年 6 月 30 日；

资料来源：公司公告、21 世纪经济报道、浦银国际整理

投资风险

- 行业竞争及价格压力加剧。国内外厂商持续迭代，平台企业依托资金与渠道扩大投入。若模型表现趋近，开放权重和低价产品可能加快客户切换、压低售价；若降价幅度超过单位成本降幅，收入及毛利改善可能不及预期。

- 客户使用与商业化增长不及预期。若编程、办公及多模态产品的效率回报不足，或企业接入与采购进度较慢，试用可能难以转为持续付费。客户调用、续费及订单交付放缓，将影响收入增长与现金回收。

- 模型迭代及研发回报不及预期。前沿模型需要持续研发投入，若新品延迟，或任务表现、可靠性落后于竞争对手，可能影响客户采用和定价。研发开支增速若持续高于收入增速，也将延缓亏损收窄。

- 算力供给与部署进度受限。芯片、数据中心及电力供应不足，或集群交付、国产芯片适配延迟，可能影响模型训练与服务。算力采购及租赁价格上涨，或临时补充高成本资源，也会增加支出。

- 毛利率改善不及预期。新增集群利用率提升较慢、低价套餐或较低毛利业务占比上升，可能抵消推理降本收益。新品上线补偿、服务投入和模型切换成本，也可能压低阶段性毛利。

- 融资、偿债及股本摊薄风险。研发与算力支出通常先于收入回收，商业化放缓或融资环境转弱可能加大现金消耗，影响后续投入。可转债到期未换股将增加偿付需求，增发或转股则可能摊薄股东权益。

- 盈利兑现及估值回落风险。部分厂商尚未盈利，若收入或盈利改善低于预期，市场可能下调预测和估值倍数。ARR 统计时点及年化口径不同，也可能影响估值比较，且不能直接等同于已确认收入。

- 信息安全、知识产权及合规风险。数据泄露、内容侵权或智能体越权执行，可能导致客户停用、诉讼赔偿及整改支出。数据保护、跨境服务及海外准入要求变化，也可能限制产品部署并增加合规成本。

浦银国际研究

首次覆盖 | 互联网行业

浦銀國際 SPDB INTERNATIONAL

智谱（2513.HK）：前沿模型加速商业化

我们首次覆盖智谱（2513.HK），给予“买入”评级，目标价 750 港元。我们看好公司在复杂编程与高价值专业任务中的商业化兑现能力，云托管分成与 Co-work 专业任务采购有望进一步拓宽变现边界。随着国产算力集群规模效应释放与收入放量，公司盈利能力有望改善。

- **开放平台成为收入主体，云端服务高速放量。**智谱是国内最早布局大模型研发的领军团队之一，技术源自清华大学知识工程实验室的深厚积累。公司主动收缩传统本地化软件授权业务，全面转向云端持续模型服务，1H26 开放平台及 API 收入同比增长 2,735.7%，占总收入比重由 1H25 的 15.2%提升至 86.5%。目前公司围绕编程与长程任务持续迭代 GLM 系列，通过旗舰模型与高性价比 Flash 版本形成产品分层，全方位覆盖不同梯度的客户需求。
- **调用量与平均售价同步提升，高价值任务支撑付费需求。**截至 2026 年 8 月 31 日，公司 MaaS 平台 Token 调用量较年初增长超过 40 倍，同期 API 平均售价提高约 101%。我们认为，复杂编程与长程智能体任务高度依赖交付质量以减少人工返工和降低总成本，为前沿模型保留了充足的溢价空间。受益于 10 万卡级国产芯片规模化推理与全栈工程优化，公司单位 Token 推理成本较年初下降 80%，驱动 1H26 开放平台毛利率较 2H25 提升 2.2 个百分点至 24.6%，降本增效成果正逐步兑现。
- **云托管强化模型分发，Co-work 切入高价值专业场景。**公司已与国内外多家头部云厂商签署托管 API 收入分成协议，GLM-5.3 已上线 Amazon Bedrock、腾讯云等主流平台，依托云厂商的算力底座与企业客群实现轻资产扩张。同时，公司通过 Co-work 将长程执行能力延伸至可验证的专业工作流，率先在网络安全领域跑通商业闭环。截至 9 月中旬，Co-work 累计订单已超人民币 10 亿元，预计于未来一至两年随交付逐步确认收入。基于强劲的业务势头，管理层于 9 月将 2026 年底 ARR 指引由 24 亿美元上调至 30 亿美元。
- **首予“买入”评级，目标价 750 港元。**我们预计公司 2026 年末 ARR 为 31 亿美元，参考海外头部模型厂商估值，给予 2026E 15x P/ARR 估值，目标价 750 港元，对应 2027E 13.2x P/S。短期催化剂包括新模型发布、年底 ARR 兑现超预期、Co-work 垂直场景拓展及 A 股上市进展。短期主要关注订单交付节奏、毛利率爬坡进度，以及 2027 年 1 月首发前股东禁售期届满带来的潜在流动性压力。
- **投资风险：**模型及行业竞争加剧；客户调用与 ARR 增长不及预期；毛利率改善及订单转化不及预期；监管合规与资金消耗风险。

图表 1：盈利预测和财务指标

人民币百万元	2024	2025	2026E	2027E	2028E
营业收入	312	724	6,413	23,667	33,779
经营利润	(2,541)	(3,780)	(4,041)	(2,114)	486
经调整净利润	(2,466)	(3,182)	(3,039)	(869)	1,671
目标 P/S			48.6	13.2	9.2

E=浦银国际预测；资料来源：公司报告、浦银国际

杨子超，CFA
科技分析师
charles_yang@spdbi.com
(852) 2808 6409

赵丹
首席科技软件分析师
dan_zhao@spdbi.com
(852) 2808 6436

2026 年 11 月 11 日

智谱（2513.HK） 买入

目标价（港元）	**750**
潜在升幅/降幅	**+76%**
目前股价（港元）	426.2
52 周内股价区间（港元）	116.1-2980
总市值（百万港元）	207,810
近 3 月日均成交额（百万港元）	4,910

注：截至 2026 年 11 月 6 日收盘价

市场预期区间



股价表现



注：截至 2026 年 11 月 6 日收盘价
资料来源：Bloomberg、浦银国际

浦银国际 首次覆盖 智谱（2513.HK）

本研究报告由浦银国际证券有限公司分析师编制，请仔细阅读本报告最后部分的分析师披露、商业关系披露及免责声明。

财务报表分析与预测

利润表

人民币百万	2024	2025	2026E	2027E	2028E
收入	312	724	6,413	23,667	33,779
销售成本	(137)	(428)	(4,606)	(16,174)	(20,646)
毛利	176	297	1,808	7,493	13,133
销售及营销开支	(387)	(391)	(451)	(1,183)	(1,689)
一般及行政开支	(134)	(505)	(267)	(710)	(1,013)
研发开支	(2,195)	(3,180)	(5,131)	(7,714)	(9,944)
经营利润	(2,541)	(3,780)	(4,041)	(2,114)	486
其他收入	19	15	283	513	311
金融资产减值亏损	(17)	(22)	(32)	(15)	(25)
融资成本	(38)	(74)	(127)	(53)	(65)
分占联营公司损益	21	55	144	-	-
公允价值变动及其他	(403)	(912)	28	-	-
除税前亏损	(2,941)	(4,697)	(3,713)	(1,654)	732
所得税	-	-	(2)	-	(37)
期内亏损	(2,941)	(4,697)	(3,715)	(1,654)	696
非控股权益	(2)	(20)	(1)	-	-
归属于母公司股东	(2,939)	(4,677)	(3,714)	(1,654)	696
经调整净亏损	(2,449)	(3,160)	(3,007)	(854)	1,696

现金流量表

人民币百万	2024	2025	2026E	2027E	2028E
净利润	(2,941)	(4,697)	(3,715)	(1,654)	696
折旧与摊销	280	281	316	1,747	3,125
股份薪酬	24	558	686	800	1,000
营运资金变动	(0)	704	(3,022)	(185)	(563)
其他非现金调整	410	928	(46)	53	65
经营活动现金流	(2,228)	(2,225)	(5,780)	761	4,323
资本开支	(126)	(11)	(1,187)	(6,734)	(5,010)
无形资产购置	(7)	(12)	(30)	(40)	(50)
投资及联营公司净变动	84	(353)	(1,385)	(75)	200
投资活动现金流	(49)	(376)	(2,602)	(6,849)	(4,860)
借款净变动	267	1,252	21,704	(20,565)	(400)
股权融资	3,011	1,579	44,604	-	-
租赁付款	(178)	(211)	(218)	(170)	(205)
其他融资活动	212	(6)	(40)	(60)	(48)
融资活动现金流	3,312	2,614	66,050	(20,795)	(653)
汇兑影响及其他	0	(5)	(76)	-	-
现金净变动	1,036	8	57,591	(26,884)	(1,190)
期初现金	1,249	2,268	2,276	59,868	32,983
期末现金	2,285	2,276	59,868	32,983	31,793

资产负债表

人民币百万	2024	2025	2026E	2027E	2028E
物业、厂房及设备	866	656	1,382	6,837	9,042
无形资产	50	51	89	108	131
商誉	39	39	284	284	284
于联营公司的权益	201	338	532	532	532
其他非流动资产	203	198	1,239	1,005	896
非流动资产合计	1,360	1,282	3,526	8,765	10,885
现金及银行结余	2,269	2,259	59,895	33,010	31,820
短期投资及定期存款	43	486	1,700	1,700	1,500
贸易及其他应收款项	667	699	3,363	5,190	6,598
存货及合约成本	32	126	95	114	137
合约资产	5	2	13	47	68
流动资产合计	3,016	3,571	65,066	40,062	40,123
资产总额	4,376	4,854	68,592	48,827	51,008
贸易及其他应付款项	603	1,329	1,526	2,461	3,073
合约负债及递延收入	110	329	686	1,356	1,524
银行借款及可转债	269	689	22,388	1,800	1,400
租赁负债	671	545	412	483	589
向投资者发行的金融工具	6,677	10,073	-	-	-
负债总额	8,331	12,965	25,012	6,101	6,586
权益总额	(3,955)	(8,111)	43,586	42,732	44,428
权益及负债总额	4,376	4,854	68,598	48,833	51,014

主要财务比率

人民币百万	2024	2025	2026E	2027E	2028E
盈利增速					
营业收入增速	150.9%	131.9%	785.4%	269.0%	42.7%
毛利增速	118.5%	68.7%	509.4%	314.5%	75.3%
经营利润增速	nm	nm	nm	nm	nm
净利润增速	nm	nm	nm	nm	nm
经调整净利润增速	nm	nm	nm	nm	nm
盈利能力比率					
毛利率	56.3%	41.0%	28.2%	31.7%	38.9%
经营利润率	-813.2%	-521.9%	-63.0%	-8.9%	1.4%
净利率	-941.4%	-648.4%	-57.9%	-7.0%	2.1%
经调整净利率	-783.8%	-436.3%	-46.9%	-3.6%	5.0%
每股指标（元）					
EPS			(8.11)	(3.39)	1.43
经调整EPS			(6.61)	(1.75)	3.48
估值（倍）					
目标P/S			48.6x	13.2x	9.2x
经调整目标P/E			nm	nm	183.7x

资料来源：公司公告、Bloomberg、浦银国际预测

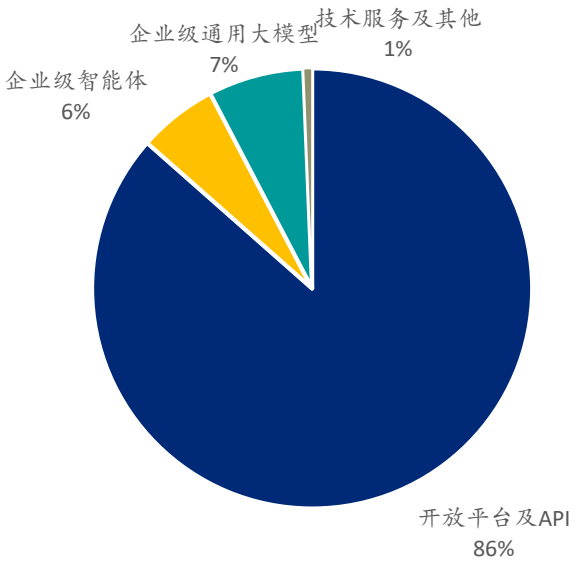
开放平台成为收入主体，云端服务高速放量

开放平台是公司模型商业化的主要渠道。企业和开发者通过 API 将模型接入自身产品与工作流程，并可持续使用后续升级版本。我们认为，模型在编程与复杂任务中的表现改善，有助于吸引新增客户、加深既有客户使用，推动平台收入增长。

