

证券研究报告·行业动态报告

月之暗面（Kimi）：从长上下文到率先迈入万亿参数，Agentic可期 ——国产模型系列深度（3）

分析师：杨艾莉

yangaili@csc.com.cn

010-85156448

SAC编号：S1440519060002

SFC编号：BQ1330

分析师：杨晓玮

yangxiaowei@csc.com.cn

SAC编号：S1440523110001

分析师：马晓婷

maxiaoting@csc.com.cn

SAC编号：S1440525070013

发布日期：2026年6月8日

本报告由中信建投证券股份有限公司在中华人民共和国（仅为本报告目的，不包括香港、澳门、台湾）提供。在遵守适用的法律法规情况下，

更多免费行业报告

并购家

www.ipoipo.cn

核心观点

月之暗面（Moonshot AI）成立于2023年3月。同年10月，公司以中文200万字长文本能力被市场关注，Kimi智能助手开始逐步在C端破圈。2024年公司加大投放并上线会员订阅，整体MAU一度高达3600万，成为第一批国产Chatbot中的顶流之一。2025年随着DeepSeek-R1的上线，公司开始停止投放流量，回归模型层面的研发，2025年7月推出Kimi k2模型，是国内首个万亿参数的开源MoE模型，开始深耕Coding与Agentic场景。2026年1月Kimi K2.5版本发布，模型能力进一步提升。进入2026年，根据媒体报道公司ARR快速增长，3月初突破1亿美元，4月继续增长至超2亿美元。根据华峰资本，Kimi最新估值超200亿美金。

一、Kimi的长上下文基因与早期对C端的探索

1、Kimi的模型Taste，很大程度来自创始人杨植麟及其清华系创始团队的技术背景。杨植麟在清华与CMU攻读计算机期间师承顶级AI学者，包括现Apple AI首席研究员与Google AI首席科学家。在2019年发布的两篇论文，是现代长上下文技术的奠基性论文之一（Transformer XL 和 XLNet，被引用超过1万次），我们认为这也是其创立月之暗面坚持长上下文思想的来源。其他核心创始人，张宇韬/周昕宇/吴育昕（联合创始人）都出身清华计算机系，团队技术功底强。

2、2024年，Kimi以长上下文率先突围国产Chatbot，获得用户认可。2023年10月，Kimi的模型就主打全球最长20万汉字无损上下文，在“百模大战”初期就凭借差异化快速突围。2024年4月到10月，是Kimi投放最激进的半年，当时Kimi、豆包、元宝并称AI三大投流大户，有数据统计24年10月单月行业三家投流总额超过30亿元人民币。24年10月Kimi的MAU冲至3600万，位列当时国产Chatbot第一梯队。

二、25年战略转型，聚焦模型训练，推出K2正式迈入万亿参数俱乐部

1、2025年2月DeepSeek推出后，Kimi宣布暂停投放，公司战略资源重新聚焦模型训练。我们认为第一阶段Kimi在C端的尝试虽然没有大获全胜，但早期就显示出与其他国产模型气质不同之处：有技术基础的同时也具备对商业的敏锐度，在与阿里/腾讯共同竞争环境下，也仍然能在用户侧获取不小声量。

2、K2模型的重要转折：首个1T参数的开源国产模型。

2024年，市场开始担心大模型智能水平和Scaling Law接近上限，部分公司转而走向小而美的性价比路线。而Kimi坚信Scaling Law，K2是首个1T参数的开源国产模型。作为一家独角兽，Kimi训练1T量级参数的大模型需要非常强的工程化能力压缩Logit，K2的MuonClip优化器引入了QK-Clip机制，是目前公开文献万亿参数级MoE模型训练稳定性的优秀记录。

3、K2.5/2.6的推理优化能力优秀，KV优化，以及Agent Swarm集群。26年的K2.5是K系列首个原生多模态旗舰，在维持万亿参数规模的同时引入自研视觉编码器MoonViT-3D，实现纯语言到原生多模态的架构跨越。能力重心向代码与Agentic方向集中发力。K2.6又引入Agent Swarm，单次会话最多调度300个子 Agent、协调4000步，相比K2.5的100个子 Agent /1500步，分别增长200%/167%。

4、对比其他国产模型：与智谱定价同梯队，Coding和Agentic能力强。K2.6模型的定价为输入\$0.95/输出\$4.00/每百万tokens，是DeepSeekV4-Pro（\$0.14输入）的7倍、MiniMaxM2.5（\$2.40输出）的1.7倍。如果对标海外仍比GPT-5.2、ClaudeSonnet4.6便宜1/2至1/4。K2.5的SWE-bench Verified得分76.8%，超越同期主流闭源模型，标志着Kimi从“规模突破”转向“能力落地”。

三、商业化&估值：26年ARR攀升，最新估值超200亿美金

1、ARR：在2025年7月K2开源之前Kimi的商业化相对平缓。**26年3月**随着K2.5的发布，ARR突破1亿美金，4月ARR进一步环比翻倍至2亿美金。

2、海外市场对Kimi非常重要，主要通过三类渠道触达用户：1) 主流公有云，有效触达B端企业客户，已入驻AWS Bedrock、GCP Vertex AI、Azure Foundry；2) 推理优化平台，对延迟和成本敏感的规模化开发者，目前已接入Groq、Together AI、Fireworks AI、SiliconFlow等；3) 各类**Coding Agent**，对接日常普通个人开发者，目前已集成在Cursor、Cline、Roo Code等AI编程工具。

3、最新估值攀升至200亿美金。2026年5月Kimi完成最新一轮20亿美元扩展轮融资，投后估值达200亿美元，股东汇聚阿里、腾讯、美团等核心投资方。目前，Kimi 创立至今已累计融资超过 50 亿美金，这一数字基本上在一众国产模型（IPO前累计融资额）中名列前茅。

展望未来，我们认为 Kimi 的潜力巨大：

- 1) 公司是国内第一批迈向1T等超大参数训练的开源模型之一，核心团队坚信 scaling law并追求智能的绝对上限。我们认为随着Kimi后续推出更大规模参数模型（K3），公司模型能力的提升将持续。
- 2) 基模能力强+Harness强。其KV与Agent Swarm等算法与工程成本优化能力强，对于未来抢占 Coding 和 Agentic 市场具备优势。
- 3) 从早年公司愿意砸钱与阿里腾讯共同抢占C端，证明其对商业敏感且愿意下注，这在后续大模型马拉松式的商业化竞争中将展现其长远的价值和意义。

