

计算机行业 AI 模型系列（七）

月之暗面：从长文本破圈到 K3，前沿模型公司的进化之路

行业评级

买入

前次评级 买入

从长文本切入，历经“流量扩张、重回基模、能力变现”。月之暗面以长上下文切入市场，早期通过 Kimi Chat 及集中投流快速建立品牌认知和用户基础。2025 年，公司主动收缩营销投放，将资源重新集中于底层模型研发；2026 年以来，随着 K2.5、K2.6、K2.7 Code 及 K3 等模型连续发布，增长逻辑逐步转向模型能力驱动的 API 调用、专业订阅与海外市场拓展。

底座周期性升级+后训练高频迭代的研发体系逐步成型。公司通过预训练抬升上限，后训练释放模型能力，K2 完成万亿参数 MoE 底座建设，后续版本围绕推理、Coding、多模态及 Agent 能力持续强化；K3 进一步扩展至 2.8 万亿总参数，引入 KDA、AttnRes 及 Stable LatentMoE 等新架构，支持原生视觉能力及 100 万 Token 上下文，整体 Scaling 效率较 K2 提升约 2.5 倍，在长程 Coding、知识工作和 Agent 任务中已进入全球前沿序列。

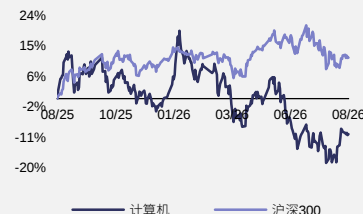
C 端产品与 B 端开放平台共同承接模型商业化，API 与海外市场成为重要增量。公司已形成 Kimi Web/App、Kimi Work、Kimi Code 及开放平台等产品矩阵，C 端承担用户入口、场景验证与订阅变现，B 端通过 API 调用和企业服务扩大模型分发，目前 API 收入在 B 端占比约 70%。此外，公司以开源模型进入海外开发者生态，并通过国际 API 和订阅产品实现付费转化；ARR 由 2026 年 3 月约 1 亿美元提升至 6 月中旬超过 3 亿美元，K3 发布后调用需求迅速增长，一度暂停 C 端新用户订阅，并披露 ARR 创下历史最大单日增幅，初步验证模型能力提升直接带动了调用及付费需求。

融资能力夯实模型迭代与算力扩容基础，算力供给仍是商业化爬坡的核心约束。据彭博报道，公司于 2026 年 7 月底完成约 35 亿美元融资，投后估值约 350 亿美元，并计划以最高约 500 亿美元估值推进上市前融资。充足资金有助于公司持续投入模型训练、人才引进及推理基础设施扩容；但 K3 上线后需求快速逼近现有服务容量上限，也表明算力获取、集群扩容、服务稳定性及单位经济性仍将约束后续收入增长的持续性。

独立模型公司定位逐步清晰，模型能力与全球商业化成为中长期核心看点。公司已逐步演进为以底层模型能力为核心、以 Agent/Coding 产品化和全球 API 变现为主要增长抓手的前沿模型公司。建议将其作为国内前沿模型核心对象持续跟踪，重点观察后续模型迭代、API 及海外收入增长的持续性、算力扩容与推理成本改善，以及融资和 IPO 推进情况。

风险提示。行业竞争加剧风险；算力供应与成本风险；全球监管政策风险；商业化不及预期风险。

相对市场表现



刘雪峰

SAC 执证号：S0260514030002
SFC CE No. BNX004
gfliuxuefeng@gf.com.cn

戴亚敏

SAC 执证号：S0260526070004
SFC CE No. BXL995
daiyamin@gf.com.cn

相关研究

计算机行业 AI 模型系列（六）：AI 基模新变化下的趋势机会	2026-08-02
计算机行业：当前位置坚定推荐国产链的几个考量点	2026-07-19
计算机行业：需求维持高景气，算租的长租重资产模式价值初显	2026-06-30

目录索引

一、月之暗面：从长文本到专注基模，前沿模型公司的快速转型.....4

 （一）从长文本切入，历经“流量扩张、重回基模、能力变现”.....4

 （二）C 端应用与 B 端开放平台双轮驱动5

 （三）高人才密度，扁平化结构.....6

二、模型端：从长上下文突破走向前沿智能，模型能力持续进化.....8

 （一）预训练抬升模型能力上限，后训练将底座能力转化为真实任务表现.....9

 （二）CODING、AGENT 与多模态能力协同演进，进入全球前沿竞争序列12

三、商业化与资本：能力变现提速，融资夯实储备17

 （一）由投流获客转向模型驱动，海外市场打开成长空间.....17

 （二）密集融资补足资金弹药，支撑前沿模型持续迭代21

四、风险提示23

 （一）行业竞争加剧风险23

 （二）算力供应与成本风险.....23

 （三）全球监管政策风险23

 （四）商业化不及预期风险.....23

图表索引

图 1: Kimi 发展历程.....	4
图 2: Kimi K1.5 的数学、代码、多模态推理能力在 long-CoT 模式下达到 OpenAI o1 正式版的水平.....	9
图 3: Kimi K2 单步训练损失曲线.....	10
图 4: K2 采用的大规模 Agent 数据合成框架.....	10
图 5: Kimi K2.7 Code 提升了 Token 效率.....	11
图 6: Kimi K3 架构.....	12
图 7: Frontend Code Arena 榜单.....	13
图 8: 多维度代码能力基准测评.....	14
图 9: Kimi K3 输出的模拟长征十号火箭发射与回收.....	14
图 10: Kimi Agent Swarm 智能集群调度架构示意图.....	15
图 11: Artificial Analysis Intelligence Index 榜单排名.....	16
图 12: Kimi App 的月活跃用户规模.....	17
图 13: Kimi 会员订阅各套餐权益详情.....	19
图 14: Kimi 算力紧缺与会员暂停开放的说明.....	20
图 15: Kimi 年化收入在 K3 发布节点录得历史最大单日增幅.....	21
表 1: Kimi 产品矩阵.....	5
表 2: 月之暗面核心创始人及管理团队.....	6
表 3: Kimi 模型迭代.....	8
表 4: Kimi 模型定价.....	18
表 5: 国内外主流模型 API 价格对比.....	18
表 6: 国内外主流模型厂商个人 AI 订阅及 Coding/Agent 套餐价格对比.....	20
表 7: 月之暗面融资历程.....	22

一、月之暗面：从长文本到专注基模，前沿模型公司的快速转型

（一）从长文本切入，历经“流量扩张、重回基模、能力变现”

月之暗面由杨植麟于 2023 年创立。杨植麟此前曾参与 Transformer-XL、XLNet 等长序列/NLP 研究，因此公司最初以长上下文作为差异化入口。公司于同年 10 月发布 Kimi Chat，凭借差异化的长上下文产品力快速积累首批原生用户，实现了 C 端破圈。

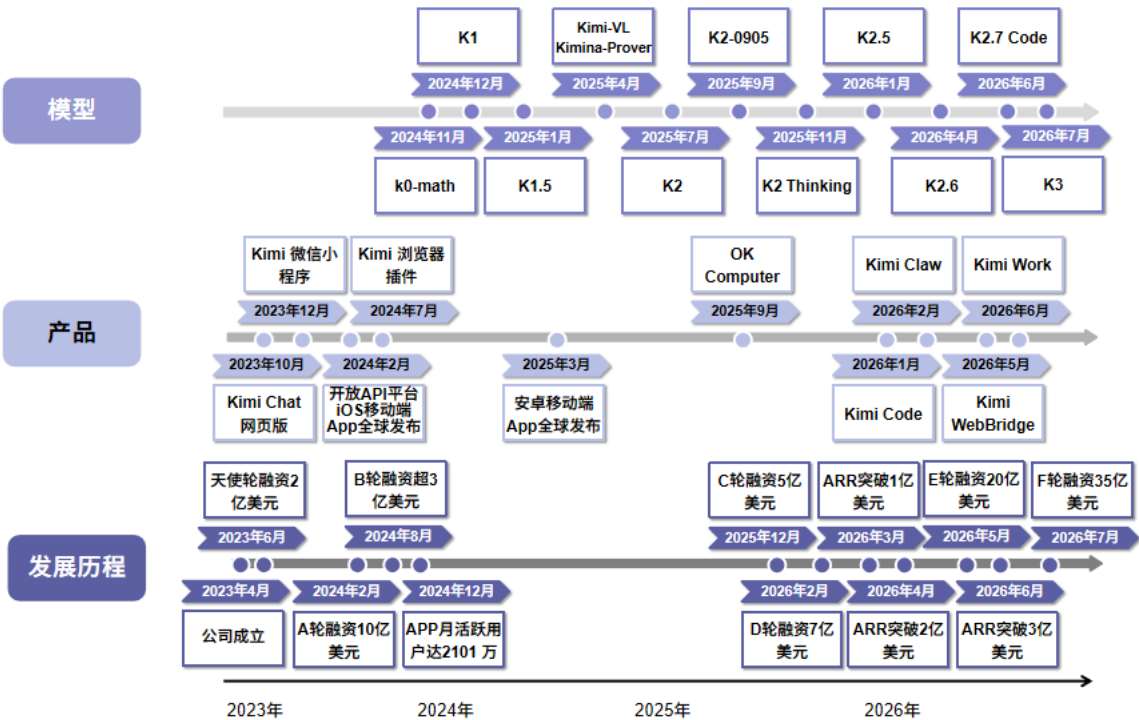
第一阶段（2023–2024）：产品与流量先行，C 端破圈。公司依靠长文本与搜索形成品牌认知，并在 2024 年投入较高营销费用争夺 C 端入口以推动用户规模快速增长。据北京日报援引 QuestMobile 数据，2024 年 12 月，Kimi App 月活跃用户数达 2101 万。通过这一阶段的集中投入，初步完成了品牌心智与用户基础积累。

第二阶段（2025）：停止大规模投流，重回基模。2025 年初，DeepSeek-R1 的发布改变了行业竞争叙事，其以较低训练成本实现接近领先闭源模型的推理能力。在此背景下，公司及时调整策略，大幅压降投流，战略重心转向底层技术研究，MAU 有所回落，但组织资源配置更接近前沿模型公司。同年 7 月，万亿参数开源 MoE 模型 Kimi K2 发布，主攻 Coding 与 Agent 场景。

第三阶段（2026 至今）：模型能力驱动商业化。2026 年，K2.5/K2.6 等模型发布后，API 与海外付费增长显著，ARR 在数月内连续跨越 1 亿、2 亿、3 亿美元。根据彭博，公司最新一轮融资最终于 7 月底超额完成，融资金额约 35 亿美元，投后估值升至约 350 亿美元。