开放平台及 API 已成为公司的收入主体。1H26 公司开放平台及 API 收入占比由 1H25 的 15.2%升至 86.5%，本地化部署收入占比相应由 84.8%降至 13.5%。公司在 Coding 阶段主动收缩本地化部署模型软件授权业务，转向持续提供模型服务。部分原采用私有化本地部署的企业级通用大模型客户，也逐步转向开放平台及 API 服务。通过平台调用，客户可减少自建部署和版本维护工作，并更便捷地使用后续升级的模型。

分部署方式看，1H26 云端部署（开放平台及 API）收入人民币 8.25 亿元，同比增长 2,735.7%；本地化部署收入人民币 1.29 亿元，同比下降 20.5%。分产品线看，开放平台及 API、企业级智能体、企业级通用大模型、技术服务及其他收入分别为人民币 8.25 亿元、0.56 亿元、0.67 亿元与 0.06 亿元。企业级通用大模型收入同比下降 54.6%，对应公司主动收缩的业务；企业级智能体收入同比增长 304.4%，需求仍在扩大。

图表 2：智谱收入结构（1H26）



资料来源：公司资料、浦银国际

编程与长程任务能力支撑产品竞争力

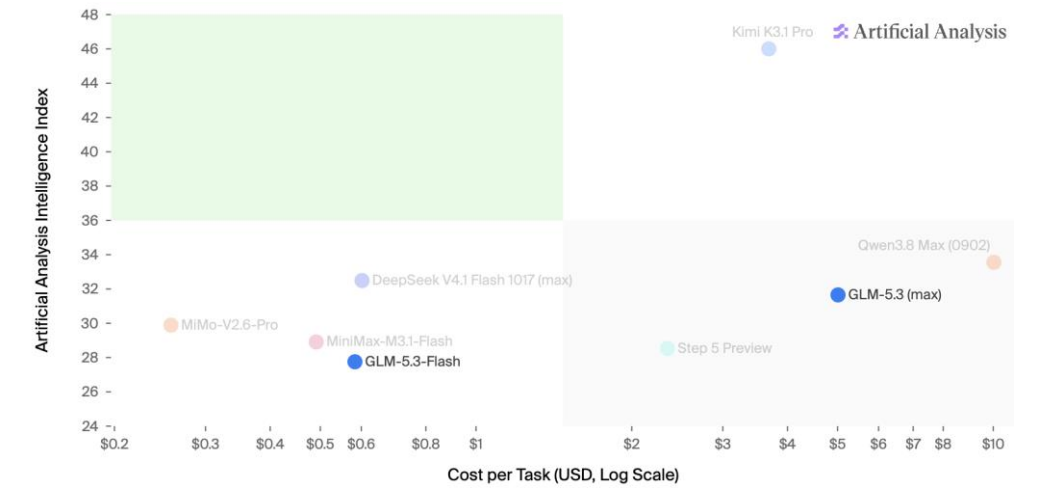
公司围绕编程和长程任务持续改进模型能力，并以旗舰和 Flash 产品覆盖不同任务及价格区间。公司提出由 Chat 到 Coding、Agent、Co-work，最终到 Autonomous AI 的能力发展路径：Coding 强调结果可验证，Agent 强调长程规划与错误恢复，Co-work 强调专业人士可复核的可靠交付，Autonomous AI 强调模型自主判断任务完成质量。

后训练正在成为公司模型升级的重要来源。2026 年 8 月发布的 GLM-5.3 与 GLM-5.2 采用相同架构、总参数与激活参数，主要通过扩展长程任务环境和强化学习提升能力。公司自建的 Z.ai Code Bench 评测显示，GLM-5.3 端到端完成率较 GLM-5.2 提升约 50%。我们认为，在同代基座上挖掘后训练潜力，使公司能够在下一代大规模预训练完成前，持续改善任务执行表现。

轻量级模型拓展了成本敏感型需求的覆盖。GLM-5.3-Flash 模型总参数 320B、激活参数 18B，采用稀疏注意力与线性注意力混合架构，结合 30T 级多模态预训练语料，按 API 标准报价，输入及输出价格均约为 GLM-5.2 的一成。据公司披露，其基准测试与实际应用表现均优于 GLM-5.2。我们认为，Flash 兼顾模型能力与调用成本，有助于吸引预算更敏感的开发者，并支持客户将模型用于更高频的日常任务。

公司同步推进新一代基座模型研发。据 1H26 业绩交流，公司在现有基座上推进后训练的同时，也在预训练更大参数的模型，该模型计划于 2026 年内发布。我们认为，新一代基座有望进一步提升复杂编程和长程任务的完成质量，使模型能够承担更高难度的工作，帮助公司拓展高价值任务的付费需求。

图表 3：智谱模型能力与任务成本



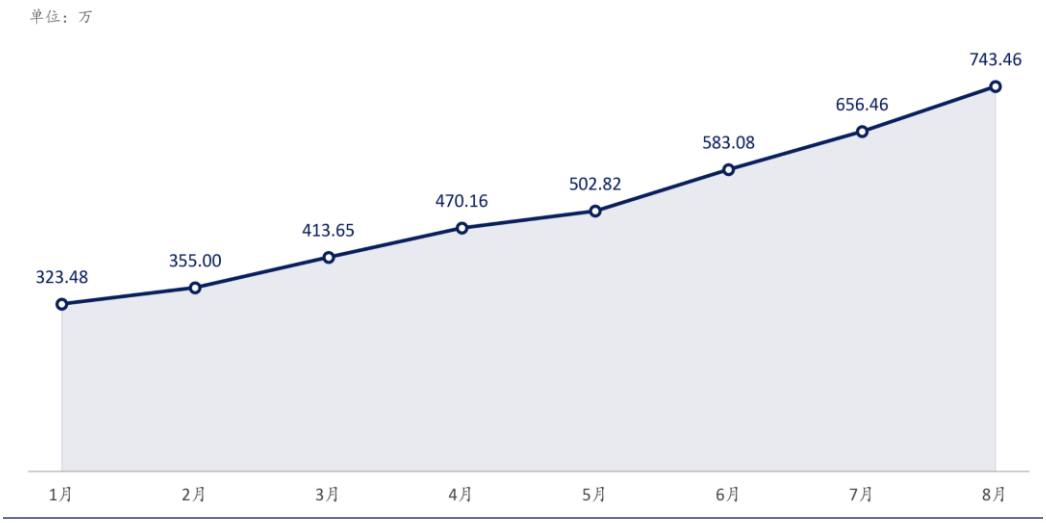
资料来源：Artificial Analysis、浦银国际

客户使用持续加深，调用量与平均售价同步增长

模型能力的商业价值已在客户使用和平均售价中体现。据公司 2026 年 8 月 31 日中期业绩披露，截至公告日，MaaS 平台 Token 调用量较年初增长超过 40 倍，同期 API 平均售价上升约 101%，前十大用户日均调用量增长 98 倍，付费日活用户数增长 603%。我们认为，量价同步增长反映了高价值任务需求扩大。客户对任务完成质量和返工成本的重视，有助于公司在复杂编程与长程任务中保留定价空间。

高付费客户群扩大，GLM 已获多家头部互联网公司采用。据公司 1H26 业绩会披露，按年化收入贡献划分，10 万美元档客户达到 115 家、100 万美元档达到 37 家；中国前十大互联网公司中 9 家为公司客户，其中 4 家已将 GLM 作为全球首要选择。

图表 4：智谱平台客户数（2026 年）



资料来源：公司资料、浦银国际

Coding API 与 Coding Plan 共同推动编程业务增长。管理层在 9 月中旬交流会上表示，Coding Plan 于 7 月底解除限售后销量增长超过 15 倍，ZCode 用户数量从 8 月中旬至 9 月中旬翻倍至约 200 万。API 按量调用便于企业将模型集成到已有产品和开发流程，Coding Plan 订阅覆盖开发者的持续使用需求；ZCode 通过上下文管理、工具调用和任务编排，将模型能力用于完整的开发任务。

云托管分成与专业任务采购丰富变现渠道

云平台合作与 Co-work 将进一步扩大智谱的收入来源。托管 API 合作借助云平台的算力和客户基础，扩大模型服务的分发规模，并为第三方部署提供分成变现方式；Co-work 则通过网络安全等专业任务，承接企业的专业服务需求。

公司已与多家海外及国内头部云服务商签署收入分成协议。GLM 系列开放权重模型以托管 API 形式提供服务，公司按约定比例获得分成，相关收入自 2026 年 10 月起确认。10 月 5 日，AWS 宣布 GLM-5.3 上线 Amazon Bedrock，符合条件的企业客户可通过托管 API 调用，并使用跨区域推理及提示词缓存等功能。腾讯云也公告，自 10 月 9 日起将部分旧版 GLM 服务切换至 GLM-5.3。模型进入企业已有的云服务采购体系，可减少接入与部署工作，便于客户在原有产品和开发流程中试用。

云平台承接推理服务，可降低新增收入对自有算力的依赖。管理层在 9 月交流会上粗略估计，GLM-5.2 原厂仅承接约 10%的全球调用，其余约 90%由第三方部署开放权重模型满足。通过合作云平台承接推理和客户接入，公司可以在自有 API 之外提供托管服务，并按约定参与收入分成。

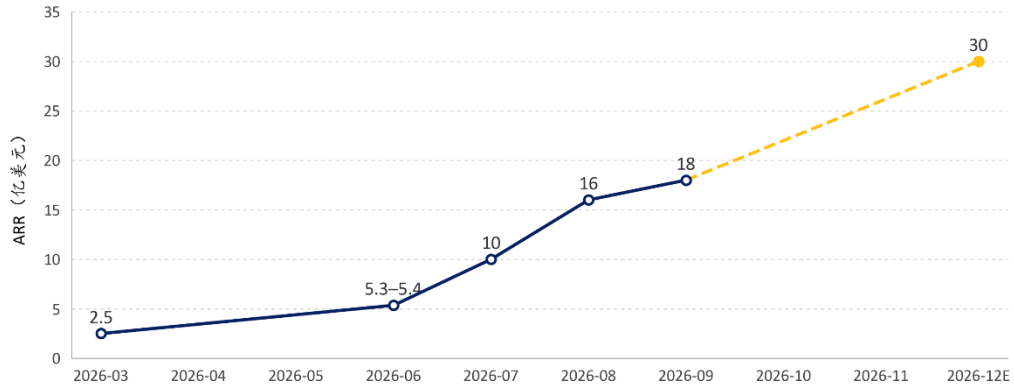
Co-work 把编程和智能体能力延伸至更多专业工作流程。据公司 1H26 业绩交流披露，公司服务需求以编程为主，约占七至八成，后续场景选择主要考虑任务的智力门槛、结果的可验证性以及经济价值。编程中积累的长程规划、工具调用和错误恢复能力，可用于需要多步骤执行、由专业人士复核结果的企业任务。网络安全是公司率先推进商业化的 Co-work 场景，10 月 12 日推出的 GLM-Legal 已开放 Beta 及 API 使用申请，将专业任务布局延伸至法律领域；金融及教育等方向仍在探索。