1

Kimi长上下文基因与早期C端探索

2

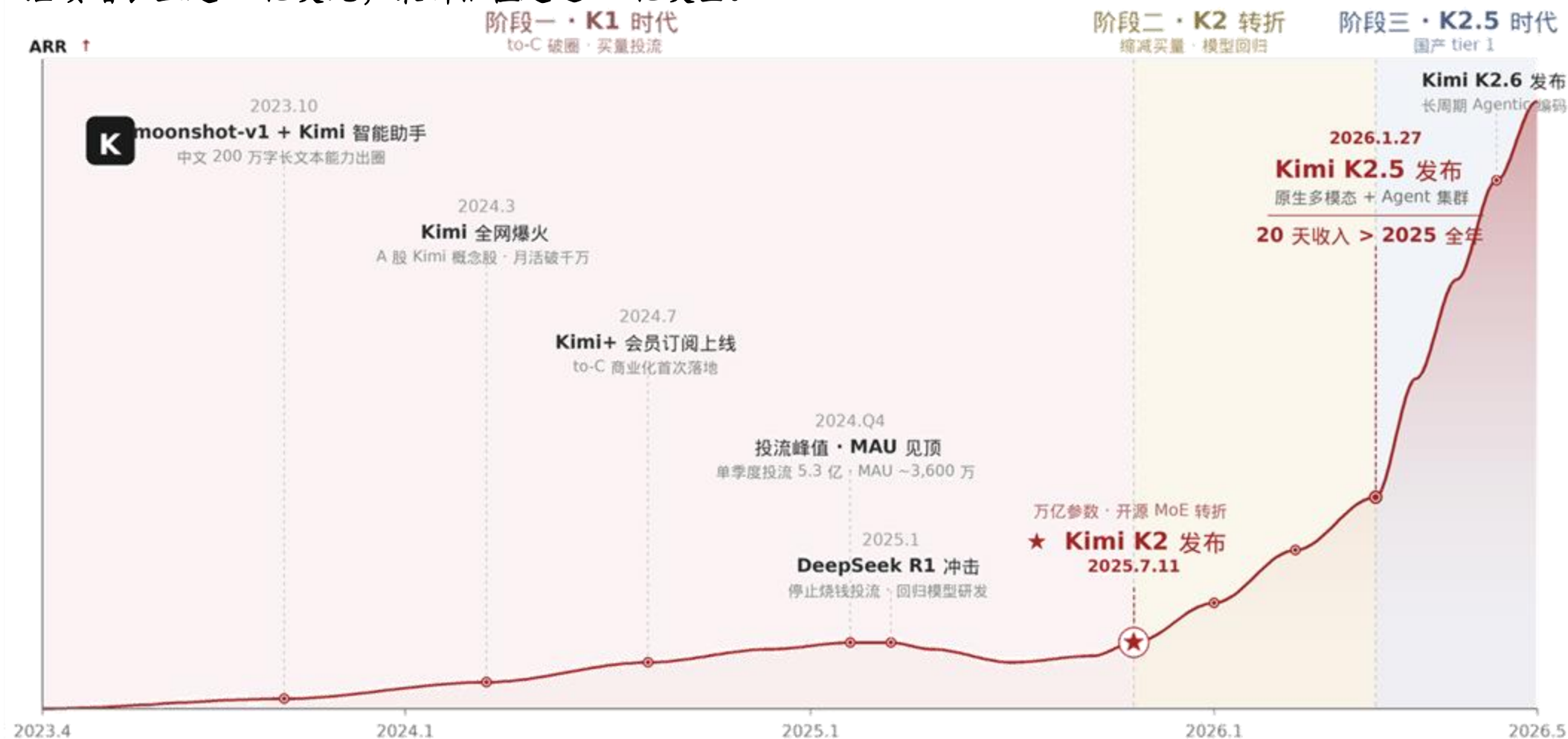
K2/2.5的转折为何重要：研发和工程化的差异优势

3

商业化与估值走到了哪一步？

（月之暗面）：国产模型，初期 端破圈，深耕 与

- ✓ 月之暗面（Moonshot AI）成立于2023年3月。同年10月公司以中文200万字长文本能力破圈，Kimi智能助手开始逐步在C端获得用户认可。2024年公司加大投放并上线会员订阅，整体MAU一度高达3600万，成为第一批国产Chatbot头部产品。2025年随着DeepSeek-R1的上线，公司开始停止投放流量，回归模型层面的研发，2025年7月推出的Kimi K2模型是国内首个万亿参数的开源MoE模型，开始深耕Coding与Agentic场景。
- ✓ 2026年1月Kimi K 2.5版本发布，模型能力进一步提升。进入2026年公司ARR快速增长：3月初突破1亿美元，4月继续增长至超2亿美元，最新估值超过200亿美金。



基因与Taste: 创始人杨植麟 “Lossless Long Context is everything” 始于长上下文

- 创始人杨植麟是全球人工智能领域的核心人才之一，多次公开讲话/访谈中提到tech vision/taste的重要性，并坚持创新和第一性原理，多次主导Kimi模型的技术创新。
- 计算机学术背景浓厚，名门正派：清华本科保送+高考双重通过，CMU仅用4年完成博士学位（标准是6年），Transformer-XL和XLNet的核心作者。他也是中国35岁以下NLP（自然语言处理）领域引用量最高的研究者。
- 其中，杨植麟在2019年发布的两篇论文是现代长上下文技术的奠基性论文之一，包括 Transformer XL 和 XLNet，论文引用超过1万次。我们认为这也是其后来创立月之暗面并一直坚持的长文本的来源。

图表：杨植麟，与其发表的核心论文（后来Kimi的模型架构和思路的基础）



杨植麟作为核心作者的重要论文	核心意义
Transformer-XL(ACL2019)	把Transformer从固定窗口扩展到可处理长序列，引入segment-level递归和相对位置编码。是现代长上下文技术的奠基性论文之一，也是kimi“Lossless Long Context”信条的来源
XLNet(NeurlPS2019)	基于排列语言建模(Permutation Language Modeling)的预训练范式，在18项NLP任务上刷新世界纪录，一度直接挑战BERT的统治地位。论文引用超过1万次