（数据来源：Kimi 官网；Github；Arxiv；Kimi 官方公众号；北京日报；QuestMobile；证券时报；Bloomberg News）

图 1：Kimi 发展历程



数据来源：Kimi 官网，Kimi 官方微信公众号，北京日报，Sensor Tower 官网，IT 之家，北京商报，证券时报，广发证券发展研究中心

(二) C 端应用与 B 端开放平台双轮驱动

月之暗面围绕自研模型能力，逐步形成面向 C 端用户的 AI 原生应用与面向企业、开发者的模型开放平台两类业务。C 端承担用户入口、产品验证与订阅变现功能，B 端则通过 API 调用及企业服务拓展模型商业化场景，两类业务共享底层模型与推理基础设施。

C 端由 ChatBot 向通用 Agent 平台演进。早期 Kimi 主要围绕长文本阅读、联网搜索和智能问答建立用户认知，随后逐步将产品边界拓展至深度研究、网站搭建、文档处理、PPT 制作、表格分析及软件开发等复杂任务。目前，公司已形成以 Kimi 网页端及移动端为通用入口，Kimi Work 和 Kimi Code 分别面向知识工作与软件开发场景，并在 Kimi 中提供 Agent、Agent Swarm 等任务执行模式的多层次产品矩阵。其中，Kimi Work 进一步接入本地文件、外部插件以及可持续更新的组件与仪表盘，Kimi Code 则通过 CLI、IDE 插件等形态进入开发者 workflow。

C 端商业模式主要为分层订阅+额度计费。Kimi 面向不同使用强度设置免费版及 19 美元、39 美元、99 美元和 199 美元/月的四档付费会员，并通过 Agent 额度、并发任务数、Agent Swarm 调用次数、Kimi Code 额度、上下文长度及模型访问权限形成产品区分。各项功能共用统一额度池，高阶会员获得更多任务额度、更高并发能力和更快执行速度，订阅额度消耗完毕后，用户还可购买额外加油包。此类模式将用户付费与模型推理资源消耗挂钩，有助于缓解重度 Agent 用户对算力成本的挤压，并通过 Coding、深度研究和多智能体任务提升专业用户 ARPU。

B 端以模型 API 为核心，并向企业级服务延伸。Kimi 开放平台向开发者提供 K3、K2.7 Code 及 K2.6 等不同能力和价格档位的模型，按 Token 模式计费，例如，K3 海外 API 缓存命中输入、普通输入和输出价格分别为 0.3 美元、3 美元和 15 美元/百万 Token，官方披露其在 Coding 场景下缓存命中率超过 90%。除标准 API 外，公司还通过工具调用、模型兼容接口、组织管理、成员权限及数据隔离等能力服务企业客户，使商业模式由单纯出售模型调用量，逐步向企业容量、团队账户和组织级 AI 基础设施延伸。

C 端与 B 端并非相互独立，而是共同承担模型能力商业化。C 端产品将底层模型封装为用户可直接感知的任务能力，帮助公司验证真实场景、完善产品体验并建立品牌入口；B 端开放平台则将模型能力分发给开发者和企业，由外部生态拓展应用边界。开源模型进一步降低了开发者试用和适配门槛，官方 API 再通过模型更新速度、推理性能、缓存效率及服务稳定性承接商业需求。整体看，月之暗面的收入重心正由流量驱动的通用应用，转向专业订阅、Coding、Agent 等高价值场景，C 端与 B 端共同构成模型能力规模化变现的双轮驱动体系。

(数据来源：Kimi 官网)

表 1：Kimi 产品矩阵

产品	产品形态	目标客户	主要能力	商业化模式
Kimi	Web/App 通用 AI 助手与 Agent 入口	个人用户及专业用户	对话、联网搜索、深度研究、多模态文件处理、Agent 及 Agent Swarm 等复杂任务执行	Freemium 与分层会员订阅，额度耗尽后可通过加油包补充；国内与海外采取差异化会员档位和价格体系
Kimi Work	macOS/Windows 桌面端本地工作流 Agent	知识工作者及专业用户	挂载本地文件、Cron 定时任务、WebBridge 浏览器自动化、Agent Swarm、Office 内容生成及全球市场数据分析	纳入 Kimi 会员体系

Kimi Code	CLI 及 VS Code 扩展	个人开发者及研发团队	代码库读取、文件编辑、命令执行、联网检索、自动化开发及第三方开发工具接入	纳入 Kimi 会员体系，不同会员档位配置差异化 Code 额度；新会员体系 Kimi 权益将与 Kimi Code 权益拆分
Kimi 开放平台	API 开放平台及企业级 AI 服务	开发者、企业及机构客户	提供 K3、K2.7 Code 等模型 API、工具调用等	API 按 Token 计费，企业版采用订阅或定制化合作模式

数据来源：Kimi 官网，广发证券发展研究中心

（三）高人才密度，扁平化结构

1. 创始人杨植麟：“技术型”CEO，兼具学术积累与创业经验

硬核学术+工程经历，深耕自然语言处理与长序列建模。杨植麟高中阶段取得清华大学保送资格，并以优异成绩毕业，随后仅用 4 年获得卡内基梅隆大学计算机科学博士学位，在 NeurIPS、ACL 等顶会发表多篇论文，是 NLP 领域具有较高学术影响力的研究者之一。2019 年，杨植麟作为核心作者发表 Transformer-XL 和 XLNet，两篇论文均成为高引用代表作，前者通过引入分段循环机制提升模型对长程依赖的建模能力，后者则探索了新的自回归预训练范式，体现了其在长序列建模和大模型预训练方面的持续积累。博士期间，杨植麟曾在 Google Brain 和 Meta FAIR 两大实验室从事研究；回国后参与华为盘古、智源悟道等国内早期大模型项目，并联合创立循环智能，形成了从前沿研究、模型工程到创业管理的复合经验。

决策风格果断，善于把握窗口期。2023 年，公司以长上下文切入市场，在国内率先推出支持约 20 万汉字输入的 Kimi 智能助手。2025 年初，面对 DeepSeek 冲击，公司收缩流量投放、重回模型研发，并于 2025 年 7 月发布强化编程和 Agent 能力的 K2，2026 年 1 月进一步推出补齐原生多模态能力的 K2.5。2026 年 7 月，公司进一步发布 K3——全球首个达到 2.8 万亿参数的开源多模态大模型，支持原生视觉与 100 万 Token 上下文，模型能力迈入全球前沿梯队。

（数据来源：GitHub；Kimi 官网；CMU 官网；清华大学官网；ACL Anthology；NeurIPS Proceedings；金融时报；Business Insider；中国新闻网；循环智能官网；中国企业家杂志）

2. 核心管理团队：清华系背景突出，技术与商业化能力互补

公司核心管理与技术团队具有鲜明的清华背景，在 AI 领域积累深厚。从公开履历及职务分工看，杨植麟负责模型技术路线与公司整体方向，联合创始人兼 CTO 张宇韬侧重技术架构与工程体系建设，周昕宇、吴育昕深度参与模型研发与技术落地，张予彤曾任金沙江创投主管合伙人并主导投资循环智能，加入月之暗面后主要负责整体战略、商业化及融资。团队整体形成“技术研究—工程落地—资本与商业化”的协同分工，有助于公司在保持技术驱动的同时提升产品转化与商业化能力。（数据来源：证券时报；innoHere）

表 2：月之暗面核心创始人及管理团队

姓名	教育背景	职位/定位	履历
杨植麟	清华大学/CMU	创始人、董事长兼 CEO	循环智能联合创始人
张宇韬	清华大学	联合创始人兼CTO	循环智能联合创始人&CTO
周昕宇	清华大学	联合创始人	原旷视科技算法研发；ShuffleNet共同第一作者

吴育昕	清华大学/CMU	联合创始人	原Meta FAIR研究员；Group Normalization作者
张予彤	清华大学/斯坦福大学	联合创始人、总裁	前金沙江创投主管合伙人；负责整体战略、商业化及融资

数据来源：杨植麟个人主页，清华官网，CMU 官网，吴育昕个人主页，广发证券发展研究中心

3. 组织理念：精简规模，高技术密度，鼓励直接沟通与跨领域协作

坚持精简团队，扁平化治理架构，强调直接沟通与跨领域协作。杨植麟曾表示，团队规模过度扩张可能损害创新效率。张予彤在北京大学公开交流中提到，公司组织层级非常扁平，员工规模约为 300 人，以相对精简的团队支撑万亿参数模型研发及大规模用户产品运营，体现出较高的人才密度和组织效率；此外，公司鼓励员工跨岗位流动，例如从预训练转向后训练、从算法转向数据，营销人员也可以参与模型评估。这种组织模式有助于缩短决策链条、促进技术与产品协同，但对人才质量和员工自驱力要求也较高。（数据来源：北京大学光华管理学院官网）

二、模型端：从长上下文突破走向前沿智能，模型能力持续进化

Kimi 的模型迭代由早期长上下文突破，逐步转向大规模预训练与高频后训练协同推进。k1.5 验证强化学习与推理时扩展能力，K2 完成万亿参数 MoE 底座建设，K2 Thinking、K2.5、K2.6 及 K2.7 Code 进一步将底座能力迁移至推理、Coding、多模态与 Agent 任务，K3 则通过近 3 万亿参数及新一代注意力、残差和 MoE 架构，再次抬升了模型能力上限。我们认为，Kimi 模型的核心价值不仅体现在当前旗舰模型的前沿竞争力，还体现在公司已初步形成底座周期性升级、后训练高频迭代及产品快速承接的研发体系。