网络安全评测与实际漏洞发现体现了公司的专业任务能力。据公司发布 GLM-5.3 时的评测表，GLM-5.3 在 CyberGym 取得 84.5%的得分，高于同表列示的 Mythos 5 和 GPT-5.6 Sol。公司模型文档列示，相较 GLM-5.2，GLM-5.3 在 CyberGym 的得分由 77.2%升至 84.5%，在 ExploitBench 的得分由 24.4%升至 54.4%；在六小时预算下，GLM-5.3 在 ExploitGym 测试中完成 130 项任务，体现了漏洞分析与多步骤执行能力。据公司 1H26 中期披露，自 GLM-5.2 发布以来，公司联合多家安全团队在真实代码库中累计发现 2,436 个漏洞，覆盖 269 个项目，其中中高危漏洞 1,097 个。漏洞陆续提交至 CNNVD、CNVD 等国家级信息安全漏洞库，并通过安全披露账本记录处置进展，为安全厂商接入和产品集成提供了实际项目依据。

客户接入和订单已体现网络安全场景的商业需求。据管理层 9 月中旬交流，GLM-5.3 发布后一个月，Co-work 相关订单金额超过人民币 10 亿元，其中网络安全是率先落地的垂直场景，超过 100 家网络安全企业已将 GLM 接入安全产品与业务流程。我们认为，网络安全客户更看重漏洞发现及处置效果，这为公司按任务价值收费提供了空间。管理层预计相关订单将在一至两年内随交付确认收入。

编程业务与新业务进展推动管理层上调 ARR 指引。管理层在 9 月中旬交流中披露，9 月上旬 ARR 约 18 亿美元，并将 2026 年底指引由 24 亿美元上调至 30 亿美元，增幅 25%。管理层将上调归因于 Coding API 与 Coding Plan 增长、云服务商分成预计自 10 月起贡献，以及 Co-work 进展快于预期。从 18 亿美元到年底 30 亿美元，对应约三至四个月增长 67%。

图表 5：智谱 ARR 趋势



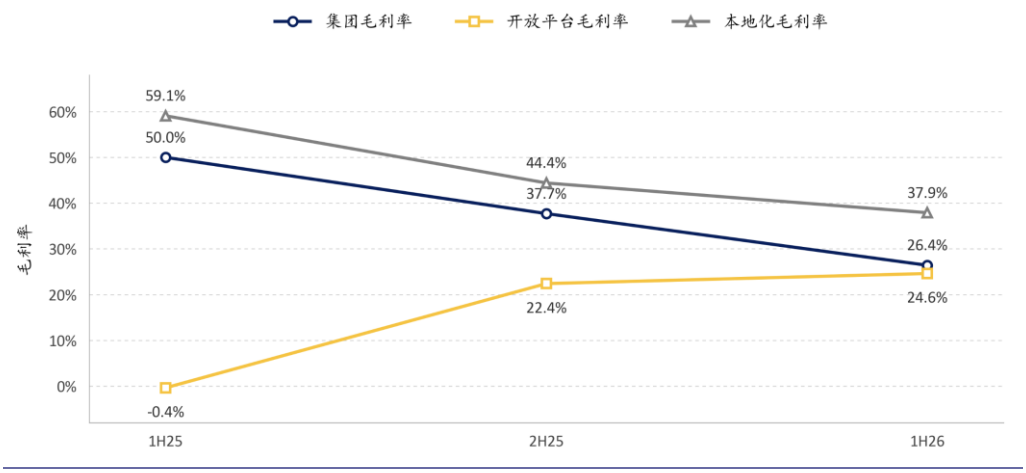
资料来源：公司资料、浦银国际

开放平台毛利改善，经营杠杆加速释放

收入放量正在扩大毛利贡献，亏损率明显下降。1H26 公司收入人民币 9.54 亿元，同比增长 399.7%，已超过 2025 全年的人民币 7.24 亿元；毛利人民币 2.52 亿元，同比增长 163.7%，毛利率 26.4%。期内亏损人民币 20.72 亿元，同比收窄 12.1%；经调整净亏损人民币 19.64 亿元，同比增加 12.1%，经调整净亏损率由 1H25 的 917.9%降至 205.9%。

集团毛利率受收入结构切换影响，开放平台毛利率持续改善。1H26，云端业务收入占比由 1H25 的 15.2%升至 86.5%，收入向毛利率较低的云端业务切换，同时本地化业务毛利率由 59.1%降至 37.9%，集团毛利率由 50.0%降至 26.4%。与此同时，开放平台毛利率由 1H25 的-0.4%升至 2H25 的 22.4%，并于 1H26 进一步升至 24.6%。

图表 6：智谱毛利率趋势



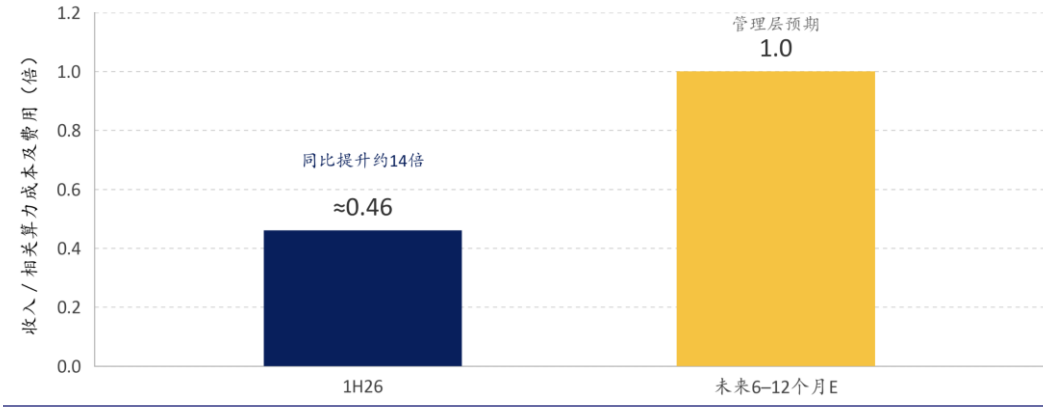
资料来源：公司资料、浦银国际

国产算力与推理工程优化带动单位推理成本下降。公司在 1H26 中期业绩披露中表示，单位 Token 推理成本较 2026 年初下降 80%。公司自年初实施 All-in-Infra 战略，已实现 10 万级国产芯片规模化推理，编码场景下的单 Token 推理性价比已与主流进口卡相当。GLM-5.3-Flash 是公司首次在超大规模真实流量中全面使用国产芯片集群提供服务。据公司披露，由 GLM-5.3 驱动的 Infra Agent 使算子开发周期缩短一半，端到端服务吞吐量达到同一硬件初始基线的约 3 倍。

规模化部署与模型分层有望推动毛利率进一步提升。据公司 1H26 业绩会，管理层提出未来 12 至 18 个月毛利率向 50%以上提升的方向性目标，主要驱动包括：1）国产大集群规模效应，2）自建超大规模算力中心与算力伙伴合作，3）模型至芯片层的全栈 Infra 优化，以及 4）模型分层和智能路由。根据任务难度选用旗舰或 Flash 模型，可减少简单任务调用高成本模型的支出。

收入对算力投入的覆盖程度持续提高。公司将算力乘数定义为开放平台及API收入与同期训练、推理相关算力成本及费用的比值。管理层在9月中旬交流中披露，1H26该指标约为0.46，即每人民币1元相关算力成本及费用对应约0.46元开放平台及API收入，该指标较去年同期提升约14倍。管理层预计该指标在未来6至12个月可达到1.0。

图表 7：智谱算力乘数



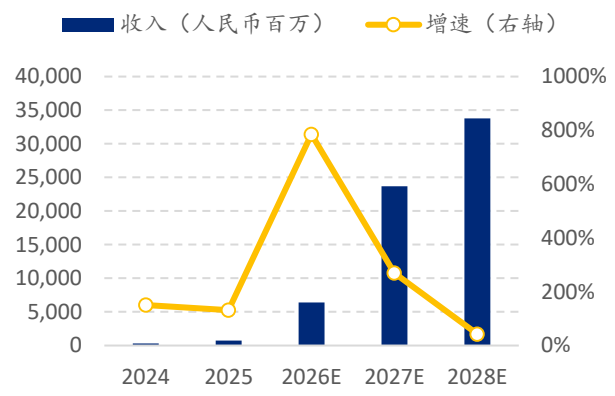
E=公司预测

资料来源：公司资料、浦银国际

收入增长与费用控制已开始体现经营杠杆。1H26研发开支人民币21.31亿元，同比增长33.6%、较2H25增长34.4%，研发费用率由1H25的835.4%降至223.4%。销售及营销开支人民币1.78亿元，同比下降14.8%，主要由于市场推广资源投放优化；一般及行政开支人民币1.03亿元，同比下降44.2%，主要由于专业服务费及以权益结算的股份支付减少。收入增速明显高于研发开支增速，销售及行政费用同时下降，带动费用率回落。