图表：公司名“月之暗面”取自英国摇滚乐队PinkFloyd1973年发行的专辑《The Dark Side of the Moon》是杨植麟最喜欢的专辑



《The Dark Side of the Moon》

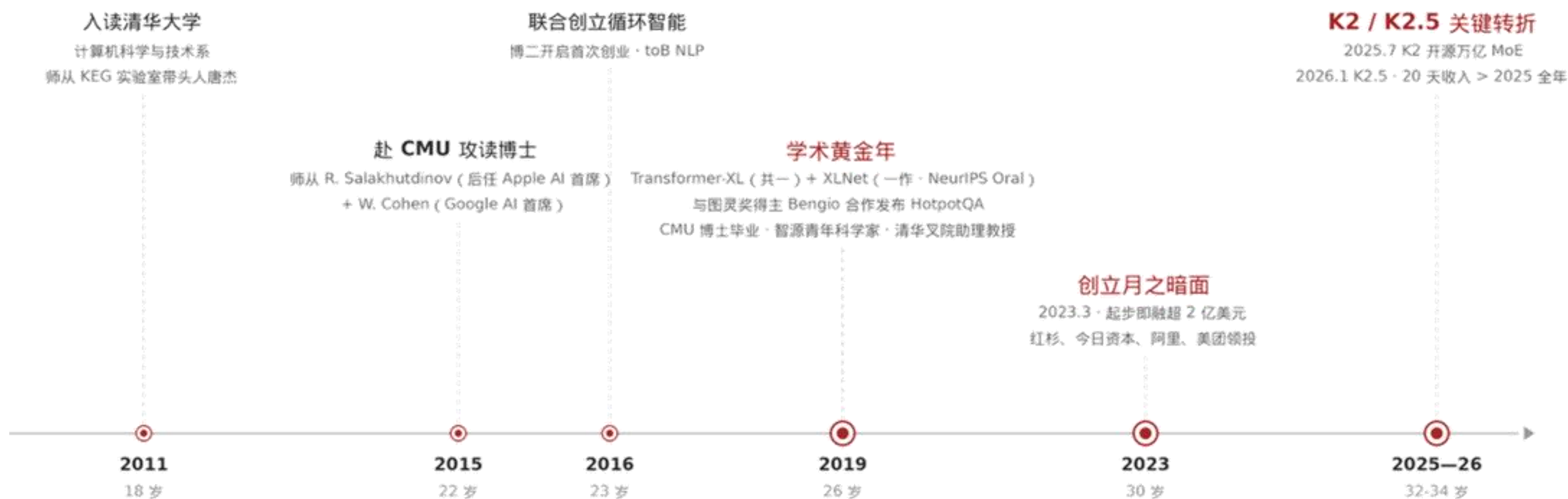
专辑封面



杨植麟：早年师从顶级AI学者，Kimi是其第二次创业

- ✓ 在清华与CMU期间师承顶级AI学者。清华大学计算机系本科（师从唐杰教授），后赴CMU攻读计算机博士，由Ruslan Salakhutdinov（现Apple AI首席研究员）与William Cohen（Google AI首席科学家）双导师指导。在校期间以一作发表XLNet（NeurIPS2019Oral），共同发表Transformer-XL，与图灵奖得主Yoshua Bengio合作发布Hotpot QA数据集；2019年取得博士学位，入选北京智源青年科学家、受聘清华交叉信息研究院助理教授。
- ✓ Kimi是杨植麟的第二次创业，首次创业为NLP方向。杨植麟并非首次创业，2016年博二期间即联合创立循环智能（Recurrent AI），做to B NLP（金融、保险对话分析），是国内最早商业化的对话NLP公司之一。2023年二次创业，创立了月之暗面。

图表：创始人杨植麟的个人成长经历：纯正计算机学术背景、产业经验丰富



组织架构：核心创始团队全清华系配置，决策执行快

- ✓ 核心创始团队：杨植麟及联合创始人张宇韬/周昕宇/吴育昕均出身清华计算机系，组织摩擦较小，团体作战能力强，团队算法等技术功底强，每人都有一篇引用1000次以上的论文。
- ✓ 2024年初公司（A+轮前后）约100人，之后也没有特别大幅的扩张，保持精干的组织矩阵。据Bismarck Analysis在2025年9月的报告，月之暗面全员扩张到约300人（含工程、产品、海外业务、销售）。公司的核心R&D团队始终保持小团队+高人效的组织风格，杨植麟多次明确表态过"组织不能太大"。

图表：Kimi的核心初创技术团队，全部出身清华计算机系



图表：Kimi团队主要成员履历

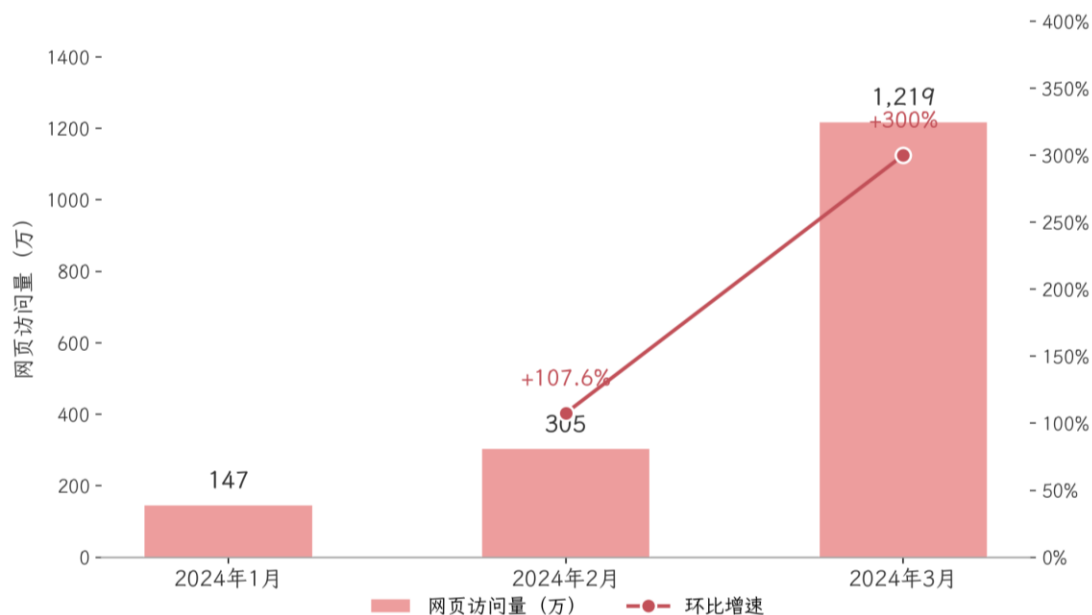
姓名	定位	年龄	介绍
杨植麟 (CEO&核心创始人)	战略及技术方向	34	清华大学本科，卡内基梅隆大学博士。深耕自然语言处理领域，曾在Google和Meta Platforms的人工智能实验室工作，是国内35岁以下NLP领域最具影响力的学者之一。
周昕宇 (联合创始人)	算法	-	清华大学本科，卡内基梅隆大学深造，专注自然语言处理与深度学习方向。早期从事AI基础研究，后加入字节跳动参与AI研发。
吴育昕 (联合创始人)	核心技术研发	-	毕业于清华大学及卡内基梅隆大学。曾在Meta AI研究院工作，主导开发开源目标检测框架Detectron2（行业主流项目之一），在计算机视觉领域具备国际领先技术能力。
张宇韬 (联合创始人)	CTO	35	清华大学计算机。曾参与Aminer等科研数据平台建设，在循环智能担任联合创始人兼CTO
张予彤 (联合创始人)	战略规划、商业化推进及融资	-	曾任金沙江创投主管合伙人，具有投资及产业资源。2025年任公司总裁，推动模型开源、Agent生态构建及算力优化的战略落地。