表 3：Kimi 模型迭代

模型	发布时间	参数	训练数据	上下文	输入输出模态	API 价格 (/1M tokens ，缓存输入/ 输入/输出)	Attention	MoE / FFN	后训练 / 对齐	关键技术迭代与定位
Kimi k1	2024-12-16	-	-	-	Text/Image→Text	-	-	-	强化学习	视觉思考模型
Kimi k1.5	2025-01-20	-	文本 + 视觉 联合训练	128K	Text/Image→Text	-	-	-	Long-CoT RL；128K RL 上下文、partial rollout、Long2Short	多模态思考模型；兼具长思维与短思维，整体能力接近 o1
Kimi K2	2025-07-11	MoE；1T total / 32B active	15.5T tokens；Web Text / Code / Mathematics / Knowledge	128K	Text→Text	\$0.15 / \$0.60 / \$2.50	MLA 64 Heads； MuonClip 约 束 Attention Logits、提升 训练稳定性	MoE；384 experts， top-8 routed + 1 shared	构建真实MCP工具+ 合成工具、Agent、任 务及多轮轨迹的数据 合成体系；联合采用 RLVR与Self-Critique Rubric Reward	万亿参数开放权重 Agentic 模型；用 MuonClip 优化器稳健支撑 万亿参数模型训练，显著 提升 token 利用效率
Kimi K2-0905	2025-09-05	MoE；1T total / 32B active	-	256K	Text→Text	\$0.15 / \$0.60 / \$2.50	继承 K2	继承 K2	重点强化真实软件工程、长程Coding及前端生成；具体增量数据和RL配方未披露	拓展上下文长度，更强的 代码能力，更快的 API
Kimi K2 Thinking	2025-11-06	MoE；1T total / 32B active	-	256K	Text→Text	\$0.15 / \$0.60 / \$2.50	继承 K2	继承 K2	后训练阶段引入 QAT；长思考与工具 调用交错训练，扩展 思考Token和工具调 用步数；	全面提升 Agent 和推理能 力
Kimi K2.5	2026-01-27	MoE；1T total / 32B active	在 K2-Base 上继续预训练 约 15T mixed visual/text tokens	256K	Text/Image/ Video→Text	\$0.10 / \$0.60 / \$3.00	继承 K2	继承 K2	多模型合成数据SFT + 文本视觉联合RL + PARL并行Agent强化 学习	全能模型，原生的多模态 架构设计；最多 100 个子 Agent、1,500 个协同步 骤，端到端耗时最高缩短 4.5x
Kimi K2.6	2026-04-20	MoE；1T total / 32B active	-	256K	Text/Image/ Video→Text	\$0.16 / \$0.95 / \$4.00	继承 K2	继承 K2	强化长程Coding、工 具调用及Agent Swarm后训练	全面精进代码和 Agent 集 群能力；Agent Swarm 扩 展至 300 个子 Agent、 4,000 个协同步骤
Kimi K2.7 Code	2026-06-12	MoE；1T total / 32B active	-	256K	Text/Image/ Video→Text	\$0.19 / \$0.95 / \$4.00	继承 K2	继承 K2	强化 Coding 场景	专项 Coding 模型，显著 提升了长上下文编程场景 的指令遵循能力、长程编 程任务的性能，大幅改善 了过度思考倾向
Kimi K3	2026-07-17	MoE； 2.8T total / 104B active	-	1M	Text/Image/ Video→Text	\$0.30 / \$3.00 / \$15.00	KDA+ Gated MLA； AttnRes 负责 跨深度选择性 提取表征	Stable LatentMoE ；896 experts、16 active（约 1.8%）	SFT阶段引入QAT， 采用MXFP4权重与 MXFP8激活；结合完 全均衡专家并行提升 超大规模MoE后训练 效率；	首个国产近 3T 级旗舰开 放模型；面向长程 Coding、知识工作与深度 推理；仅次于Claude Fable 5 和 GPT 5.6 Sol

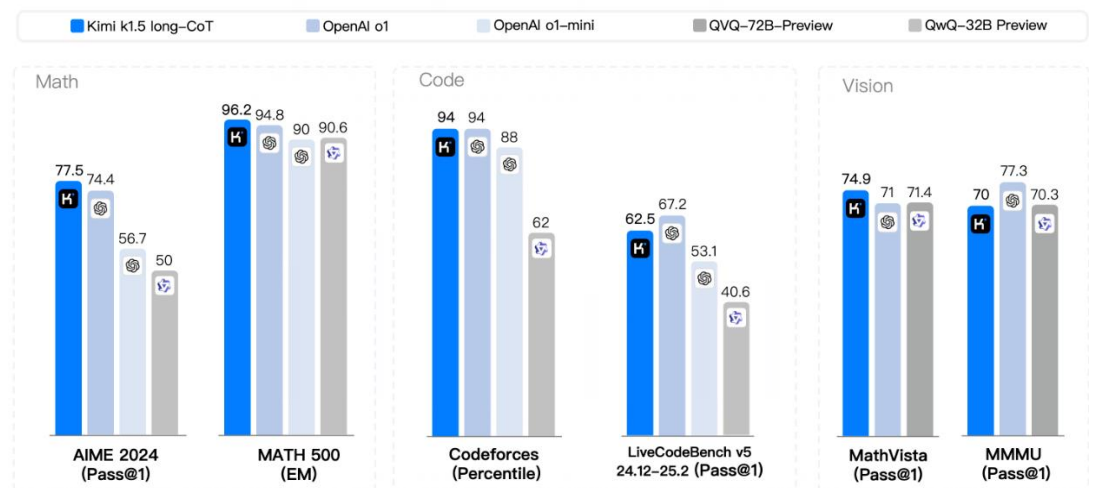
数据来源：Kimi 官网，Kimi 官方微信公众号，Arxiv，广发证券发展研究中心

（一）预训练抬升模型能力上限，后训练将底座能力转化为真实任务表现

Kimi 早期以长上下文作为差异化切口。2023 年 10 月，Kimi Chat 推出并开启内测，初期即支持约 20 万汉字的上下文输入，这一长度达到全球已产品化大模型服务中的较高水平。而长上下文能力由于更容易在研报阅读、资料总结和搜索问答中被用户感知，因此快速建立了长文本处理工具的产品认知。但从长期看，标称上下文窗口长度本身较容易被行业追赶，其竞争力更取决于模型能否在长文本中准确提取和关联信息，以及长上下文推理的速度、成本与并发效率。随着国内外头部模型陆续扩展上下文窗口，Kimi 需要由产品功能领先进一步转向基础模型和训练方法领先。（数据来源：Kimi 官方公众号）

2025 年初发布的 k1.5 成为这一转型的节点之一。不同于早期主要围绕长文本输入构建产品差异化，k1.5 将强化学习训练的上下文窗口扩展至 128K，并通过 partial rollout 提升长推理轨迹的训练效率，通过 Long2Short 将长链推理能力迁移至更短的推理输出，推动模型在数学、代码和多模态推理任务上持续提升，也为后续向 Coding 及复杂任务执行方向演进积累了方法。

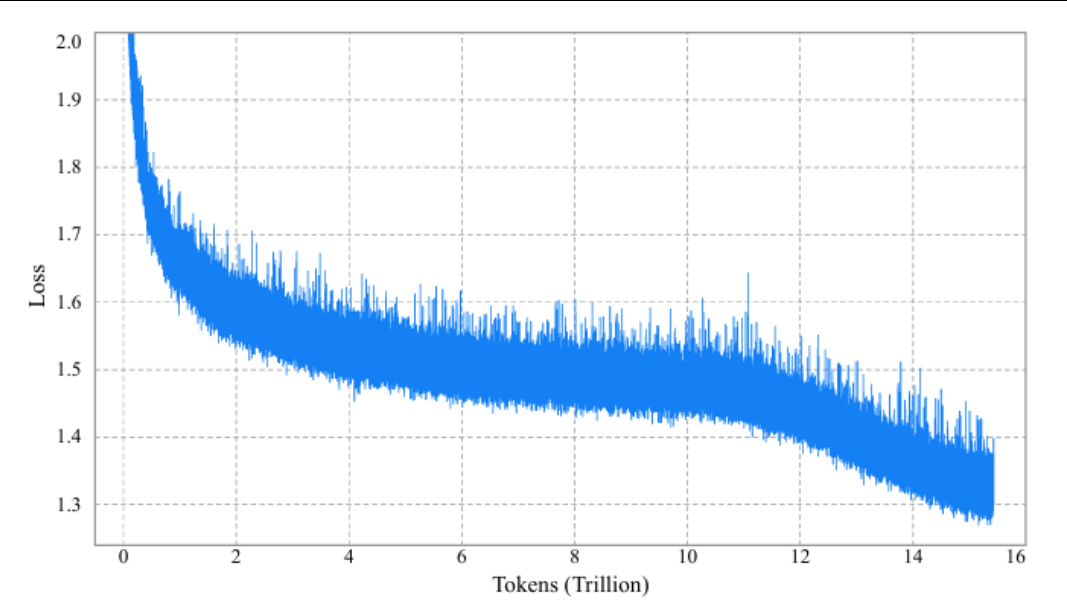
图 2：Kimi K1.5 的数学、代码、多模态推理能力在 long-CoT 模式下达到 OpenAI o1 正式版的水平



数据来源：Kimi K1.5 技术报告，广发证券发展研究中心

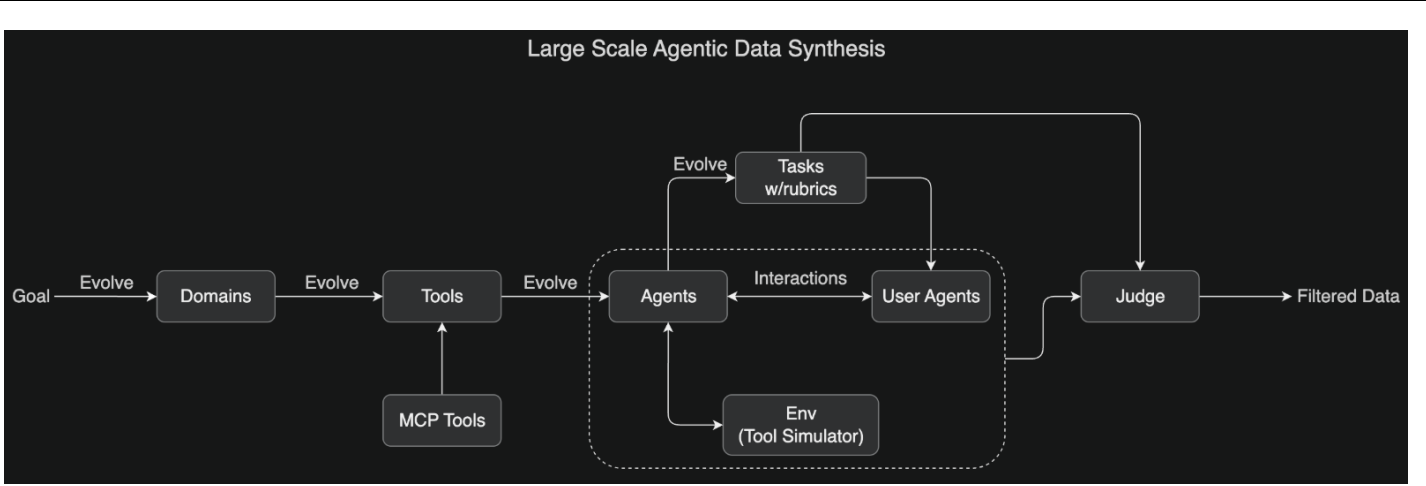
K2 完成了万亿参数底座建设。2025 年 7 月，公司发布 K2 模型，采用 MoE 架构，总参数约 1.04 万亿、单 Token 激活参数约 326 亿，设置 384 个专家、每次激活 8 个，此外还有 1 个共享专家，在提升模型容量的同时兼顾推理成本。训练端，K2 引入 MuonClip 优化器，并通过其中的 QK-Clip 机制约束注意力 logit 异常增长，最终在 15.5 万亿 Token 预训练过程中实现零 loss spike，验证了公司在万亿参数模型的数据组织、分布式训练及稳定性控制能力。后训练端，公司进一步搭建 Agent 数据合成和联合强化学习体系，使模型由回答问题向任务规划、工具调用和自主纠错延伸。K2 形成了可复用的通用底座，后续 Thinking、多模态、Coding 及 Agent 模型可在其基础上持续迭代，有助于缩短更新周期并提高预训练投入的复用效率。