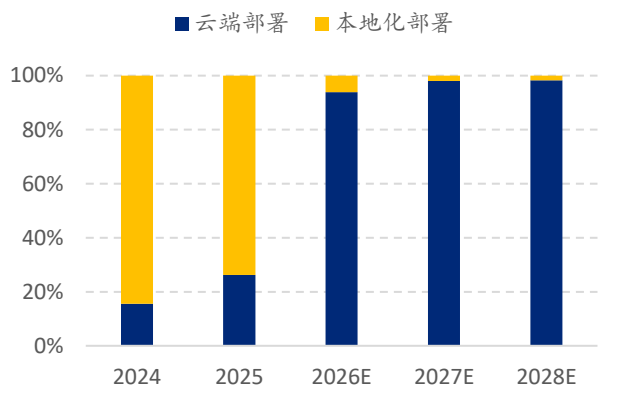
盈利预测

图表 8：智谱收入预测



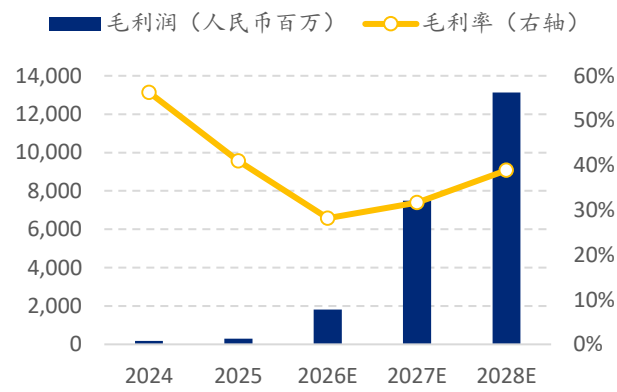
E=浦银国际预测，资料来源：公司财报、浦银国际

图表 9：智谱收入结构预测



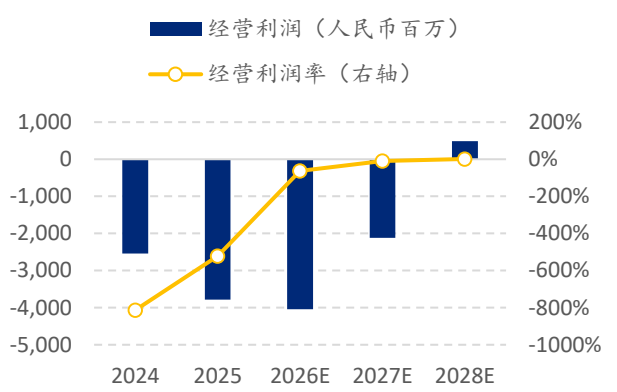
E=浦银国际预测，资料来源：公司财报、浦银国际

图表 10：智谱毛利润预测



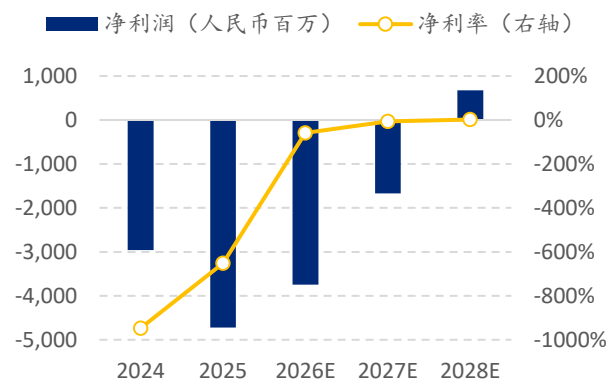
E=浦银国际预测，资料来源：公司财报、浦银国际

图表 11：智谱经营利润预测



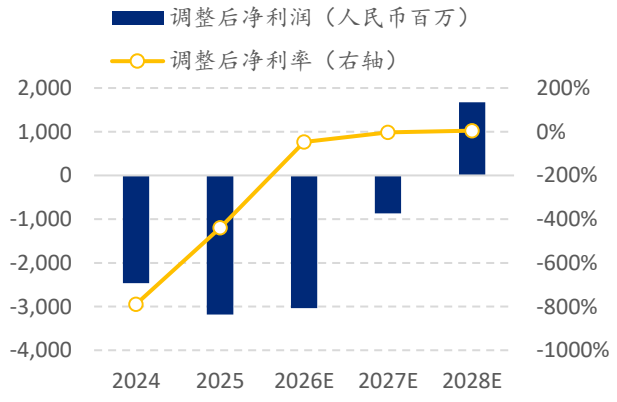
E=浦银国际预测，资料来源：公司财报、浦银国际

图表 12：智谱净利润预测



E=浦银国际预测，资料来源：公司财报、浦银国际

图表 13：智谱调整后净利润预测



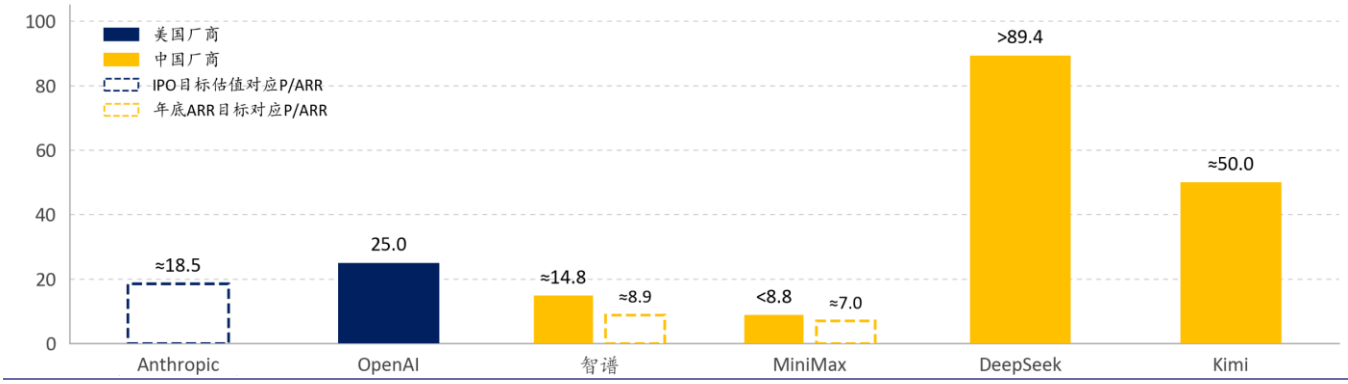
E=浦银国际预测，资料来源：公司财报、浦银国际

估值分析与预测

首予“买入”评级，目标价 750 港元

我们预计智谱 2026 年末 ARR 为 31 亿美元，略高于管理层 9 月提出的 30 亿美元目标。参考海外领先模型厂商估值，并结合公司业务增长及盈利预期，给予 15 倍 P/ARR，目标价 750 港元，对应 2027 年预测收入的 13.2 倍 P/S。我们看好公司在复杂编程与专业任务中的变现能力，以及云托管合作带来的新增收入，首次覆盖给予“买入”评级。

图表 14：中美主要独立模型厂商 P/ARR



注：市值截至 2026 年 11 月 6 日；资料来源：The Information、Bloomberg、Financial Times、Reuters、WSJ、Wind、公司资料、浦银国际

图表 15：可比公司估值比较

股票代码	公司	市值 (百万美元)	P/E (x) 2026E	2027E	2028E	P/S (x) 2026E	2027E	2028E
基础模型								
2513 HK Equity	智谱	26,507	N/A	N/A	N/A	27.5	9.0	4.3
100 HK Equity	MiniMax	7,005	N/A	N/A	N/A	14.9	5.1	2.4
平均			N/A	N/A	N/A	21.2	7.0	3.3
企业 AI 与模型服务								
20 HK Equity	商汤	6,252	N/A	N/A	76.0	6.5	5.1	4.1
6682 HK Equity	范式智能	1,481	80.6	28.2	12.9	0.9	0.7	0.5
002230 CH Equity	科大讯飞	13,768	87.9	62.1	44.8	3.0	2.5	2.2
PLTR US Equity	Palantir	386,364	94.8	63.7	42.6	45.8	29.7	20.2
平均			87.8	51.3	44.1	14.1	9.5	6.8
模态与创意应用								
1024 HK Equity	快手	15,863	12.7	11.6	9.7	0.8	0.7	0.7
1357 HK Equity	美图公司	2,461	13.2	10.5	8.3	3.4	2.8	2.4
ADBE US Equity	Adobe	93,136	9.7	8.6	7.6	3.5	3.2	3.0
平均			11.8	10.2	8.6	2.6	2.3	2.0
综合平台								
BABA US Equity	阿里巴巴	272,284	21.5	12.0	10.1	1.7	1.5	1.3
BIDU US Equity	百度	27,975	14.8	10.5	8.3	1.5	1.4	1.2
700 HK Equity	腾讯	486,893	12.4	11.9	10.9	3.9	3.6	3.3
GOOGL US Equity	谷歌	3,495,992	15.4	18.0	14.3	8.1	6.1	4.9
平均			16.0	13.1	10.9	3.8	3.1	2.7

注：E=Bloomberg 一致预测，股价截至 2026 年 11 月 6 日；资料来源：Bloomberg、浦银国际

图表 16：SPDBI 目标价：智谱（2513.HK）



注：截至 2026 年 11 月 6 日收盘

资料来源：Bloomberg、浦银国际

投资风险

- 模型及行业竞争加剧。若国内外厂商迭代速度加快，替代模型在编程与专业任务中表现接近，公司可能面临客户流失、价格竞争和研发投入增加的风险。

- 客户调用与 ARR 增长不及预期。客户扩容、订阅续费或头部用户使用放缓，可能影响业务增长；管理层 ARR 包含订单年化因素，相关订单交付及收入确认进度也会影响财报收入。

- 毛利率改善不及预期。新增集群利用率提升较慢、算力采购及租赁价格上涨、新旧模型切换成本及产品结构变化，可能抵消部分推理降本收益。

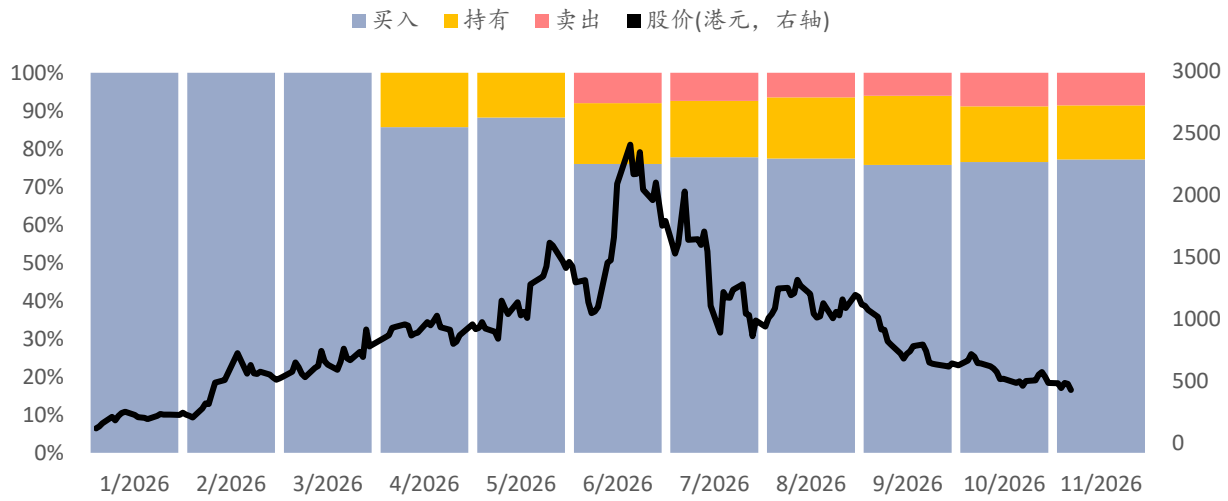
- Co-work 场景拓展及订单转化不及预期。专业任务的可靠性、企业产品集成和验收进度将影响采购与交付，网络安全订单转化或新增垂直场景落地慢于预期，可能影响新增收入贡献。

- 监管及合规风险。训练数据与知识产权、企业数据安全、模型安全和跨境业务相关监管要求，可能影响产品部署、客户采用及合作渠道。

- 现金消耗与债务偿付压力。研发和算力扩张需要持续资金投入；若收入及现金回收慢于投入进度，或 2027 年到期可转债未完成换股，公司可能面临偿债、再融资及新增股本摊薄压力。

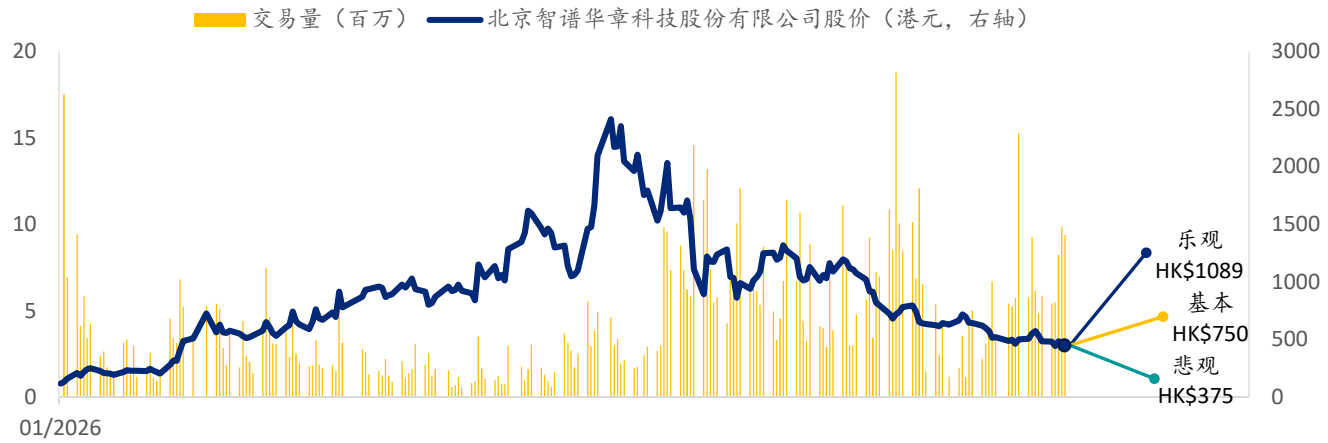
SPDBI 乐观与悲观情景假设

图表 17：市场普遍预期：智谱（2513.HK）



资料来源：Bloomberg、浦银国际

图表 18：SPDBI 情景假设：智谱（2513.HK）



乐观情景：调用与订单转化、毛利改善超预期（概率：20%）	悲观情景：调用、订单收入确认及毛利改善低于预期（概率：20%）
目标价：1,089 港元	目标价：375 港元
• 客户持续扩容，API 调用及订阅增长超预期； • 云平台分成与 Co-work 订单转化更快； • 新增集群利用率及单位推理成本改善超预期； • 新增毛利支撑研发投入，亏损收窄快于预期。	• 客户调用及订阅增长放缓； • 模型竞争加剧，净售价或毛利改善不及预期； • 专业任务订单交付延迟，现金消耗高于预期。

资料来源：Bloomberg、浦银国际预测

浦银国际研究

首次覆盖 | 互联网行业

浦銀國際 SPDB INTERNATIONAL

MiniMax（100.HK）：多模态驱动全球扩张

我们首次覆盖 MiniMax（100.HK），给予“买入”评级，目标价 245 港元。公司通过底层推理工程优化持续下探模型服务成本，凭借高性价比优势带动企业端调用需求释放。我们看好公司在编程与智能体场景下的企业级服务放量，以及视频、语音和音乐等多模态产品在全球商业内容制作市场的广阔变现空间。