资料来源：企查查，新浪财经，新智元，极客公园，钛媒体等。
注：上述股权比例是截至2024年

Kimi初破圈：24年初，中文首个百万级别上下文Chatbot，访问量曾登顶

- ✓ 2023年10月，Kimi首次发布就主打全球**最长20万汉字无损上下文**。当时国内仍处在“百模大战”初期，Kimi凭借这一特点迅速引发关注，2024年2月，Kimi访问量月环比增长107.6%，位列当月全球AI产品增速榜第一。
- ✓ 2024年3月18日，Kimi宣布将长上下文文本升级到**200万字**，“**扔几篇论文/一本小说/几十份简历让Kimi读**”一度成为Kimi差异化的功能，体验一度好于ChatGPT，被认为是ChatGPT国内版平替，Kimi因为访问量过载而宕机2天。
- ✓ 据Similar Web的统计，2024年1月Kimi网页访问量不到150万，到3月已经超过1200万，稳居当时国产Chatbot的第一梯队，其当时的估值已上升到25亿美元。

图表：24年初Kimi网页访问量飙升，一度成为国产Chatbot第一名



图表：Kimi当时简约的首页设计+读文件，还有联动热点的运营能力，显示出kimi的独特性



营销投放一度激进：24年大幅投流，25年回归专注在模型与Infra

- 2024年4~10月是Kimi投放最激进的半年。当时Kimi、豆包、元宝并称AI三大投流大户。第三方数据显示，24年10月单月，这三家的投流总额超过30亿元人民币。24年10月，Kimi的MAU冲至3600万，位列前三。
- 2025年2月DeepSeek引爆后，Kimi宣布暂停投放，转为依赖自然增长。同时战略资源重新聚焦到两件事：①万亿参数K2基础模型的训练；②在医疗、法律等高壁垒垂直场景做API商业化。
- 我们认为第一阶段的Kimi在C端的尝试虽然没有大获全胜，但却在早期就显示出与其他国产模型气质上的不同——有技术基础的同时也具备对商业的敏锐度，落地并执行了移动互联网时代最经典运营策略，即大量投放换取C端用户。

图表：Kimi App端月活跃用户趋势——24年上半年激进投放，MAU快速增长。25年初暂停投放，转向K2模型的训练



Kimi历代模型迭代：K2是关键拐点，从此跃上1T参数模型俱乐部

- **K2是首个1T参数的开源国产模型**：25年7月发布的Kimi-K2，是首个公开达万亿参数规模的国产模型。截至今年1月K2.5发布时，其他主要披露参数的国产模型（GLM-5745B、Qwen3.5397B、MiniMaxM2.5230B）均未达到万亿。
- **K2.5是K系列首个原生多模态旗舰**：26年1月发布的Kimi-K2.5在维持万亿参数规模的同时引入自研视觉编码器MoonViT-3D，实现纯语言到原生多模态的架构跨越。能力重心向代码与Agentic方向集中发力。SWE-bench Verified得分76.8%，超越同期主流闭源模型，标志着Kimi从“规模突破”转向“能力落地”。

维度	K1 (moonshot-v1)	K1.5 (新RL模型)	K2 (关键转折)	K2.5 (旗舰基座模型)	K2.6 (Agentic专项迭代版)
发布时间	2023年11月（产品上线） 2024年初（API开放）	2025年1月22日	2025年7月11日	2026年1月27日	2026年4月20日
能力重心	长文档理解 文件处理	数学推理、代码、多模态推理，对标24年9月发布的OpenAI o1。	Agentic、代码、数学	Agentic+原生多模态推理	长周期Agentic编码
开源/闭源	闭源	闭源（发布技术报告）	开源✓	开源✓	开源✓
总参数量	未披露	未披露	~1T	~1T	~1T
激活参数量	未披露	未披露	32B	32B	32B
输入定价 (\$/百万tokens)	\$0.2	未独立发布API定价	\$0.6	\$0.6	\$0.95
输出定价 (\$/百万tokens)	\$2.0	未独立发布API定价	\$2.5	\$3.0	\$4.0
多模态	X	X	X	✓原生多模态（理解）	✓原生多模态（理解）
训练数据规模	未披露	未披露	15.5Ttokens	15Ttokens（含视觉）	未披露
模型架构	Dense Transformer	Dense Transformer	MoE（混合专家）	MoE+MLA注意力	MoE+MLA注意力
专家配置（MoE）	—	—	384专家/8激活+1共享	384专家/8激活，61层	384专家/8激活，61层
训练架构亮点	Lossless Long Context 无损长文本处理	长上下文RL扩展 （128K窗口）	MuonClip优化器 （零loss spike）	视觉-文本原生融合 （预训练起联合训练）	Agent Swarm扩展 （300子智能体）