图 3：Kimi K2 单步训练损失曲线



数据来源：Kimi K2 技术报告，广发证券发展研究中心

图 4：K2 采用的大规模 Agent 数据合成框架

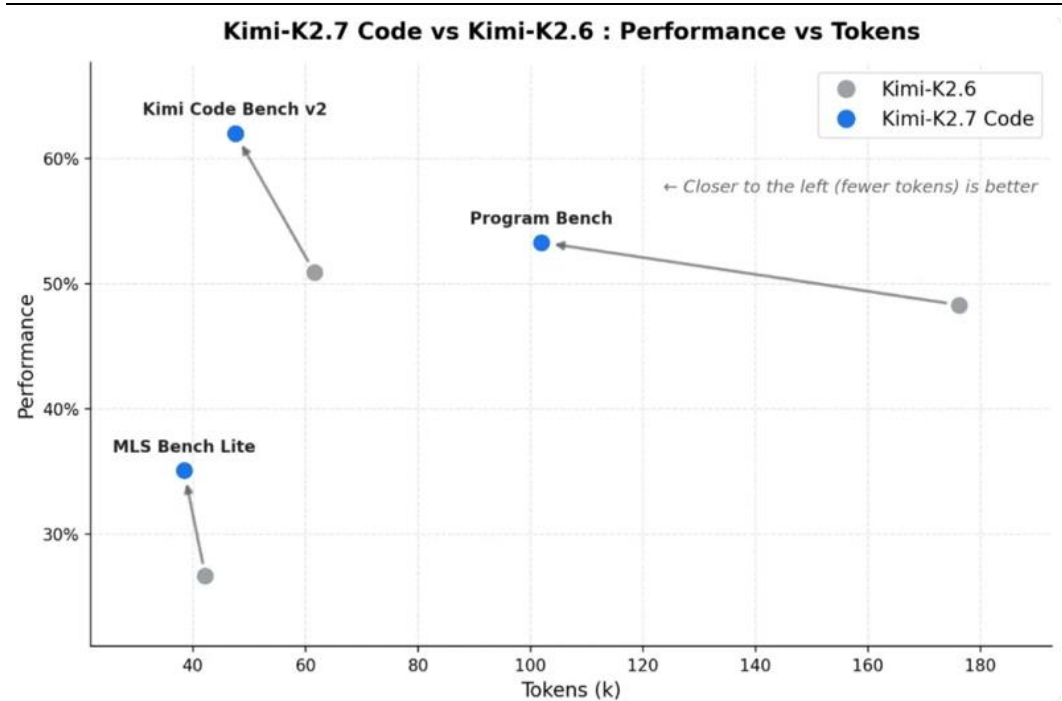


数据来源：Kimi K2 技术博客，广发证券发展研究中心

高频后训练推动底座能力持续向真实场景迁移。K2 发布后，公司围绕同一底座模型进行了多轮后训练迭代增强。K2 Thinking 强化长链推理与工具调用，并通过量化感知训练改善生成速度和部署效率；K2.5 进一步扩展至原生多模态，通过大规模图文联合预训练，使模型能够结合图像、视频和代码运行结果完成网页生成、视觉调试等任务；K2.6 进一步强化长程 Coding、主动执行和复杂 Agent 任务，并推动 Agent 集群由 100 个子 Agent、1,500 个协同步骤扩展至 300 个子 Agent 和 4,000 个步骤。K2.7 Code 则针对编程场景进行专项后训练，相较 K2.6 在 Program-Bench、MCP Mark Verified 和 SWE Marathon 等内部或特定评测上继续提升，同时将推理 Token 使用量降低约 30%。（数据来源：Kimi 官网）

以上迭代体现出后训练在模型竞争中的重要性持续提升。预训练主要决定模型的知识容量、理解能力和潜在能力上限，而后训练决定这些能力能否在 Coding、搜索、工具调用和办公任务中被稳定调用。特别是在 Agent 场景中，高质量资产已不只是静态语料，还包括可交互环境、任务验证器、工具调用轨迹和失败纠错数据。随着模型在真实环境中持续执行任务，数据和训练基础设施有望形成比传统问答数据更难复制的积累。

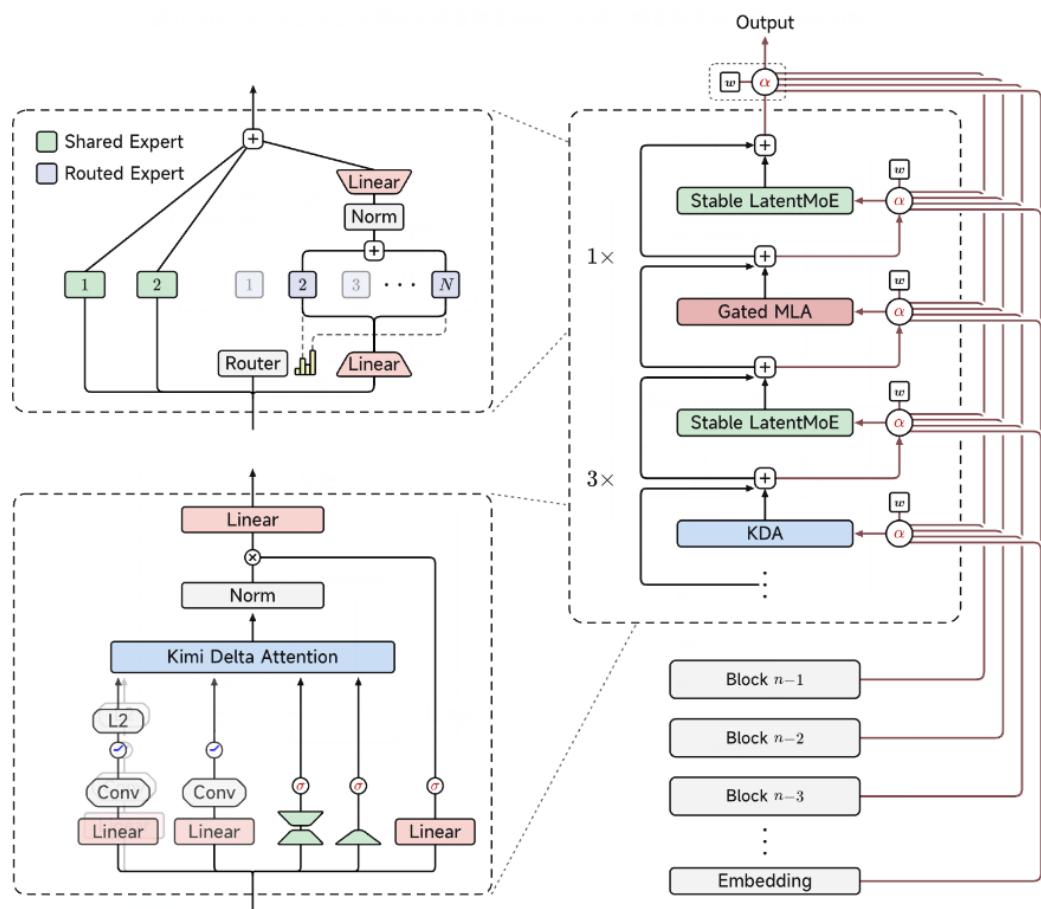
图 5: Kimi K2.7 Code 提升了 Token 效率



数据来源: Kimi K2.7 Code 技术博客, 广发证券发展研究中心

K3 重新进入底座模型升级周期。K3 并非单纯依靠扩大参数规模提升模型能力，而是围绕长序列处理、深层网络信息传递、MoE 参数利用效率及训练稳定性，对模型架构与训练体系进行协同优化。K3 总参数达到 2.8 万亿，采用 Kimi Delta Attention、Attention Residuals 和 Stable LatentMoE 等架构，其中 MoE 包含 896 个路由专家，每个 Token 有效激活 16 个专家。伴随模型架构、训练方法与数据配方的共同改进，K3 整体 Scaling 效率较 K2 提高约 2.5 倍，并成为全球首个 3T 参数量级的开源模型。（数据来源：Kimi 官网）

图 6: Kimi K3 架构



数据来源: Kimi K3 技术报告, 广发证券发展研究中心

预训练抬升上限，后训练释放能力。一方面，K2 和 K3 通过参数规模、模型架构、训练数据和方法升级，抬升通用底座的能力上限；另一方面，K2 Thinking、K2.5、K2.6 和 K2.7 Code 围绕推理、多模态、Agent 和 Coding，通过持续预训练与更高频率的后训练，使基础模型能力持续迁移至真实应用。这一模式使公司无需为每次版本升级都重新开展完整预训练，即可在既有底座上通过任务数据、监督微调、强化学习和交互环境快速迭代，并将训练过程中积累的任务样本、工具调用轨迹和结果反馈沉淀为下一阶段模型优化的重要资产。