- **模型升级与推理降本驱动企业端业务高速增长。**MiniMax 是面向全球市场提供基础模型及 AI 原生应用的领先企业。受益于付费客群扩容与 Token 订阅计划普及，1H26 公司开放平台及其他企业服务收入同比增长 703.1%至 7,393 万美元，占总收入比重达 63.4%。7 月平台 Token 调用量达今年 1 月的 20 倍，8 月集团 ARR 达 8 亿美元，其中企业端贡献超 80%，管理层维持年底 10 亿美元 ARR 目标。公司 10 月推出的 M3.1 Flash 将智能体任务使用成本大幅降低 70%，结合性价比优势与 MiniMax Code 等工具生态，有望进一步带动企业日常开发与高频智能体调用需求释放。
- **全模态矩阵拓宽商业创作场景，全球商业化进程加速。**公司拥有文本、视频、语音及音乐的多模态布局，其中 H3 系列视频生成能力稳居全球盲测榜单前列，能够满足广告营销、电商展示及影视娱乐的高频素材制作需求。公司依托开放平台 API 及自有 AI 原生应用，广泛服务全球企业、开发者与个人用户。1H26 公司业务已覆盖全球逾 230 个国家和地区，海外收入同比增长 224.0%至 7,083 万美元，占总收入比重达 60.8%，全球商业化进程加速。
- **推理效率与服务优化有望驱动下半年毛利率环比修复。**1H26 公司实现毛利 2,081 万美元，毛利率同比提升 5.8 个百分点至 17.9%。受 M3 上线初期用户 Token 补偿、推理服务磨合及收入结构变化影响，1H26 毛利率较 2H25 出现阶段性环比回落。随着一次性补偿因素消退、Token 套餐权益优化及底层推理效率提升，管理层预计下半年毛利率将迎来环比改善。收入规模扩张亦将有效摊薄研发及运营费率，推动经调整净亏损率逐步收窄。
- **首予“买入”评级，目标价 245 港元。**我们预计 2026 年末集团 ARR 为 11 亿美元，给予 2026E 10x P/ARR 估值，目标价 245 港元，对应 2027E 7.9x P/S。短期催化剂包括 M3 Pro 旗舰及多模态新品发布、年底 ARR 超预期、下半年毛利率环比回升及 A 股上市进展。短期主要关注新品商业化转化、研发与算力投入节奏，以及 2027 年 1 月部分股东禁售期届满带来的潜在流通盘增加。
- **投资风险：**模型迭代及商业化不及预期；行业竞争加剧；毛利率改善不及预期；海外监管及数据合规风险；研发投入与债务偿付压力。

图表 1：盈利预测和财务指标

美元百万	2024	2025	2026E	2027E	2028E
营业收入	31	79	484	1,390	2,116
经营利润	(287)	(321)	(843)	(941)	(817)
经调整净利润	(244)	(251)	(759)	(824)	(681)
目标 P/S			22.7	7.9	5.2

E=浦银国际预测；资料来源：公司公告、浦银国际

杨子超，CFA
科技分析师
charles_yang@spdbi.com
(852) 2808 6409

赵丹
首席科技软件分析师
dan_zhao@spdbi.com
(852) 2808 6436

2026 年 11 月 11 日

MiniMax（100.HK） 买入

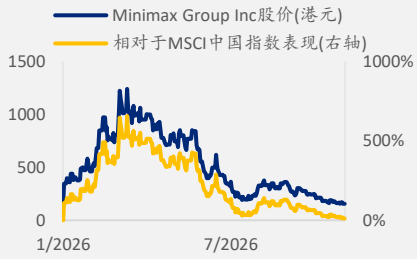
目标价（港元）	**245**
潜在升幅/降幅	**+57%**
目前股价（港元）	155.8
52 周内股价区间（港元）	148.4-1330
总市值（百万港元）	54,916
近 3 月日均成交额（百万港元）	3,484

注：截至 2026 年 11 月 6 日收盘价

市场预期区间



股价表现



注：截至 2026 年 11 月 6 日收盘价
资料来源：Bloomberg、浦银国际

浦银国际 | 首次覆盖 | MiniMax（100.HK）

本研究报告由浦银国际证券有限公司分析师编制，请仔细阅读本报告最后部分的分析师披露、商业关系披露及免责声明。

财务报表分析与预测

利润表

百万美元	2024	2025	2026E	2027E	2028E
收入	31	79	484	1,390	2,116
销售成本	(27)	(59)	(368)	(927)	(1,304)
毛利	4	20	116	463	812
销售及分销开支	(87)	(52)	(75)	(125)	(169)
行政开支	(14)	(37)	(67)	(97)	(127)
研发开支	(189)	(253)	(817)	(1,182)	(1,333)
经营亏损	(287)	(321)	(843)	(941)	(817)
其他收入及收益净额	36	40	31	43	18
金融负债公允价值亏损	(214)	(1,590)	(31)	-	-
财务开支	(1)	(1)	(19)	(21)	(6)
金融资产减值亏损	(0)	(0)	(2)	(4)	(6)
税前亏损	(465)	(1,872)	(864)	(924)	(811)
所得税开支	-	-	-	-	-
期内亏损	(465)	(1,872)	(864)	(924)	(811)
经调整净亏损	(244)	(251)	(759)	(824)	(681)

现金流量表

百万美元	2024	2025	2026E	2027E	2028E
税前亏损	(465)	(1,872)	(864)	(924)	(811)
折旧与摊销	2	3	17	46	71
股份薪酬	7	24	68	100	130
金融负债公允价值亏损	214	1,590	31	-	-
营运资金变动	14	6	66	58	48
其他非现金调整	(30)	(31)	20	26	12
经营活动现金流	(258)	(280)	(662)	(695)	(550)
资本开支	(1)	(1)	(238)	(100)	(100)
投资及存款净变动	(431)	73	161	100	75
投资活动现金流	(431)	72	(77)	-	(25)
借款及可转债净变动	19	16	926	(860)	-
股权融资	753	412	1,926	-	-
租赁付款	(2)	(2)	(2)	(2)	(2)
其他融资活动	(0)	(2)	(19)	(5)	(5)
融资活动现金流	771	423	2,831	(867)	(8)
汇兑影响	1	3	5	-	-
现金净变动	83	219	2,097	(1,562)	(583)
期初现金	206	289	508	2,605	1,043
期末现金	289	508	2,605	1,043	460

资产负债表

百万美元	2024	2025	2026E	2027E	2028E
物业、厂房及设备	1	2	112	169	200
使用权资产	3	2	5	6	7
其他非流动资产	101	77	207	207	212
非流动资产合计	105	81	325	382	419
现金及现金等价物	289	508	2,605	1,043	460
贸易应收款项	7	11	68	181	254
其他流动资产	510	489	466	502	485
流动资产合计	806	1,007	3,139	1,726	1,200
资产总额	911	1,088	3,463	2,108	1,618
计息银行借款	19	35	134	134	134
可转债及其他债务	-	-	844	-	-
贸易应付款项及应付票据	51	58	296	527	659
其他负债	56	38	65	98	128
合同负债	2	8	29	76	106
可转换可赎回优先股	1,582	3,598	-	-	-
负债总额	1,710	3,737	1,367	835	1,027
权益总额	(799)	(2,648)	2,096	1,273	592
权益及负债总额	911	1,088	3,463	2,108	1,618

主要财务比率

百万美元	2024	2025	2026E	2027E	2028E
盈利增速					
营业收入增速	782.2%	158.9%	512.3%	187.2%	52.2%
毛利增速	-537.7%	437.2%	478.8%	298.4%	75.4%
经营利润增速	nm	nm	nm	nm	nm
净利润增速	nm	nm	nm	nm	nm
经调整净利润增速	nm	nm	nm	nm	nm
盈利能力比率					
毛利率	12.2%	25.4%	24.0%	33.3%	38.4%
经营利润率	-939.0%	-406.6%	-174.1%	-67.7%	-38.6%
净利率	-1524.2%	-2368.0%	-178.5%	-66.4%	-38.3%
经调整净利率	-800.2%	-317.4%	-156.8%	-59.3%	-32.2%
每股指标（美元）					
EPS			(2.64)	(2.64)	(2.32)
经调整EPS			(2.32)	(2.36)	(1.95)
估值（倍）					
目标P/S			22.7x	7.9x	5.2x
经调整目标P/E			nm	nm	nm

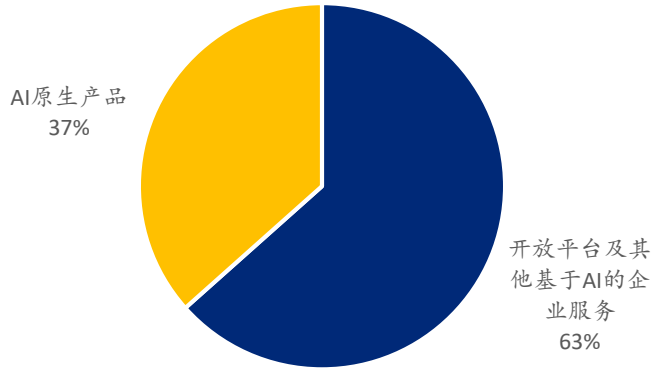
资料来源：公司公告、Bloomberg、浦银国际预测

能力提升与成本优化带动企业端增长

开放平台成为收入主体，企业客户贡献主要增量。公司 1H26 开放平台及其他基于 AI 的企业服务收入为 7,393 万美元，同比增长 703.1%，占总收入的 63.4%；AI 原生产品收入 4,264 万美元，同比增长 100.9%，占比 36.6%。开放平台增长主要来自付费用户增加、API 调用扩大及 Token 计划的采用。据业绩会披露，今年 7 月 Token 调用量达到 1 月的 20 倍。公司持续改进编程与智能体任务表现，并通过推理工程优化降低服务成本，10 月发布的 M3.1 Flash 进一步降低了智能体任务的使用成本。我们认为，模型能够完成的工作增加、客户使用成本下降，有助于企业扩大调用，推动业务增长。

AI 原生产品保持增长，与开放平台形成互补。公司表示，C 端业务已实现正毛利，并转向用户自然增长策略。自有应用直接满足内容创作等终端需求，开放平台则供开发者和企业将模型集成至自身产品，分别通过应用付费和模型调用变现。

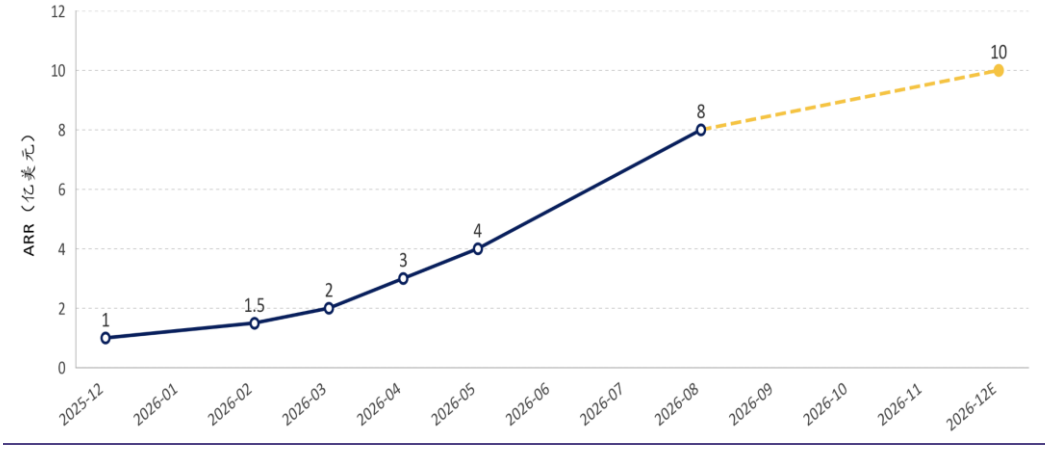
图表 2：MiniMax 收入结构（1H26）



资料来源：公司资料、浦银国际

企业端也已成为 ARR 的主要来源。据公司今年 8 月披露，按周度收入年化计算，公司 8 月 ARR 约 8 亿美元；其中企业端占比超过 80%，去年同期约为 30%；公司维持 2026 年底 10 亿美元的预期，并将 M3.1、M3 Pro 及 H3.1 列为后续增长的重要产品。