资料来源：K1.5技术报告，KimiK2技术报告

1

Kimi长上下文基因与早期C端探索

2

K2/2.5的转折为何重要：研发和工程化的差异优势

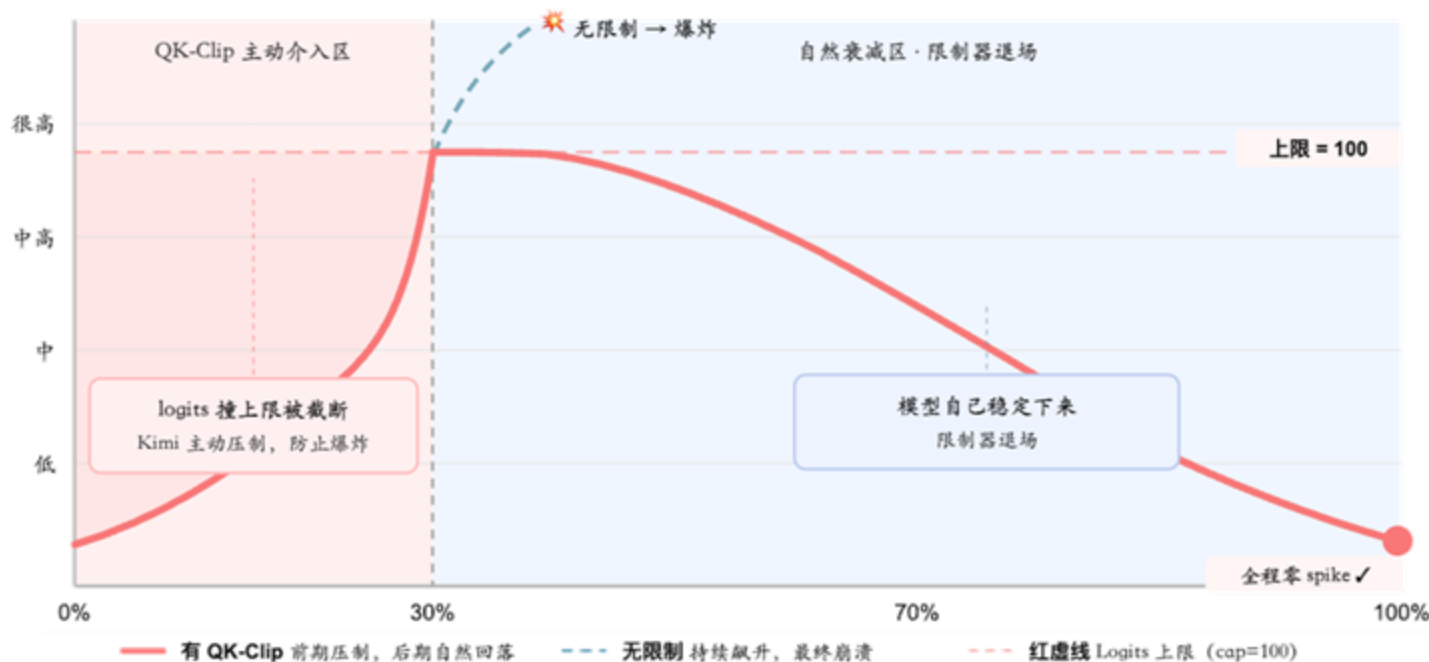
3

商业化与估值走到了哪一步？

K2为何重要——早期就认知到Scaling Law的上限很高，迈向1T参数

- 我们认为大模型行业对Scaling Law的认知经历过波折，尤其是2024年开始市场担心大模型智能水平和Scaling Law接近上限，部分公司转而走向小而美的性价比模型路线。而Kimi坚信Scaling Law，K2是首个1T参数的开源国产模型。
- 作为一家创业公司，训练1T量级参数的大模型需要非常强的工程化能力去压缩Logit，K2的MuonClip优化器引入了QK-Clip机制，是目前公开文献万亿参数级MoE模型训练稳定性的优秀记录。
- 在训练前期对这组数值设置硬性上限，强制压制其过度增长；待模型进入稳定阶段后，该约束自然失效，不干预后续训练。最终，K2在15.5万亿tokens的超长训练周期内实现了全程零loss spike。官方博客称，相比Adam W，MuonClip在大规模LLM训练上的token效率约为2倍。

图表：K2的Muon Clip优化器引入了QK-Clip机制，极大压缩Logit，提升了万亿参数模型的训练效率

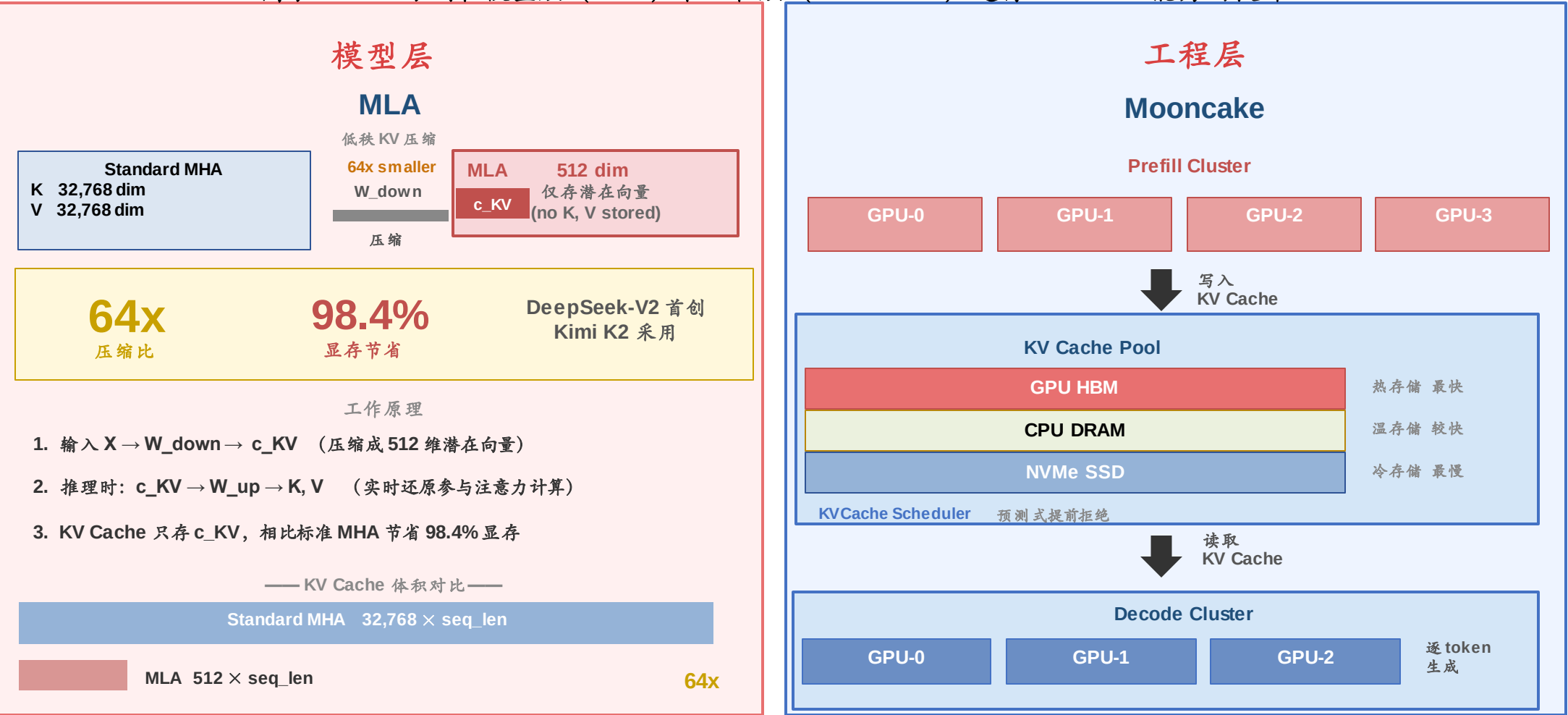


推理侧优化：KV Cache高于同业，通过模型层的MLA和工程层的Mooncake

Kimi 在模型层（MLA）与工程层（Mooncake）共同优化 KV Cache，推理效率实现跃升。模型层采用 MLA 将显存压缩 64×（节省 98.4%），工程层 Mooncake 三级分层调度使吞吐量提升 525%，KV Cache 显著高于同业。

- ✓ 模型层，Kimi K2 采用 MLA 架构，将 KV Cache 压缩至 512 维潜在向量（vs 标准 MHA 的 32,768 维），压缩比 64×、显存节省 98.4%，长上下文场景推理成本大幅下降。
- ✓ 工程层，Kimi 自研 Mooncake 平台以 GPU HBM / CPU DRAM / NVMe SSD 三级分层调度 KV Cache，配合预测式提前拒绝机制，系统吞吐 +525%、请求处理 +75%。