(二) Coding、Agent 与多模态能力协同演进，进入全球前沿竞争序列

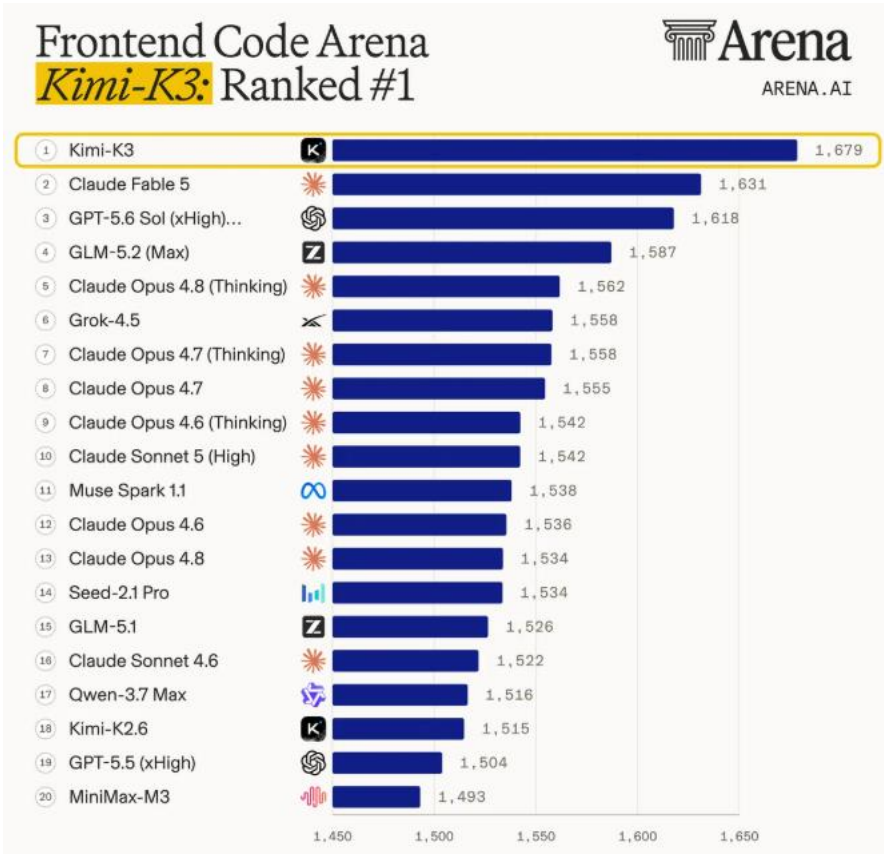
Coding 成为模型能力与商业化落地的首要结合点。一方面，代码任务具有较清晰的运行结果和测试标准，便于通过编译、单元测试和执行反馈建立可验证奖励；另一方面，开发者使用频率较高，复杂工程任务需要持续读取文件、修改代码、执行命令和修复错误，单用户 Token 消耗及付费意愿普遍高于普通问答。因此，Coding 是当前大模型率先实现商业落地的场景。

在 Coding 方面，K2 初步建立软件工程与工具调用能力；K2 Thinking 进一步融合长链推理与环境交互，在

官方评测中 SWE-bench Verified 达到 71.3%，并支持连续 200-300 次工具调用；K2.7 Code 则针对 Agentic Coding 进行专项后训练，在提高多项编程评测表现的同时，较 K2.6 平均减少约 30%的推理 Token。（数据来源：Kimi 官网）

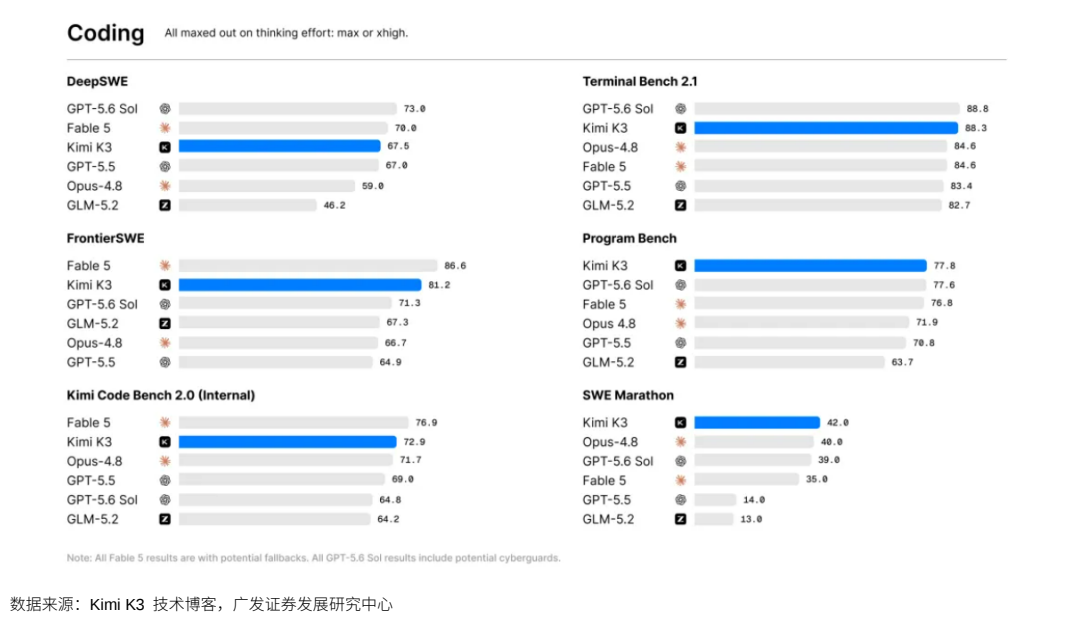
K3 进一步将 Coding 能力延伸至长程工程任务，可持续理解大型代码库、调用终端工具，并结合运行结果与视觉反馈进行多轮修改，在游戏开发、前端设计、GPU Kernel 优化等复杂任务中展现出更强的持续执行能力。2026 年 7 月 17 日，Arena.ai 公布的当期 Frontend Code Arena 最新评测结果，K3 以 1679 分位列第一，较上一代 Kimi-K2.6 的第 18 名跃升 17 位，分别领先 Claude Fable 5 和 GPT-5.6 Sol 48 分、61 分，并在 7 个前端细分场景中的 6 项排名第一。（数据来源：Arena）

图 7：Frontend Code Arena 榜单



数据来源：Arena.AI，广发证券发展研究中心
备注：数据截至 2026 年 7 月 17 日

图 8：多维度代码能力基准测评



原生多模态补齐环境感知能力，模型由生成代码转向理解结果。纯文本 Coding 模型可以生成代码和读取日志，但难以直接判断网页布局是否正确、游戏画面是否符合要求或应用界面是否出现视觉异常。引入视觉能力后，模型能够观察代码运行结果，并依据截图和视频反馈继续修改代码，为 Agent 提供理解外部环境的能力，形成更完整的执行闭环。公司此前已通过 Kimi-VL 探索多模态 MoE、长上下文视觉理解和视觉 Agent 能力；K2.5 进一步在 K2 Base 基础上开展大规模图文联合训练，使视觉理解与文本、Coding 和工具调用能力在同一模型中协同优化；K3 则在此基础上升级为原生多模态旗舰模型，并将视觉能力嵌入长程 Coding、知识工作和数字内容生产，模型不再只接收用户最初提供的图片，而可以在任务执行过程中持续读取代码运行结果，通过“vision in the loop”反复修改网页、游戏和其他可视化内容。

图 9：Kimi K3 输出的模拟长征十号火箭发射与回收



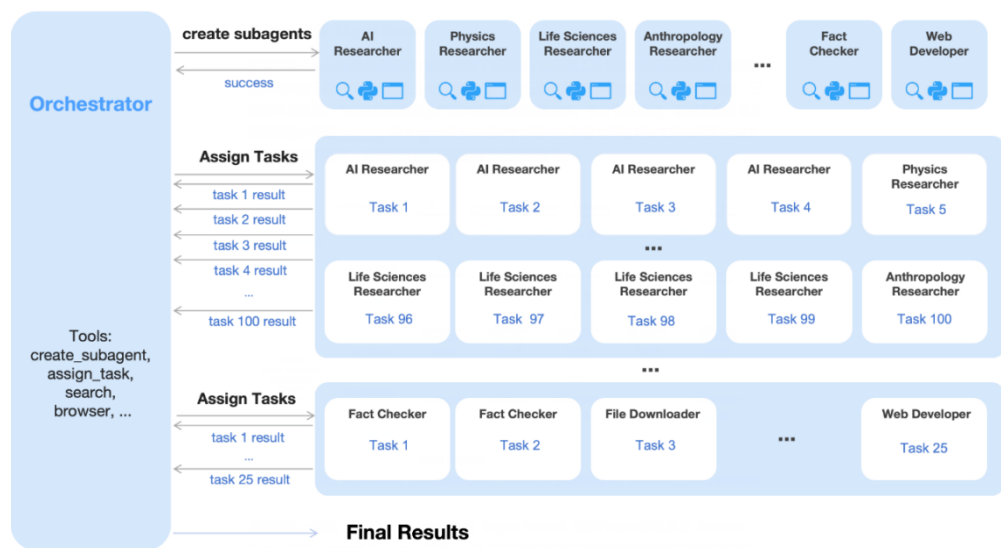
数据来源：Kimi 官方公众号，广发证券发展研究中心

Agent 集群在纵向 Scaling 之外引入横向扩展，提高复杂任务并行能力。传统 Agent 通常由单个模型顺序完成规划、搜索、阅读和写作，当任务包含大量可以同时进行的子任务时，单 Agent 会受到顺序执行的时间限制。Kimi 的 Agent 集群尝试通过横向扩展来解决这一问题，由主 Agent 自主判断任务是否适合拆分、动态创建子 Agent 并分配工作，最后对结果进行汇总，而非由开发者预先设定固定角色和工作流。

具体来看，K2.5 首次引入 Parallel-Agent Reinforcement Learning，可自主调度最多 100 个子 Agent 并执行约 1,500 次协同步骤。训练过程中，系统不仅奖励最终任务结果，也鼓励合理的并行拆分和子任务完成，避免主 Agent 虽然具备并行能力，却退化为单 Agent 顺序执行。官方测试显示，在适合并行处理的任务中，Agent Swarm 最高可将端到端时间缩短约 4.5 倍。K2.6 将规模扩展至最多 300 个子 Agent 和 4,000 个协同步骤，并允许不同子 Agent 分别承担广泛搜索、深度分析、文档处理和多格式内容生成。K3 则进一步升级了底层模型的推理、多模态理解和长程任务执行能力，当前 Kimi 的 Agent 集群已由 K3 驱动，用于大规模搜索、批量资料处理和复杂知识工作。（数据来源：Kimi 官网）

我们认为，Agent 集群代表 Kimi 在模型能力之外的系统级探索。当单模型纵向扩展面临边际收益和执行时间限制时，可以通过增加并行 Agent 数量扩大任务覆盖范围，对于数据收集、批量公司研究和多文档分析等任务具有较强的应用潜力。但 Agent 数量增加并不天然等同于经济价值提升，多 Agent 并行会显著增加 Token 消耗、工具调用和瞬时算力需求，仅当任务完成质量提高、交付时间缩短带来的用户价值超过新增推理成本时，Agent 集群才能形成可持续商业模式。