图表 3：MiniMax ARR 趋势



E=公司目标

资料来源：公司资料、浦银国际

文本与多模态并进，拓展全球商业化机会

文本与多模态产品使公司能够覆盖编程、智能体及内容制作等多类需求。公司持续提升文本模型能力、降低调用成本，并通过视频、语音和音乐产品拓展商业内容制作场景。

M3.1 Flash 降低专业任务使用成本，支持企业客户扩大调用。2026 年 10 月 19 日，公司正式推出 M3.1 Flash，通过 API、M Plan 及 MiniMax Code 提供服务。据公司披露，M3.1 Flash 的智能体任务使用成本较前代下降 70%；API 每百万 Token 输入、输出价格分别为 0.15 美元和 0.60 美元。我们认为，任务使用成本下降，有助于企业将模型用于更高频的代码修改、调试和测试，带动调用量增长。

Pro 研发规划面向更复杂的任务需求。管理层在 8 月业绩会提出，M3 Pro 将进一步扩大模型规模，总参数预计约 3 万亿，为 M3 的 6 至 7 倍，采用 MoE 2.0，并计划作为公司开放权重的最高版本；其预训练已进入后期，单 Token 成本增幅预计远小于参数增幅，定价和毛利率预计高于 M3。公司在 Flash 发布说明中表示 Pro 将于后续推出。Flash 侧重广泛使用，Pro 面向更复杂的任务，产品分层有助于覆盖不同需求和预算。

现有算力与新增集群支持后续模型研发，按管理层 8 月业绩会当时披露的计划，算力储备可覆盖 M3 Pro 预训练、M3.1 及 H3 更新等训练需求，新增国内与海外自建集群计划用于 5 至 10 万亿参数下一代模型的训练储备。推理端通过采购外部 Token 资源补充服务能力，管理层预计 2026 年第四季度国产芯片占比将提升。

图表 4：MiniMax 模型与产品路线

产品矩阵	发布时间	参数	定位
M3	2026 年 6 月 1 日	总参数约 428B、激活约 23B；最高 1M 上下文，原生多模态	强化编程、Agentic 工作流与专业工作
H3	2026 年 7 月 31 日	核心模型约 33B 参数，支持音视频联合生成；开放 Base 权重	商业创作视频生成
M3.1 Flash	2026 年 10 月 19 日	428B 总参数、23B 激活；全模型稀疏注意力、量化感知训练	编程、智能体、Computer Use 与网络安全
M3 Pro	2026 年 8 月披露研发计划；10 月 19 日公司称后续推出	预计约 3 万亿参数、MoE 2.0；预训练当时已进入后期	开放权重旗舰版本，提升复杂任务能力

资料来源：公司资料、浦银国际

H3 支持多种素材输入与音画联合生成。除 M 系列文本模型外，公司还布局了 H3、Speech 等产品，其中 H3 于今年 7 月 31 日推出。据公司披露，H3 可综合理解文本、图片、视频与音频输入，支持生成最长 15 秒、带原生立体声音频的视频，通过完整工作流程输出 2K 画面，并支持参考素材生成、编辑及动作迁移。客户可结合已有画面、声音和文字要求完成创作，更好地控制人物、场景及视听效果。管理层在 8 月业绩会上表示，文本与多模态能力相互促进，并预计 2026 年第四季度至 2027 年第一季度将出现更多文本与视频生成融合的机会。

图表 5：MiniMax 多模态产品矩阵



资料来源：公司资料，浦银国际

图表 6：MiniMax 视频生成能力处于行业领先水平

厂商	前沿模型	Elo 等级分	发布时间	API 价格 (USD/每分钟)
文生视频（带音频）				
阿里巴巴	Wan 3.0	1172	2026 年 8 月	$12.00
Utopai Studios	Utopai X (Based on **MiniMax H3**)	1169	2026 年 9 月	未提供 API
MiniMax	**MiniMax H3 (768p)**	**1160**	**2026 年 7 月**	**$4.80**
fal	FLUX 3 Max	1156	2026 年 10 月	$4.35
字节跳动	Dreamina Seedance 2.5	1155	2026 年 7 月	$34.12
图生视频（带音频）				
字节跳动	Dreamina Seedance 2.5	1159	2026 年 7 月	$34.12
Utopai Studios	Utopai X (Based on **MiniMax H3**)	1147	2026 年 9 月	未提供 API
字节跳动	Dreamina Seedance 2.0	1135	2026 年 3 月	$22.45
MiniMax	**MiniMax H3 (768p)**	**1131**	**2026 年 7 月**	**$4.80**
HeyGen	HeyGen Video 1 (Based on **MiniMax H3**)	1124	2026 年 10 月	$2.40

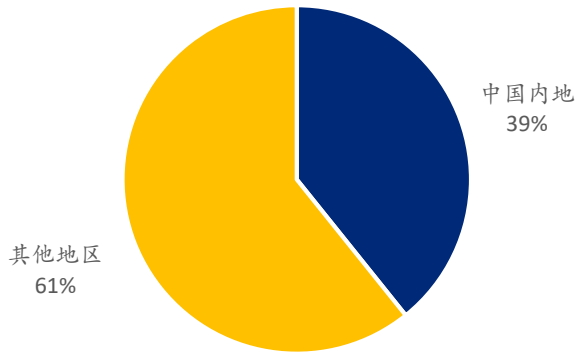
注：数据截至 2026 年 11 月 6 日；

资料来源：Artificial Analysis、浦银国际

多模态产品可服务广告、电商、娱乐等内容制作需求。广告、电商和品牌营销通常需要围绕同一商品或创意，制作不同画面、配音和版本，客户既关注生成效果，也重视素材一致性、修改效率与制作成本。我们认为，视频、配音等能力结合，可减少工具切换与人工后期工作；此外，营销素材需要持续更新，也为模型调用及创作工具订阅带来重复付费需求。

海外收入保持快速增长，全球业务覆盖模型服务与终端应用。1H26 公司通过开放平台和 AI 原生应用覆盖超过 230 个国家和地区，同期海外收入 7,083 万美元，同比增长 224.0%，占总收入的 60.8%。公司在 1H26 中期业绩中表示，将继续加强开放平台，拓展 AI 原生产品及工具，并降低模型接入、整合和部署的门槛，以服务更广泛的企业、开发者及个人用户。

图表 7：MiniMax 收入地区分布（1H26）

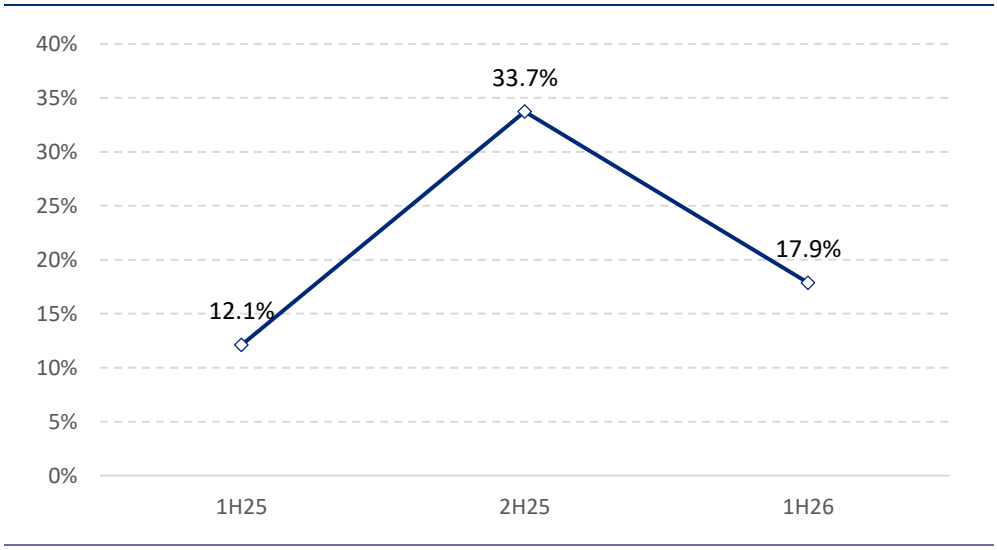


资料来源：公司资料、浦银国际

推理与服务优化带动毛利修复

收入与毛利额快速增长，毛利率同比提升。1H26 公司收入 1.166 亿美元，同比增长 283.1%；毛利 2,081 万美元，同比增长 464.8%，毛利率 17.9%，高于 1H25 的 12.1%，较 2H25 约 33.7%环比下降约 16 个百分点，主要因为：1）M3 上线初期出现服务体验及定价权益问题，公司向用户发放额外 Token 补偿；2）推理服务仍在优化，输出速度由上线时约 20 Token/s 提升至一个月后接近 100 Token/s；3）企业端 Token 订阅占比上升，其毛利率低于 API；4）文本模型收入占比提升，其毛利率低于多模态产品。据 1H26 业绩会，管理层称 M3 的 API 毛利率已高于 M2，API 稳态毛利率可达 40%至 50%。管理层预计，一次性补偿影响退出、套餐权益优化及推理效率提升，将推动下半年毛利率环比改善。

图表 8：MiniMax 毛利率趋势



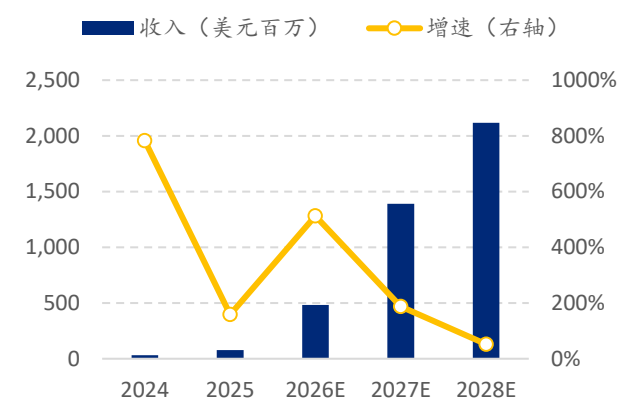
资料来源：公司资料、浦银国际

训练效率提升，M3 单 Token 成本低于 M2。今年 8 月公司在业绩会上表示，训练基础设施的有效训练时间比率（ETTR）达 97%，最大参数规模下计算效率较第一代架构提升至约 3 倍，单位算力吞吐亦有约 3 倍提升。公司还表示，M3（总参数约 428B）的单 Token 成本已低于 M2（230B）。

主要费率下降，经营杠杆逐步体现。1H26 销售及分销开支同比下降 17.9%至 2,697 万美元，主要因获客策略转向用户自然增长；行政开支同比增长 103.7%至 3,023 万美元，占收入比重由 48.8%降至 25.9%。同期研发开支 2.969 亿美元，同比增长 138.8%，约为当期收入的 2.5 倍，增速低于收入增长。