图表：Kimi 分别在模型层（MLA）和工程层（Mooncake）进行 KV Cache 能力的提升

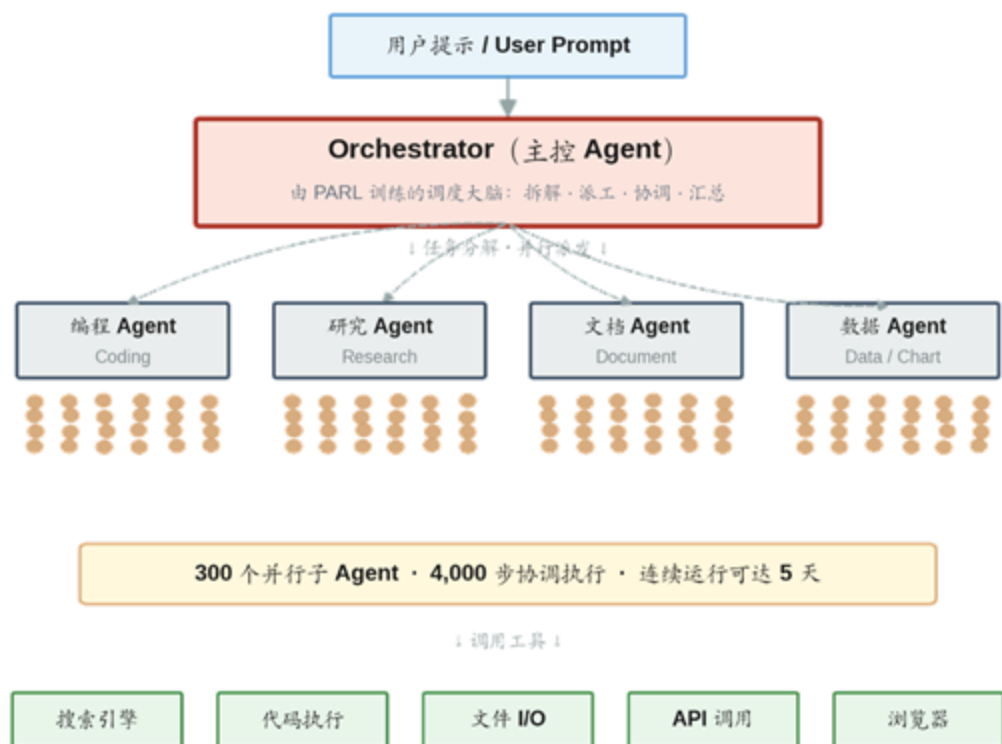


资料来源：K1.5 技术报告，KimiK2 在

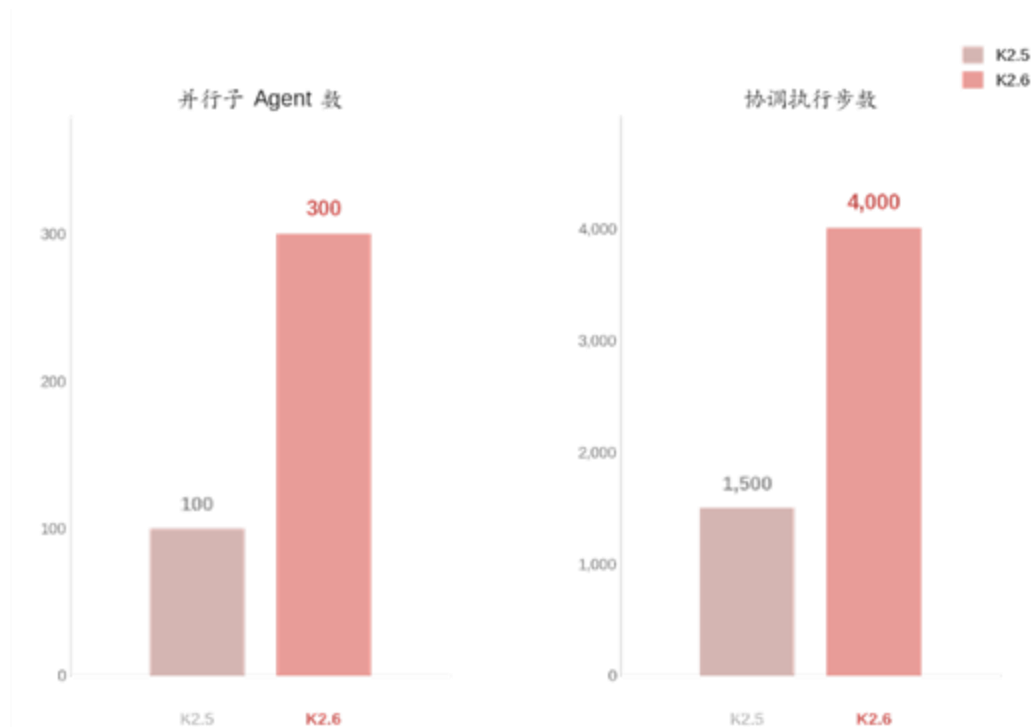
K2.5/2.6: 独特Agent Swarm架构，子Agent与协调步数大幅增长200%

- ✓ **Agent 集群 (Agent Swarm)**，模型自动把复杂任务拆成多个子Agent 并行处理、再由“主控Agent”汇总——K2.6 单次会话最多调度 300 个子Agent、协调 4,000 步，相比 K2.5 的 100 个子Agent / 1,500 步，分别增长 200% / 167%。
- ✓ 根据Browse Comp (OpenAI 2025 年 4 月发布的浏览/深搜能力评测)，K2.6 启用 Agent Swarm 模式 (多子Agent 并行检索) 得 86.3 分，显著比 GPT-5.4 单 Agent 模式的 78.4 分高，验证 Agent Swarm 架构在结构化检索任务上有效。但与 GPT-5.5 Pro 的 90.1 分仍有小幅差距。