图 10：Kimi Agent Swarm 智能集群调度架构示意图

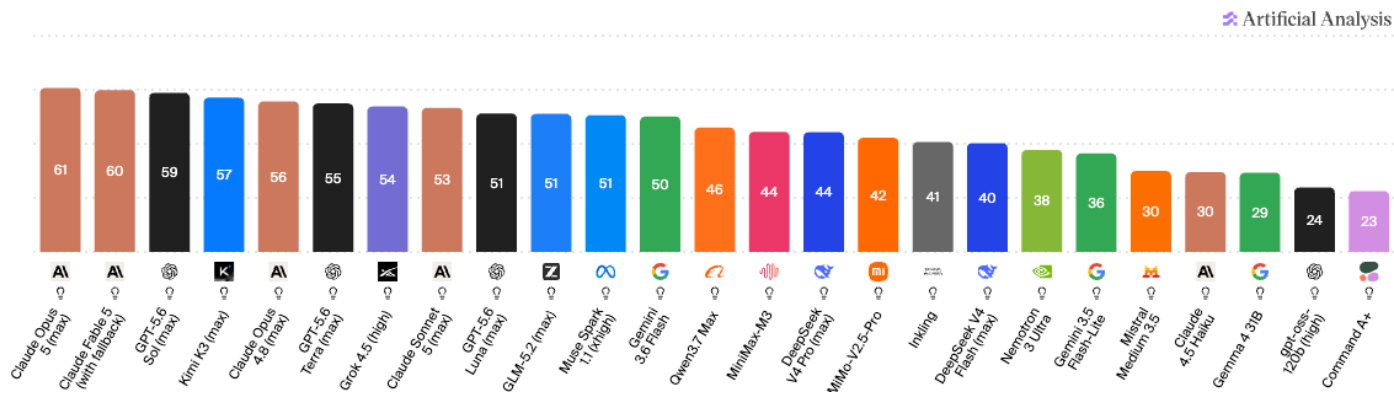


数据来源：Kimi K2.5 技术报告，广发证券发展研究中心

整体看，公司模型的综合能力已经进入全球前沿。Artificial Analysis 于 2026 年 7 月的测试显示，K3 在其 Intelligence Index 中获得 57 分、排名第四，综合能力与 Claude Opus 4.8 和 GPT-5.6 Sol 处于相近区间。在更贴近知识工作的 AA-Briefcase 评测中，K3 排名第三，仅次于 Claude Fable 5 与 Claude Opus 5 max，并

领先 GPT-5.6 Sol；在 GDPval-AA v2 中排名第七。上述结果说明，K3 的优势不仅体现在传统数学和知识问答，也开始延伸至资料分析、复杂交付物生成和长程知识工作。但第三方数据也显示 K3 的效率仍有改进空间，Artificial Analysis 测得 K3 输出速度约 36 Token/s，低于同规模大型开放权重模型约 61 Token/s 的中位水平；完成 Intelligence Index 测试时生成约 1.30 亿 Token，高于样本中位数约 9,900 万 Token，反映其在 max 推理配置下推理轨迹较长、Token 使用效率仍有进一步优化空间。

图 11：Artificial Analysis Intelligence Index 榜单排名



数据来源：Artificial Analysis，广发证券发展研究中心
备注：数据截至 2026.07.31

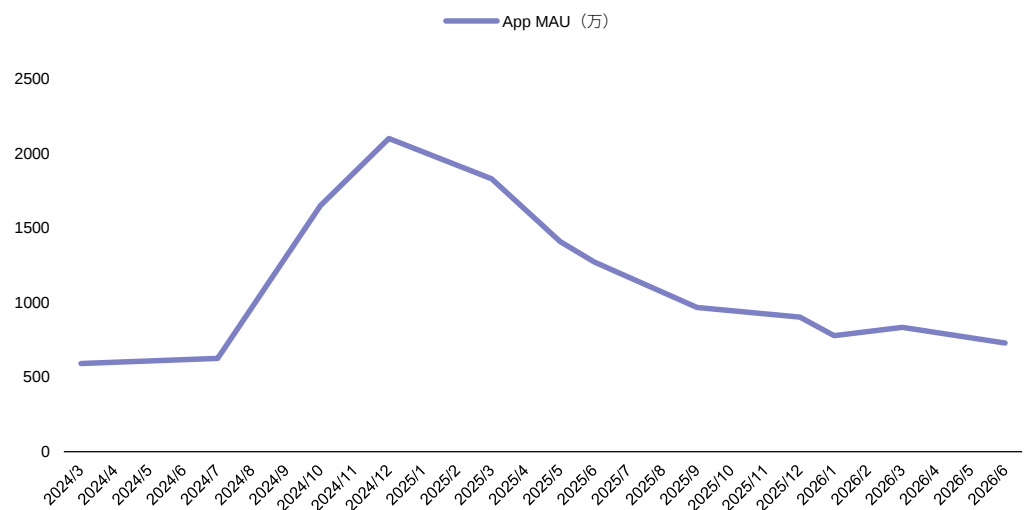


三、商业化与资本：能力变现提速，融资夯实储备

（一）由投流获客转向模型驱动，海外市场打开成长空间

早期流量先行，推动用户规模快速增长。公司依靠长文本与搜索形成品牌认知，2024 年采取较为积极的用户扩张策略，投入较高营销费用争夺 C 端入口，全年投流规模超过 7 亿元，其中第四季度单季达到 5.3 亿元（数据来源：投中网援引 AppGrowing 数据）。通过早期集中投入，Kimi 初步完成了品牌心智与用户基础积累。2025 年，公司主动收缩 C 端投流，将资源配置重心转向底层模型研发，整体用户活跃度有所回落，其中 APP 月活跃用户数由 2024 年 12 月的 2101 万下降至 2025 年 12 月的 903 万。

图 12：Kimi App 的月活跃用户规模



数据来源：QuestMobile，广发证券发展研究中心

商业化结构转型，增长重心由流量获取转向模型能力驱动的 B 端 API 调用和 C 端订阅，以及海外市场拓展。 C 端用户流量回落并不代表整体商业化进程的放缓，相较于延续高投入、低留存的买量增长模式，公司正逐步将增长逻辑切换至以底层模型能力为核心的内生驱动模式，推动商业化重心由国内 C 端用户规模扩张，转向 B 端 API 调用和 C 端订阅，以及海外市场拓展。

API 业务在 B 端收入占比约 70%。随着模型能力增强，Kimi 的 API 价格中枢逐步上移，输入输出价格分别从 K2 时期的每百万 Token 0.6/2.5 美元提升至 K3 的 3/15 美元。尽管单位 Token 价格明显提高，K3 发布后整体需求快速放量，公司一度因算力供给承压暂停新用户订阅，表明价格上调并未抑制需求，模型迭代反而带动使用强度进一步提升。横向来看，K3 的输入输出价格，较 Claude Opus 5 低约 40%，较 GPT-5.6 Sol 低约 40%-50%，虽然相较前代模型绝对价格优势有所收窄，但考虑到模型能力同步提升，K3 在高性能模型中仍具备较强的性价比。与此同时，K3 缓存命中价格仅为常规输入价格的十分之一。据公司官方数据，其 API 在编程场景中的缓存命中率超过 90%，有望显著降低长上下文、多轮交互及 Agent 任务的实际调用成本，缓解名义价格上涨对客户总使用成本的影响。（数据来源：香港新闻网，Kimi 官网）

表 4：Kimi 模型定价

模型	发布时间	缓存命中输入 (USD/1M)	缓存未命中输入 (USD/1M)	输出价格 (USD/1M)
Kimi K2	2025-07-11	0.15	0.60	2.50
Kimi K2-0905	2025-09-05	0.15	0.60	2.50
Kimi K2 Thinking	2025-11-06	0.15	0.60	2.50
Kimi K2.5	2026-01-27	0.10	0.60	3.00
Kimi K2.6	2026-04-20	0.16	0.95	4.00
Kimi K2.7 Code	2026-06-12	0.19	0.95	4.00
Kimi K3	2026-07-17	0.30	3.00	15.00

数据来源：Kimi 官网，广发证券发展研究中心

表 5：国内外主流模型 API 价格对比

公司	模型	缓存命中输入 (USD/1M)	缓存未命中输入 (USD/1M)	输出价格 (USD/1M)	备注
Anthropic	Claude Fable 5	1	10	50	-
	Claude Opus 5	0.5	5	25	-
	Claude Sonnet 5	0.2	2	10	-
OpenAI	GPT-5.6 Sol	0.5	5	30	长上下文时输入价为 2x,输出价为 1.5x
	GPT-5.6 Terra	0.2	2	12	长上下文时输入价为 2x,输出价为 1.5x
	GPT-5.6 Luna	0.02	0.2	1.2	长上下文时输入价为 2x,输出价为 1.5x
月之暗面 AI	Kimi K3	0.3	3	15	-
智谱 AI / Z.AI	GLM-5.2	0.26	1.4	4.4	-
DeepSeek	DeepSeek-V4-Pro	0.003625	0.435	0.87	-
	DeepSeek-V4-Flash	0.0028	0.14	0.28	-
MiniMax	MiniMax-M3	0.06	0.3	1.2	上下文>512K 时为 2x
阿里巴巴	Qwen3.8-Max	0.206	1.65	4.951	
腾讯	Hy3	0.033	0.132	0.528	-
字节跳动	Doubao-Seed-2.1-Turbo	0.1	0.5	2.5	-

数据来源：各公司官网，广发证券发展研究中心

C 端会员订阅是公司的另一类重要变现模式。公司目前已形成“免费层+四档付费计划”的分层产品体系，国内四档会员月费分别为 49 元、99 元、199 元和 699 元，海外市场对应月费分别为 19 美元、39 美元、99 美元和 199 美元，通过差异化的价格与权益配置，覆盖从轻度体验到高频、专业使用的个人用户及开发者需求。

(数据来源：Kimi 官网)

现阶段 Kimi 深度研究、PPT 生成、网站部署、Kimi Work 及 Kimi Code 等会员权益主要共享月度总额度池，并按照实际 Token 消耗扣减，其中 Kimi Code 另设每周额度及滚动频率限制，以控制短期高并发使用。与此同时，公司计划后续将通用会员权益与 Kimi Code 权益拆分销售，订阅体系有望由综合会员包进一步向专业功能独立定价演进。一方面，权益拆分有助于降低非 Coding 用户的付费门槛、提升通用会员的覆盖范围，另一方面，也有利于针对高频 Coding 用户实施更精细的额度管理和价值定价，打开单用户付费深度及收入上限。

(数据来源：Kimi 官网)