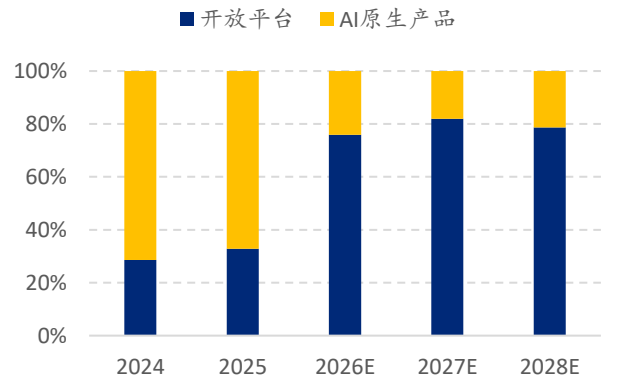
盈利预测

图表 9：MiniMax 收入预测



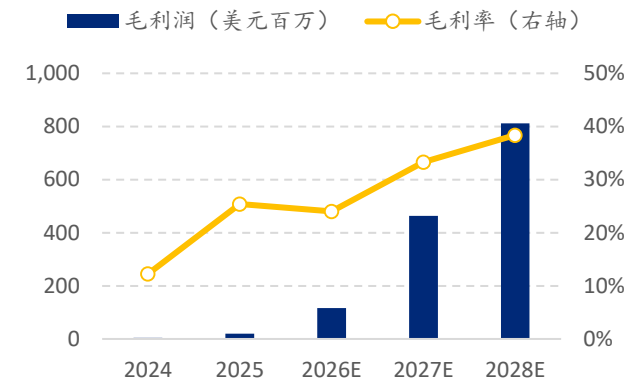
E=浦银国际预测，资料来源：公司财报、浦银国际

图表 10：MiniMax 收入结构预测



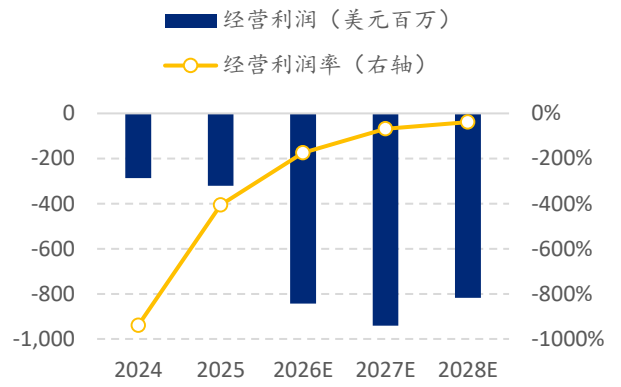
E=浦银国际预测，资料来源：公司财报、浦银国际

图表 11：MiniMax 毛利润预测



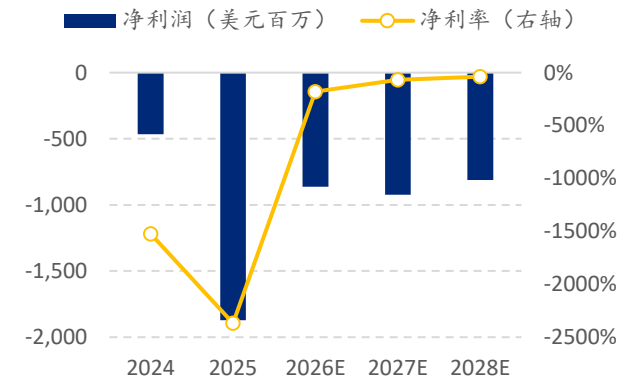
E=浦银国际预测，资料来源：公司财报、浦银国际

图表 12：MiniMax 经营净利润预测



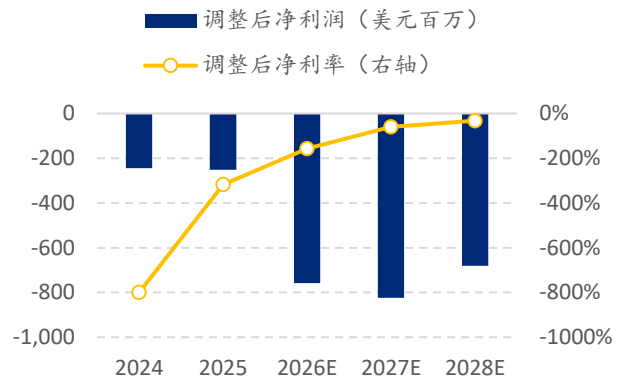
E=浦银国际预测，资料来源：公司财报、浦银国际

图表 13：MiniMax 净利润预测



E=浦银国际预测，资料来源：公司财报、浦银国际

图表 14：MiniMax 调整后净利润预测



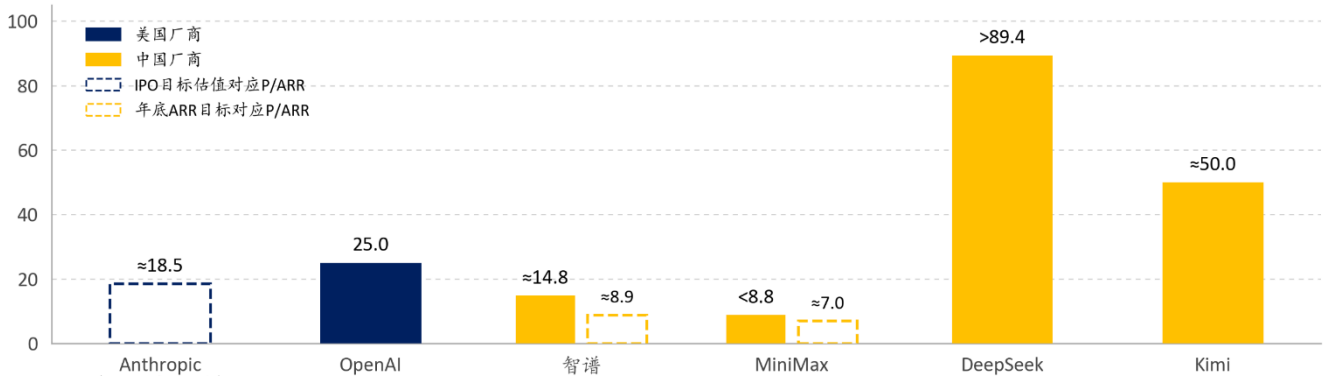
E=浦银国际预测，资料来源：公司财报、浦银国际

估值分析与预测

首予“买入”评级，目标价 245 港元

我们预计 MiniMax 2026 年末集团 ARR 为 11 亿美元，略高于管理层提出的 10 亿美元目标。参考海外领先模型厂商估值，并结合公司业务增长及盈利预期，给予 10 倍 2026E P/ARR，目标价 245 港元，对应 2027 年预测收入的 7.9 倍 P/S。我们看好企业端模型服务与多模态内容制作需求的增长，首次覆盖给予“买入”评级。

图表 15：中美主要独立模型厂商 P/ARR



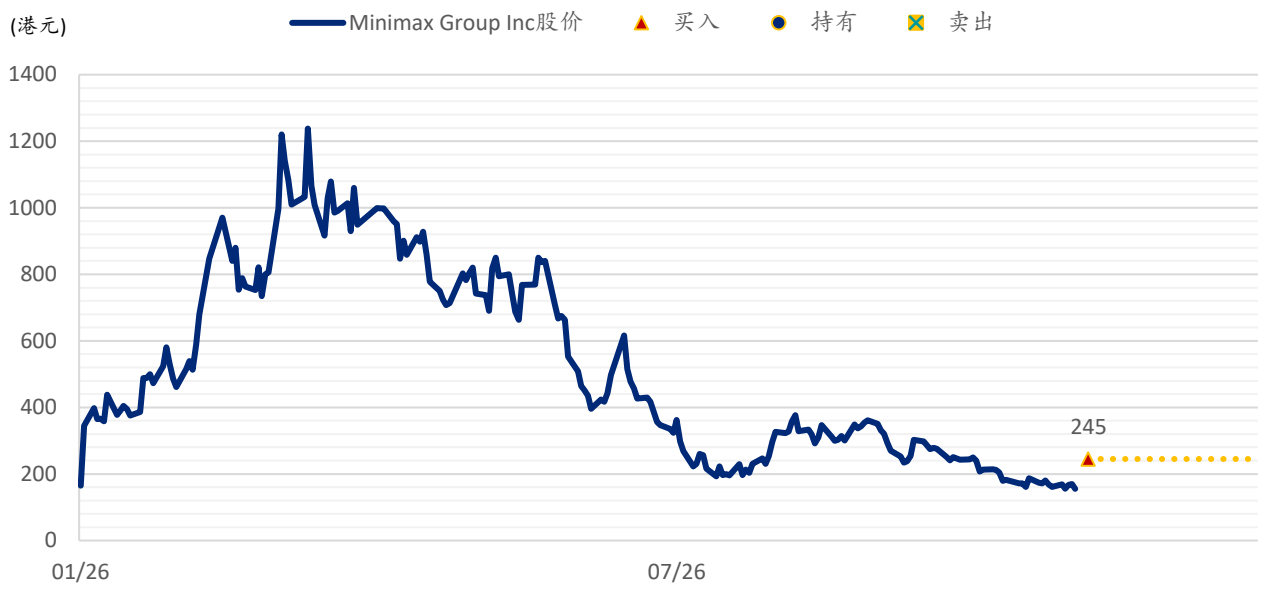
注：市值截至 2026 年 11 月 6 日；资料来源：The Information、Bloomberg、Financial Times、Reuters、WSJ、Wind、公司资料、浦银国际

图表 16：可比公司估值比较

股票代码	公司	市值 (百万美元)	P/E (x) 2026E	2027E	2028E	P/S (x) 2026E	2027E	2028E
基础模型								
2513 HK Equity	智谱	26,507	N/A	N/A	N/A	27.5	9.0	4.3
100 HK Equity	MiniMax	7,005	N/A	N/A	N/A	14.9	5.1	2.4
平均			N/A	N/A	N/A	21.2	7.0	3.3
企业 AI 与模型服务								
20 HK Equity	商汤	6,252	N/A	N/A	76.0	6.5	5.1	4.1
6682 HK Equity	范式智能	1,481	80.6	28.2	12.9	0.9	0.7	0.5
002230 CH Equity	科大讯飞	13,768	87.9	62.1	44.8	3.0	2.5	2.2
PLTR US Equity	Palantir	386,364	94.8	63.7	42.6	45.8	29.7	20.2
平均			87.8	51.3	44.1	14.1	9.5	6.8
模态与创意应用								
1024 HK Equity	快手	15,863	12.7	11.6	9.7	0.8	0.7	0.7
1357 HK Equity	美图公司	2,461	13.2	10.5	8.3	3.4	2.8	2.4
ADBE US Equity	Adobe	93,136	9.7	8.6	7.6	3.5	3.2	3.0
平均			11.8	10.2	8.6	2.6	2.3	2.0
综合平台								
BABA US Equity	阿里巴巴	272,284	21.5	12.0	10.1	1.7	1.5	1.3
BIDU US Equity	百度	27,975	14.8	10.5	8.3	1.5	1.4	1.2
700 HK Equity	腾讯	486,893	12.4	11.9	10.9	3.9	3.6	3.3
GOOGL US Equity	谷歌	3,495,992	15.4	18.0	14.3	8.1	6.1	4.9
平均			16.0	13.1	10.9	3.8	3.1	2.7

注：E=Bloomberg 一致预测，股价截至 2026 年 11 月 6 日；资料来源：Bloomberg、浦银国际

图表 17：SPDBI 目标价：MiniMax（100.HK）



注：截至 2026 年 11 月 6 日收盘

资料来源：Bloomberg、浦银国际

投资风险

- **行业竞争加剧。**国内外模型厂商持续推出新产品，编程、视频生成及语音等领域的竞争可能进一步加剧。若竞争对手在任务效果、调用价格或分发渠道方面取得优势，公司可能面临客户转用其他模型、实际售价下降及获客投入增加，影响开放平台与 AI 原生产品的收入增长。

- **模型迭代及服务表现不及预期。**新模型训练、后训练与推理部署需要持续研发和工程投入。若模型发布延迟，或在复杂任务完成率、生成质量及服务稳定性方面未达到客户要求，可能影响新品采用与付费转化。算力供给不足、集群部署或芯片适配进度落后，也可能限制模型迭代和新增调用的承接能力。

- **客户使用及 ARR 增长不及预期。**开放平台增长依赖付费客户增加、调用量扩大及 Token 计划续费。若企业客户试用后未持续使用，或编程与智能体任务的使用频率、付费意愿低于预期，收入增速可能放缓；自有应用的用户留存和付费转化不足，也会影响增长。