图表：Kimi的Agent Swarm架构



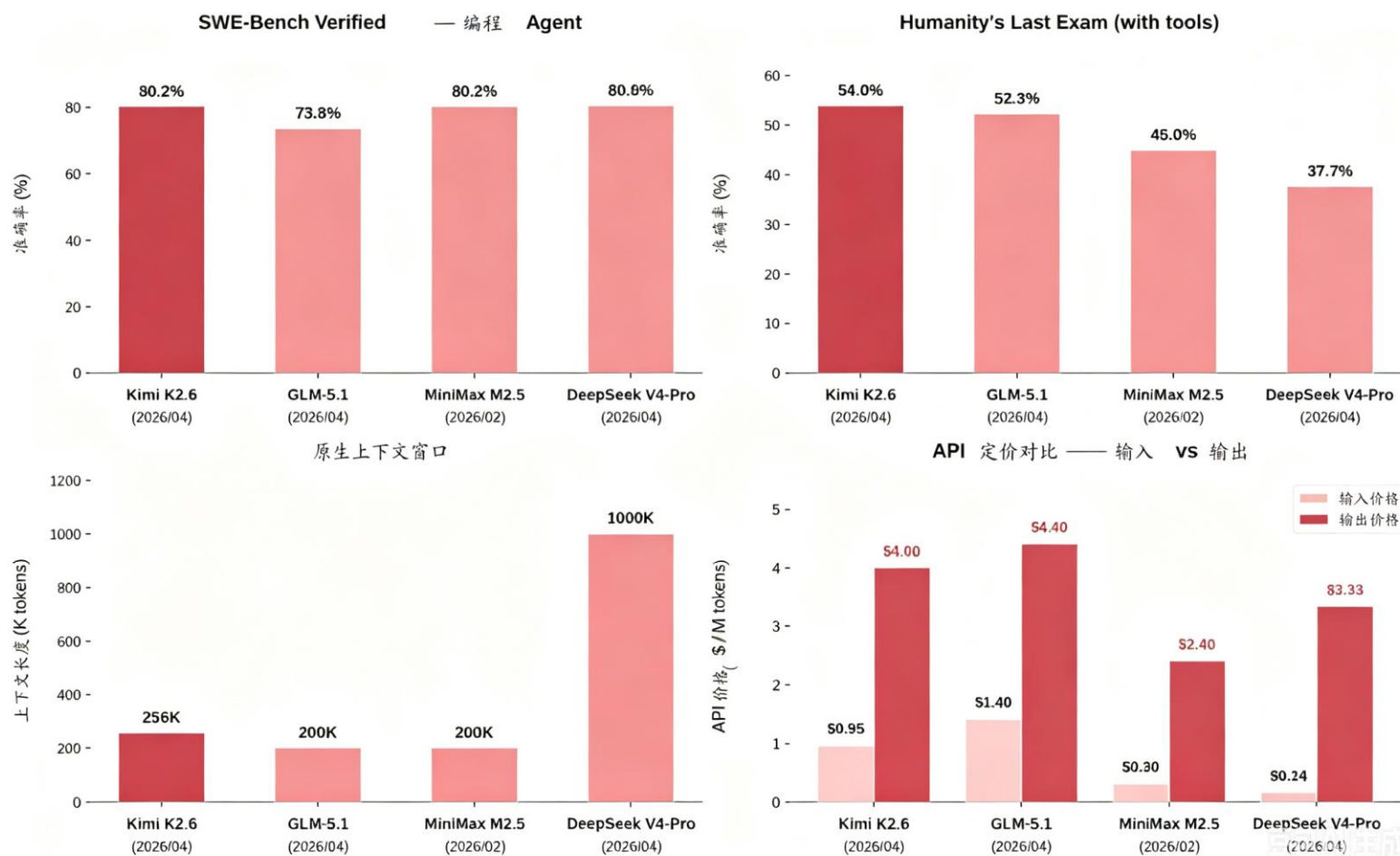
图表：Kimi的K2.6模型，并行子Agent数量和协调执行步数增长~200%



对比国产模型：定价与智谱在同一梯队，Agent与编程能力领先（1）

- ✓ **API定价相对较高**，与智谱定价位于同一梯队。输入\$0.95/输出\$4.00/每百万tokens，是DeepSeekV4-Pro（\$0.14输入）的7倍、MiniMaxM2.5（\$2.40输出）的1.7倍。对标海外，仍比GPT-5.2、ClaudeSonnet4.6便宜1/2至1/4。