付费转化方面，据《上海证券报》援引 Stripe 数据，2026 年 1 月，Kimi 个人订阅支付订单数环比增长约 8280%，2 月订单数环比进一步增长 123.8%，进入 Stripe 全球榜单前十。尽管该数据主要反映 Stripe 渠道的支付表现，不能完全代表 Kimi 整体订阅规模，但其订单增速及排名快速提升，仍反映了随着模型和 Agent 产品能力升级，C 端用户付费意愿正加速释放。

图 13：Kimi 会员订阅各套餐权益详情

Andante 日常使用 ¥49 / 月 按月计费 订阅 ✓ 更多 Agent 额度 └ Office 文件处理 └ 深度研究 └ 网站部署 ✓ Kimi Code 可调用 实验性功能 ▣ 定制化看板，包含小组件任务 ▣ 定时任务，自动执行推送 ✱ 精选插件 └ 支持同花顺、天眼查等付费数据库 ▣ 制作并发布带数据库的网站	Moderato 效率升级 ¥99 / 月 按月计费 订阅 ✓ 2 倍 Agent 额度 └ Office 文件处理 └ 深度研究 └ 网站部署 ✓ Kimi Code 4 倍额度 实验性功能 ▣ 定制化看板，包含小组件任务 ▣ 定时任务，自动执行推送 ✱ 集群功能 └ 多 Agent 并行，数倍生产力 ✱ 精选插件 └ 支持同花顺、天眼查等付费数据库 ▣ 制作并发布带数据库的网站	Allegretto 专业优选 ¥199 / 月 按月计费 订阅 ✓ 4 倍 Agent 额度 └ Office 文件处理 └ 深度研究 └ 网站部署 ✓ Agent 多任务并行 ✓ Kimi Code 20 倍额度 实验性功能 ▣ 定制化看板，包含小组件任务 ▣ 定时任务，自动执行推送 ⦿ 目标模式，Agent 自主推进完成 ✱ 集群功能 └ 多 Agent 并行，数倍生产力 ✱ 专属 Kimi Claw ✱ 精选插件 └ 支持同花顺、天眼查等付费数据库 ▣ 制作并发布带数据库的网站	Allegro 全能尊享 ¥699 / 月 按月计费 订阅 ✓ 10 倍 Agent 额度 └ Office 文件处理 └ 深度研究 └ 网站部署 ✓ Agent 多任务并行 ✓ Kimi Code 60 倍额度 实验性功能 ⦿ 支持百万 Tokens 超长对话容量 ▣ 定制化看板，包含小组件任务 ▣ 定时任务，自动执行推送 ⦿ 目标模式，Agent 自主推进完成 ✱ 集群功能 └ 多 Agent 并行，数倍生产力 ✱ 专属 Kimi Claw ✱ 精选插件 └ 支持同花顺、天眼查等付费数据库 ▣ 制作并发布带数据库的网站
---	---	---	---

数据来源：Kimi 官网，广发证券发展研究中心

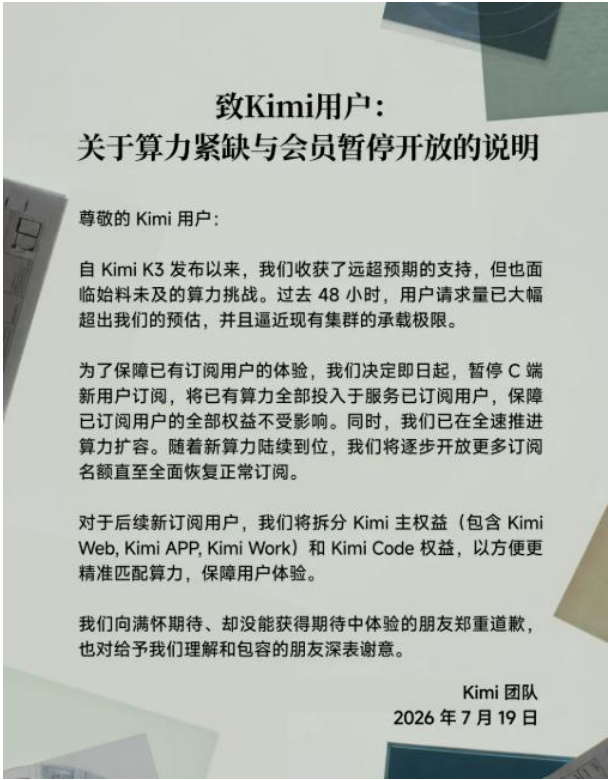
表 6：国内外主流模型厂商个人 AI 订阅及 Coding/Agent 套餐价格对比

公司/产品	价格（每月）	产品形态	主要权益及口径
OpenAI / ChatGPT	Go：\$8；Plus：\$20；Pro 5x：\$100；Pro 20x：\$200	通用AI会员	覆盖ChatGPT对话、推理、Deep Research、图像生成和Codex等能力；Codex并非独立订阅。
Anthropic / Claude	Pro：\$20；Max 5x：\$100；Max 20x：\$200	通用AI会员	覆盖Claude网页端、Claude Code、Cowork、Research等；不同档位主要区别为使用额度。
月之暗面 / Kimi	Andante：¥49；Moderato：¥99；Allegretto：¥199；Allegro：¥699	通用AI会员	Kimi Web、App、Work、Agent、PPT、研究与Kimi Code共用一套会员体系；Kimi Code额度分别约为1倍、4倍、20倍和60倍。
MiniMax / MiniMax Code	Token Plan Plus：¥49；Max：¥119；Ultra：¥469	个人Coding/Agent订阅	Token Plan主要面向MiniMax Code、编程Agent及个人开发者，同时覆盖M3/M2.7、图像、语音、音乐等模型，并非传统聊天App会员。
智谱AI / ZCode	GLM Coding Plan Lite：¥118；Pro：¥538；Max：¥1078	个人Coding/Agent订阅	实际购买的是GLM Coding Plan；套餐额度可直接在ZCode及官方支持的其他Coding Agent中使用，API调用另行计费。

数据来源：各公司官网，广发证券发展研究中心

需求快速增长推动算力扩容，短期暂停新客订阅并优化权益分层。根据官方公众号，K3 发布后模型调用需求显著超出预期，过去 48 小时用户请求量已接近现有集群承载上限。为优先保障存量付费用户的使用体验，公司自即日起暂停 C 端新用户订阅，并将现有可用算力优先配置给已订阅用户。针对后续新增用户，Kimi 计划将通用会员权益与 Kimi Code 权益拆分，其中通用权益覆盖 Kimi Web、Kimi App 及 Kimi Work，Kimi Code 则采用独立权益体系。此次调整一方面反映 K3 发布后用户需求及付费转化表现强劲，另一方面也有助于公司根据不同场景的算力消耗特征实施差异化定价和资源配置，提升有限算力的使用效率及订阅业务的变现弹性。

图 14：Kimi 算力紧缺与会员暂停开放的说明

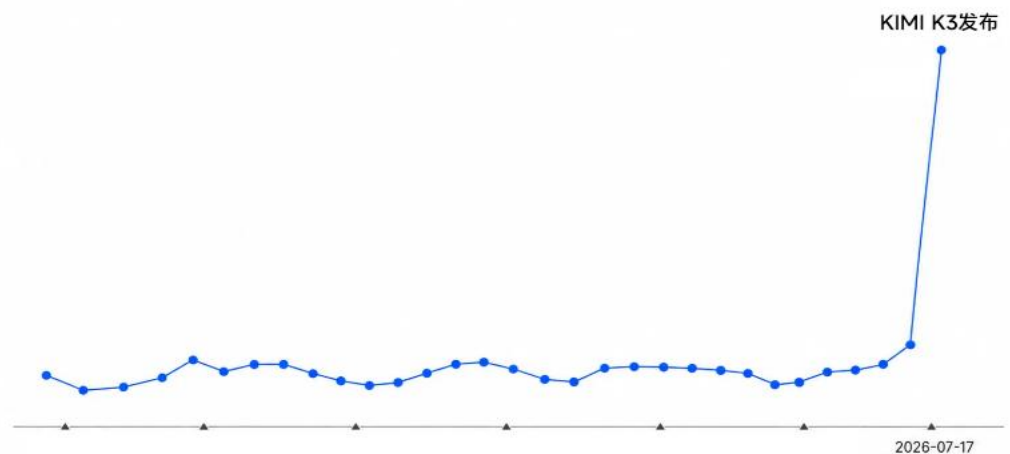


数据来源：Kimi 官方公众号，广发证券发展研究中心

以开源模型为全球获客入口，基于 Coding、Agent 等场景进入海外开发者生态，并通过国际 API 和订阅产品实现商业转化。传统互联网公司出海，需要买量、投广告以及做本地运营，但大模型公司可以通过开源模型和开发者生态，显著降低对早期大规模买量及本地化获客的依赖。2025 年 7 月，公司发布 Kimi K2 并开放模型权重，海外开发者可以下载权重、通过第三方推理平台调用，也可以把模型接入自己的 Coding Agent 和应用。2025 年 11 月，公司发布并开源 K2 Thinking，海外 API 收入增长约 4 倍，海内外付费用户数月度环比增速超过 170%。2026 年 1 月，K2.5 进一步补齐了多模态、Coding 和 Agent 能力，发布后数日内，全球付费用户增长至此前的约 4 倍，海外收入超过国内收入，初步验证了模型能力驱动的全球商业化路径。（数据来源：Kimi 官方技术博客，Github，OpenRouter 官网，界面新闻，经济观察网，南方日报）

ARR 由 2026 年 3 月约 1 亿美元提升至 6 月约 3 亿美元，K3 发布后录得历史 ARR 最大增幅。2026 年，公司 ARR 呈现明显加速态势，3 月初突破 1 亿美元，4 月进一步超过 2 亿美元，至 6 月中旬突破 3 亿美元。2026 年 7 月 17 日，公司发布 K3 模型，在技术能力和复杂任务表现上实现进一步突破，并迅速转化为开发者调用和付费需求。发布后，公司年化收入录得历史最大单日增幅，进一步体现出模型能力对用户需求、API 调用及收入增长的直接拉动。（数据来源：北京商报，证券时报，经济日报）

图 15：Kimi 年化收入在 K3 发布节点录得历史最大单日增幅



数据来源：每日经济新闻，经济日报，证券时报，广发证券发展研究中心

（二）密集融资补足资金弹药，支撑前沿模型持续迭代

前沿模型能力为公司提供了融资基础，多元资本接力入场。红杉中国、真格基金等头部 VC 较早参与，为公司提供早期资金支持；阿里巴巴、腾讯、美团龙珠等产业资本多轮加码，在云计算、开发者生态及应用场景等方面具备潜在协同空间；中国移动、CPE 源峰等机构则在 2026 年 5 月的新一轮融资中入场，进一步丰富了公司的股东结构，亦反映出其对公司技术路线和商业前景的认可。