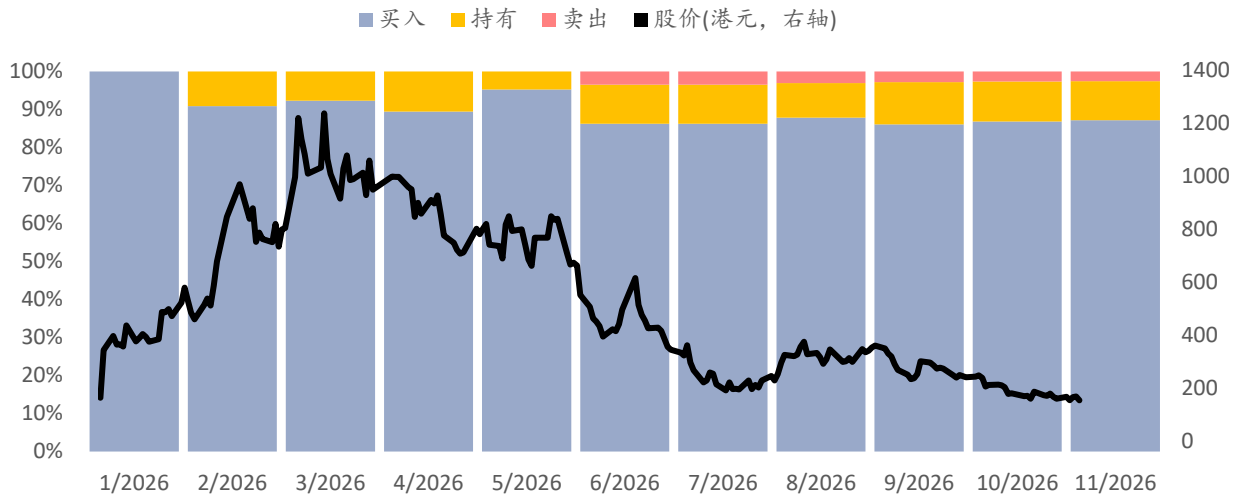
- **毛利率改善不及预期。**推理效率提升向实际毛利的传导，还受到集群利用率、算力采购及租赁价格、产品售价和业务结构影响。若新增集群利用率提升较慢，或较低毛利的 Token 计划及文本服务占比继续上升，降本收益可能被部分抵消。新模型上线期间的服务投入或用户补偿，也可能影响阶段性毛利，减缓收入增长对亏损收窄的推动作用。

- **监管及合规风险。**公司业务覆盖多个海外市场，模型训练数据、生成内容、知识产权、个人信息及跨境数据处理等要求可能影响产品开发和服务提供。中报披露，截至 2026 年 6 月 30 日，部分子公司涉及的知识产权争议诉讼仍在进行，结果及潜在赔偿难以估计，未计提相关拨备。相关规则变化、争议处理或内容审核不足，可能增加合规和诉讼支出，并导致产品调整、服务受限或客户信任受损，影响多模态产品及海外业务拓展。

- **现金消耗及债务偿付压力。**前沿模型研发和算力建设需要持续投入，相关支出通常先于收入及现金回收。若商业化增长或毛利改善慢于投入，公司可能面临更高的现金消耗及再融资需求。可转债若到期未完成转换，将增加现金偿付压力；若通过换股或增发补充资金，则可能摊薄现有股东权益。研发投入回报不足或融资条件转弱，也可能影响后续模型和算力扩张。

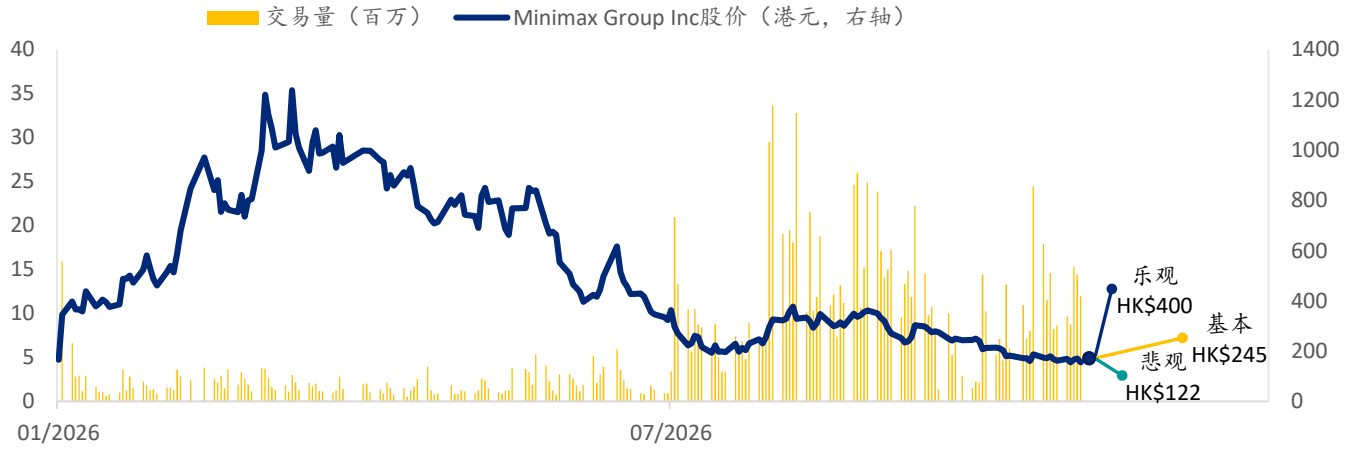
SPDBI 乐观与悲观情景假设

图表 18：市场普遍预期：MiniMax（100.HK）



资料来源：Bloomberg、浦银国际

图表 19：SPDBI 情景假设：MiniMax（100.HK）



乐观情景：模型采用及商业化好于预期 概率：20%		悲观情景：商业化及毛利改善低于预期 概率：20%
目标价：400 港元		目标价：122 港元
• 企业客户及调用量增长较快； • 多模态产品与海外付费需求扩大； • 推理效率提升，毛利贡献改善。		• 客户采用或调用增长较慢； • 价格竞争及成本压力延缓亏损收窄。

资料来源：Bloomberg、浦银国际预测

免责声明

本报告之收取者透过接受本报告（包括任何有关的附件），表示及保证其根据下述的条件下有权获得本报告，且同意受此中包含的限制条件所约束。任何没有遵循这些限制的情况可能构成法律之违反。

本报告是由从事证券及期货条例(香港法例第 571 章)中第一类(证券交易)及第四类(就证券提供意见)受规管活动之持牌法团－浦银国际证券有限公司 （统称“浦银国际证券”）利用集团信息及其他公开信息编制而成。所有资料均搜集自被认为是可靠的来源，但并不保证数据之准确性、可信性及完整性，亦不会因资料引致的任何损失承担任何责任。报告中的资料来源除非另有说明，否则信息均来自本集团。本报告的内容涉及到保密数据，所以仅供阁下为其自身利益而使用。除了阁下以及受聘向阁下提供咨询意见的人士（其同意将本材料保密并受本免责声明中所述限制约束）之外，本报告分发给任何人均属未授权的行为。

任何人不得将本报告内任何信息用于其他目的。本报告仅是为提供信息而准备的，不得被解释为是一项关于购买或者出售任何证券或相关金融工具的要约邀请或者要约。阁下不应将本报告内容解释为法律、税务、会计或投资事项的专业意见或为任何推荐，阁下应当就本报告所述的任何交易涉及的法律及相关事项咨询其自己的法律顾问和财务顾问的意见。本报告内的信息及意见乃于文件注明日期作出，日后可作修改而不另通知，亦不一定会更新以反映文件日期之后发生的进展。本报告并未包含公司可能要求的所有信息，阁下不应仅仅依据本报告中的信息而作出投资、撤资或其他财务方面的任何决策或行动。除关于历史数据的陈述外，本报告可能包含前瞻性的陈述，牵涉多种风险和不确定性，该等前瞻性陈述可基于一些假设，受限于重大风险和不确定性。

本报告之观点、推荐、建议和意见均不一定反映浦银国际证券的立场。浦银国际控股有限公司及其联属公司、关联公司（统称“浦银国际”）及/或其董事及/或雇员，可能持有在本报告内所述或有关公司之证券、并可能不时进行买卖。浦银国际或其任何董事及/或雇员对投资者因使用本报告或依赖其所载信息而引起的一切可能损失，概不承担任何法律责任。

浦银国际证券建议投资者应独立地评估本报告内的资料，考虑其本身的特定投资目标、财务状况及需要，在参与有关报告中所述公司之证券的交易前，委任其认为必须的法律、商业、财务、税务或其它方面的专业顾问。惟报告内所述的公司之证券未必能在所有司法管辖区或国家或供所有类别的投资者买卖。对部分的司法管辖区或国家而言，分发、发行或使用本报告会抵触当地法律、法则、规定、或其它注册或发牌的规例。本报告不是旨在向该等司法管辖区或国家的任何人或实体分发或由其使用。

美国

浦银国际不是美国注册经纪商和美国金融业监管局(FINRA)的注册会员。浦银国际证券的分析师不具有美国金融监管局(FINRA)分析师的注册资格。因此，浦银国际证券不受美国就有关研究报告准备和分析师独立性规则的约束。

本报告仅提供给美国 1934 年证券交易法规则 15a-6 定义的“主要机构投资者”，不得提供给其他任何个人。接收本报告之行为即表明同意接受协议不得将本报告分发或提供给任何其他人。接收本报告的美国收件人如想根据本报告中提供的信息进行任何买卖证券交易，都应仅通过美国注册的经纪交易商来进行交易。

英国

本报告并非由英国 2000 年金融服务与市场法(经修订)(「FSMA」)第 21 条所界定之认可人士发布,而本报告亦未经其批准。因此，本报告不会向英国公众人士派发，亦不得向公众人士传递。本报告仅提供给合资格投资者(按照金融服务及市场法的涵义），即(i)按照 2000 年金融服务及市场法 2005 年(金融推广)命令(「命令」)第 19(5)条定义在投资方面拥有专业经验之投资专业人士或(ii)属于命令第 49(2)(a)至(d)条范围之高净值实体或(iii)其他可能合法与之沟通的人士(所有该等人士统称为「有关人士」)。不属于有关人士的任何机构和个人不得遵照或倚赖本报告或其任何内容行事。

本报告的版权仅为浦银国际证券所有，未经书面许可任何机构和个人不得以任何形式转发、翻版、复制、刊登、发表或引用，浦银国际证券对任何第三方的该等行为保留追述权利，并且对第三方未经授权行为不承担任何责任。

权益披露

1) 浦银国际并没有持有本报告所述公司逾 1%的财务权益。
2) 浦银国际跟本报告所述公司（快手科技 1024.HK）在过去 12 个月内有投资银行业务的关系。
3) 浦银国际并没有跟本报告所述公司为其证券进行庄家活动。

评级定义

证券评级定义：

“买入”：未来 12 个月，预期个股表现超过同期其所属的行业指数
“持有”：未来 12 个月，预期个股表现与同期所属的行业指数持平
“卖出”：未来 12 个月，预期个股表现逊于同期其所属的行业指数

行业评级定义（相对于 MSCI 中国指数）：

“超配”：未来 12 个月优于 MSCI 中国 10%或以上
“标配”：未来 12 个月优于/劣于 MSCI 中国少于 10%
“低配”：未来 12 个月劣于 MSCI 中国超过 10%

分析师证明

本报告作者谨此声明：(i) 本报告发表的所有观点均正确地反映作者有关任何及所有提及的证券或发行人的个人观点，并以独立方式撰写；(ii) 其报酬没有任何部分曾经，是或将会直接或间接与本报告发表的特定建议或观点有关；(iii) 该等作者没有获得与所提及的证券或发行人相关且可能影响该等建议的内幕信息 / 非公开的价格敏感数据。

本报告作者进一步确定 (i) 他们或其各自的关联人士（定义见证券及期货事务监察委员会持牌人或注册人操守准则）没有在本报告发行日期之前的 30 个历日内曾买卖或交易过本报告所提述的股票，或在本报告发布后 3 个工作日（定义见《证券及期货条例》(香港法例第 571 章)）内将买卖或交易本文所提述的股票；(ii) 他们或其各自的关联人士并非本报告提述的任何公司的雇员；及 (iii) 他们或其各自的关联人士没有拥有本报告提述的证券的任何金融利益。

浦银国际证券机构销售团队

杨增希

essie_yang@spdbi.com
(852) 2808 6469

浦银国际证券财富管理团队

张帆

vane_zhang@spdbi.com
(852) 2808 6467

浦银国际证券有限公司

SPDB International Securities Limited
网站：www.spdbi.com
地址：香港轩尼诗道 1 号浦发银行大厦 33 楼