公司融资和估值在 2026 年明显加速。2025 年 12 月，公司完成 5 亿美元 C 轮融资，投后估值约 43 亿美元。2026 年 5 月，公司完成约 20 亿美元融资，投后估值超过 200 亿美元。2026 年 6 月，公司启动新一轮融资，

最初计划融资约 10 亿至 20 亿美元，对应约 315 亿美元投前估值。随着 K3 发布后市场关注度及融资需求提升，据彭博报道，该轮融资最终于 7 月底超额完成，融资金额约 35 亿美元，投后估值升至约 350 亿美元。公司随后提前启动上市前最后一轮融资洽谈，目标投前估值约 500 亿美元，并推进赴港上市准备。（数据来源：IT 之家，Bloomberg）

前沿模型研发具备较强的资本密集属性，融资使公司能够提前储备、为模型 Scaling 及全球化商业拓展提供资源保障。随着模型规模持续扩张，训练与推理对算力、数据、人才的要求同步提升。截至 2025 年 12 月 31 日，公司现金储备超过 100 亿元人民币，相关资金将主要用于扩大 GPU 储备、加快 K3 研发，并为后续业务扩张提供资金保障。充足的资金储备也有利于公司持续投入研发，有望持续保持前沿的模型能力并将技术优势转化为商业化收入，形成正向循环。

（数据来源：南华早报，南方新闻）

表 7：月之暗面融资历程

时间节点	轮次/融资事项	融资金额 (亿美元)	估值 (亿美元)	估值口径	融资状态	关键投资方
2023/06	天使轮	2	3	投后	已完成	红杉中国、真格基金、今日资本、砺思资本
2024/02	A轮	10	25	投后	已完成	阿里巴巴（领投）、红杉中国、美团、小红书、砺思资本
2024/08	B轮	3	33	投后	已完成	腾讯、高榕创投、阿里巴巴
2025/12	C轮	5	43	投后	已完成	IDG（领投）、阿里巴巴、腾讯
2026/02	D轮	>7	100	投后	已完成	阿里巴巴、腾讯、五源资本、九安医疗
2026/05	E轮	20	200	投后	已完成	美团龙珠（领投）、中国移动、CPE源峰、国智投、顺禧基金
2026/07	F轮	>35	350	投后	已完成	未披露
2026/08E	Pre-IPO轮	未披露	500	投前	启动谈判	未披露

数据来源：EE Times China，界面新闻，爱企查，企查查，广发证券发展研究中心

四、风险提示

（一）行业竞争加剧风险

全球大模型赛道竞争趋于白热化，海外受 OpenAI、Anthropic 等头部厂商技术与市场挤压，国内互联网巨头及同业玩家持续加码研发、场景落地与商业化布局，或对行业公司市场份额及盈利空间造成压制。

（二）算力供应与成本风险

全球高端 GPU 算力紧缺，算力供给紧张或制约行业公司模型迭代、推理扩容及业务扩张进度。同时算力采购与运维成本较高，若商业化变现效率无法覆盖成本压力，将直接挤压行业公司盈利空间、拖累盈利水平。

（三）全球监管政策风险

全球各国 AI 监管政策持续收紧，数据安全、内容合规、跨境数据流动、技术出口管制等合规要求不断升级，各国监管规则存在显著差异与不确定性，或对行业公司全球化布局、海外业务落地及常态化经营形成约束。

（四）商业化不及预期风险

模型公司商业化仍处于早期扩张阶段，若后续海内外付费用户增长、B 端客户拓展、垂直行业渗透及海外扩张节奏不及预期，将导致业绩增速承压。

广发计算机行业研究小组

刘雪峰：首席分析师，东南大学工学士，中国人民大学经济学硕士，1997年起先后在数家IT行业跨国公司从事技术、运营与全球项目管理
工作。2010年就职于招商证券研究发展中心负责计算机组行业研究工作，2014年加入广发证券发展研究中心。

周源：资深分析师，慕尼黑工业大学硕士，2021年加入广发证券，曾任职于TUMCREATE自动驾驶科技公司，负责大数据相关工作。

王钰翔：高级分析师，哥伦比亚大学运筹学硕士，2024年加入广发证券发展研究中心。

戴亚敏：高级分析师，北京大学金融硕士，2024年加入广发证券发展研究中心。

分析师及联系人



刘雪峰

SAC 执证号：S0260514030002
SFC CE No. BNX004
gfliuxuefeng@gf.com.cn



戴亚敏

SAC 执证号：S0260526070004
SFC CE No. BXL995
daiyamin@gf.com.cn

广发证券

买入：预期未来 12 个月内，股价表现强于大盘 10%以上。

行业投资评级说明

持有：预期未来 12 个月内，股价相对大盘的变动幅度介于-10%~+10%。

卖出：预期未来 12 个月内，股价表现弱于大盘 10%以上。

广发证券

买入：预期未来 12 个月内，股价表现强于大盘 15%以上。

公司投资评级说明

增持：预期未来 12 个月内，股价表现强于大盘 5%-15%。

持有：预期未来 12 个月内，股价相对大盘的变动幅度介于-5%~+5%。

卖出：预期未来 12 个月内，股价表现弱于大盘 5%以上。

联系我们

地址

邮政编码

客服邮箱

广州市： 广州市天河区马场路 26 号广发证券大厦 47 楼

510627

gfzqyf@gf.com.cn

深圳市： 深圳市福田区益田路 6001 号太平金融大厦 31 层

518026

北京市： 北京市西城区月坛北街 2 号月坛大厦 18 层

100045

上海市： 上海市浦东新区南泉北路 429 号泰康保险大厦 37 楼

200120

香港： 香港湾仔骆克道 81 号广发大厦 27 楼

-

法律主体声明

本报告由广发证券股份有限公司或其关联机构制作，广发证券股份有限公司及其关联机构以下统称为“广发证券”。本报告的分销依据不同国家、地区的法律、法规和监管要求由广发证券于该国家或地区的具有相关合法合规经营资质的子公司/经营机构完成。

广发证券股份有限公司具备中国证监会批复的证券投资咨询业务资格，接受中国证监会监管，负责本报告于中国（港澳台地区除外）的分销。

广发证券（香港）经纪有限公司具备香港证监会批复的就证券提供意见（4 号牌照）的牌照，接受香港证监会监管，负责本报告于中国香港地区的分销。

本报告署名研究人员所持中国证券业协会注册分析师资质信息和香港证监会批复的牌照信息已于署名研究人员姓名处披露。

重要声明

广发证券股份有限公司及其关联机构可能与本报告中提及的公司寻求或正在建立业务关系，因此，投资者应当考虑广发证券股份有限公司及其关联机构因可能存在的潜在利益冲突而对本报告的独立性产生影响。投资者不应仅依据本报告内容作出任何投资决策。投资者应自主作出投资决策并自行承担投资风险，任何形式的分享证券投资收益或者分担证券投资损失的书面或者口头承诺均为无效。

本报告署名研究人员、联系人（以下均简称“研究人员”）针对本报告中相关公司或证券的研究分析内容，在此声明：（1）本报告的全部分析结论、研究观点均精确反映研究人员于本报告发出当日的关于相关公司或证券的所有个人观点，并不代表广发证券的立场；（2）研究人员的部分或全部的报酬无论在过去、现在还是将来均不会与本报告所述特定分析结论、研究观点具有直接或间接的联系。

研究人员制作本报告的报酬标准依据研究质量、客户评价、工作量等多种因素确定，其影响因素亦包括广发证券的整体经营收入，该等经营收入部分来源于广发证券的投资银行类业务。

本报告仅面向经广发证券授权使用的客户/特定合作机构发送，不对外公开发布，只有接收人才可以使用，且对于接收人而言具有保密义务。广发证券并不因相关人员通过其他途径收到或阅读本报告而视其为广发证券的客户。在特定国家或地区传播或者发布本报告可能违反当地法律，广发证券并未采取任何行动以允许于该等国家或地区传播或者分销本报告。

本报告所提及证券可能不被允许在某些国家或地区内出售。请注意，投资涉及风险，证券价格可能会波动，因此投资回报可能会有所变化，过去的业绩并不保证未来的表现。本报告的内容、观点或建议并未考虑任何个别客户的具体投资目标、财务状况和特殊需求，不应被视为对特定客户关于特定证券或金融工具的投资建议。

本报告发送给某客户是基于该客户被认为有能力独立评估投资风险、独立行使投资决策并独立承担相应风险。

本报告所载资料的来源及观点的出处皆被广发证券认为可靠，但广发证券不对其准确性、完整性做出任何保证。报告内容仅供参考，报告中的信息或所表达观点不构成所涉证券买卖的出价或询价。广发证券不对因使用本报告的内容而引致的损失承担任何责任，除非法律法规有明确规定。客户不应以本报告取代其独立判断或仅根据本报告做出决策，如有需要，应先咨询专业意见。

广发证券可发出其它与本报告所载信息不一致及有不同结论的报告。本报告反映研究人员的不同观点、见解及分析方法，并不代表广发证券的立场。广发证券的销售人员、交易员或其他专业人士可能以书面或口头形式，向其客户或自营交易部门提供与本报告观点相反的市场评论或交易策略，广发证券的自营交易部门亦可能会有与本报告观点不一致，甚至相反的投资策略。报告所载资料、意见及推测仅反映研究人员于发出本报告当日的判断，可随时更改且无需另行通告。广发证券或其证券研究报告业务的相关董事、高级职员、分析师和员工可能拥有本报告所提及证券的权益。在阅读本报告时，收件人应了解相关的权益披露（若有）。

本研究报告可能包括和/或描述/呈列期货合约价格的事实历史信息（“信息”）。请注意此信息仅供用作组成我们的研究方法/分析中的部分论点/依据/证据，以支持我们对所述相关行业/公司的观点的结论。在任何情况下，它并不（明示或暗示）与香港证监会第5类受规管活动（就期货合约提供意见）有关联或构成此活动。

权益披露

(1) 广发证券（香港）跟本研究报告所述公司在过去12个月内并没有任何投资银行业务的关系。

版权声明

未经广发证券事先书面许可，任何机构或个人不得以任何形式翻版、复制、刊登、转载和引用，否则由此造成的一切不良后果及法律责任由私自翻版、复制、刊登、转载和引用者承担。