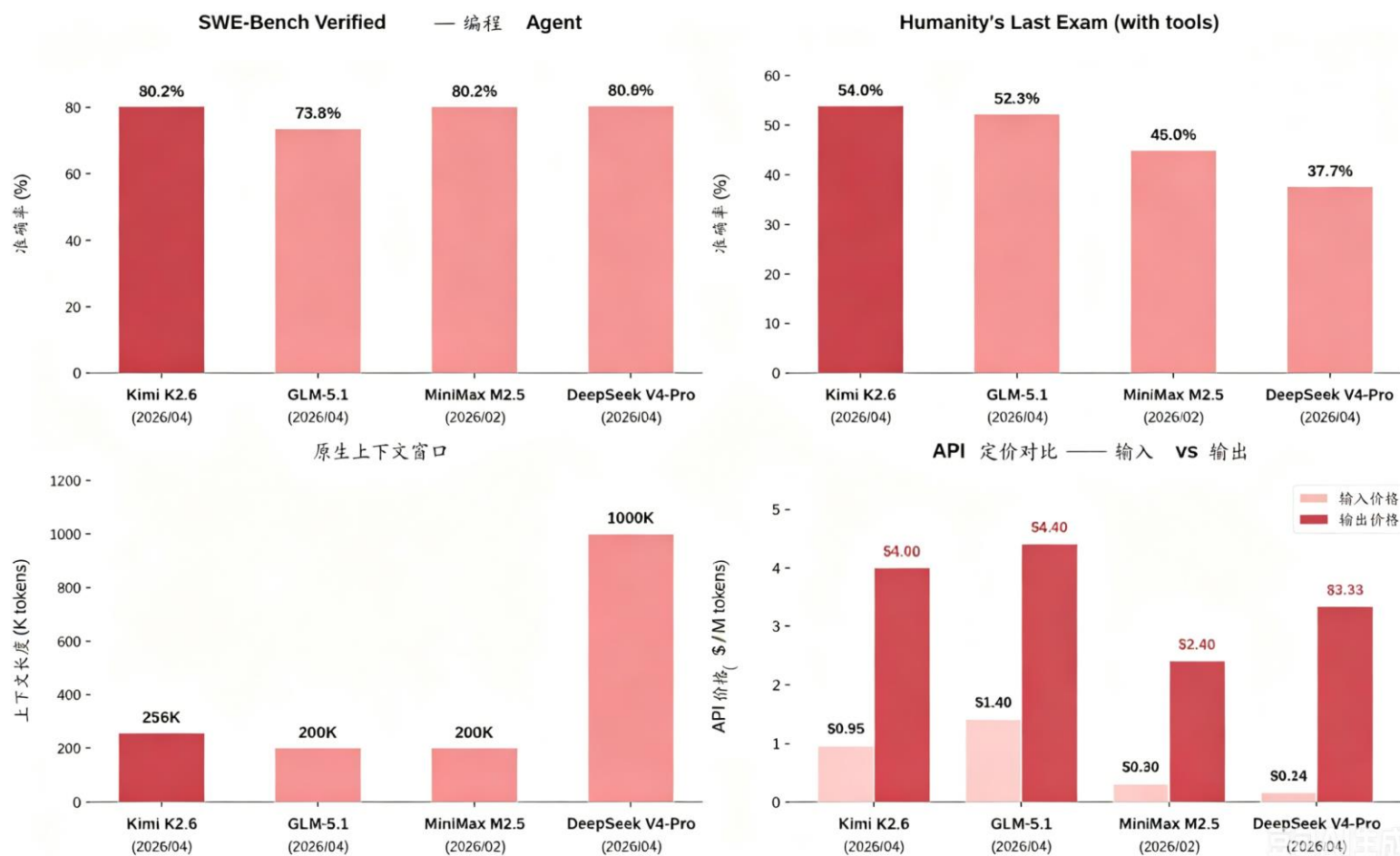
图表：K2.6模型与其他国产模型最新版本的对比



对比国产模型：定价与智谱在同一梯队，Agent与编程能力领先（2）

- ✓ **推理/工具调用/编程能力领先。** HLE衡量推理深度与工具调用，K2.6的分数为54.0%，国产唯一破50%。分别领先GLM-5.1、MiniMaxM2.5、DeepSeekV4-Pro1.7/9.0/16.3pp；SWE-Bench Verified衡量真实编程实战，分数为80.2%，与MiniMax并列、仅落后DeepSeekV4-Pro0.4pp，位列第一梯队。

图表：K2.6模型与其他国产模型最新版本的对比



1

Kimi长上下文基因与早期C端探索

2

K2/2.5的转折为何重要：研发和工程化的差异优势

3

商业化与估值走到了哪一步？

商业化：核心变现方式是API调用和会员订阅

Kimi的核心变现方式为API调用和会员订阅。

- ✓ 1、API调用：主要面向B端和P端（专业用户），可用于编程、垂类场景任务交付（Agent）等场景，定价基于消耗的token量。
- ✓ 2、会员订阅：会员体系将各功能权益整合进统一额度，额度按月刷新，以百分比形式呈现。核心权益包括Agent额度、Kimi Code、Kimi Chat、Kimi Claw及专业数据库、Agent集群。这些权益被整合进统一额度池，额度消耗取决于处理任务的token消耗。

图表：月之暗面API定价

K2.5		K2 0905		K2 Thinking	
kimi-k2.5 是 Kimi 最智能的模型，支持视觉与文本输入、思考与非思考模式、对话与 Agent 任务		kimi-k2 是一款具备超强代码和 Agent 能力的 MoE 架构基础模型，总参数 1T，激活参数 32B		kimi-k2-thinking 是具有通用 Agentic 能力和推理能力的思考模型，它擅长深度推理	
缓存命中	¥0.70 / MTok	缓存命中	¥ 1.00 / MTok	缓存命中	¥ 1.00 / MTok
输入	¥4.00 / MTok	输入	¥ 4.00 / MTok	输入	¥ 4.00 / MTok
输出	¥21.00 / MTok	输出	¥ 16.00 / MTok	输出	¥ 16.00 / MTok

商业化：订阅分为四层套餐，包含Code/Chat/Claw/Agent等权益

- 会员订阅面向C端和企业端提供不同的方案：
- 1) C端会员体系分4档，不同档位对应不同权益，更高阶的订阅版本每月的用量更多，所能使用的服务更全，包括专业数据库、Agent集群等。
 - 2) 企业版提供更高等级的数据安全保障和偏定制化服务，支持批量购买会员坐席（需购买5个及以上坐席）。
- 定价与国内大模型公司的Coding Plan接近。

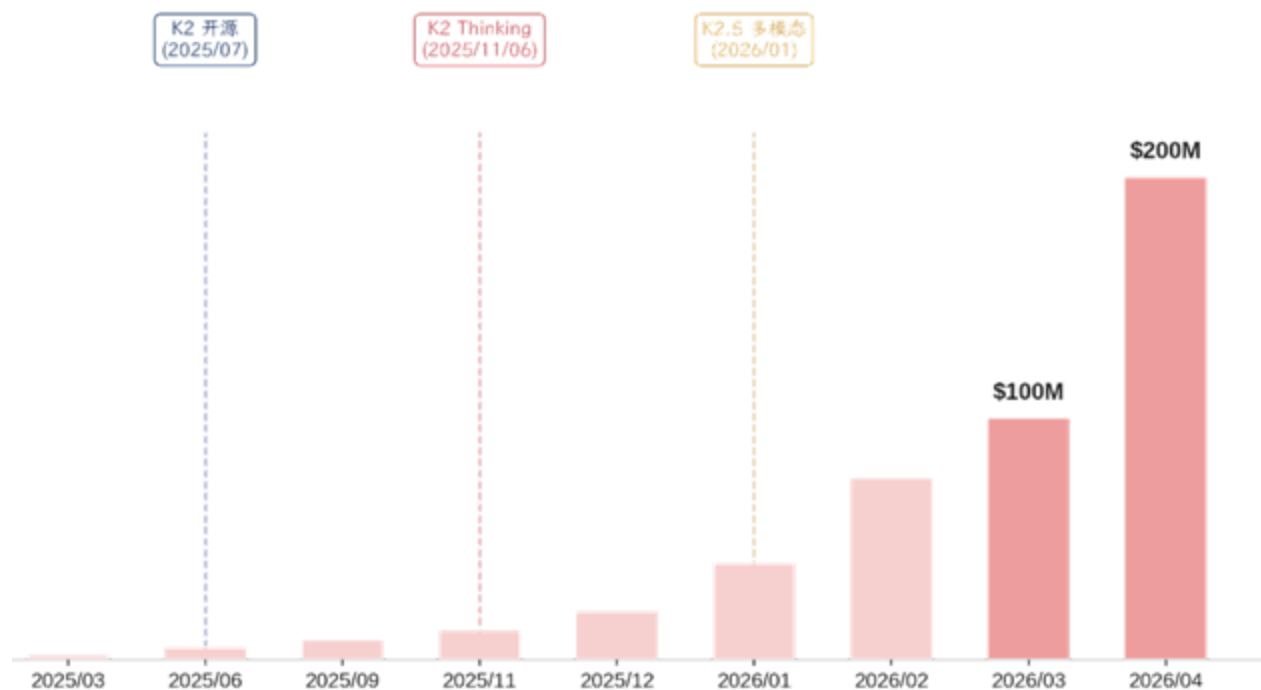
图表：月之暗面订阅定价

订阅版本	定价 (元/月)	权益	备注
Andante (日常使用)	39	1、Agent额度：更多（包括Office文件处理、深度研究、网站部署等） 2、Kimi Code：可调用 3、Kimi Chat 4、实验性功能：专业数据库；制作并发布带数据库的网站	1、kimi升级了会员权益体系，将原本分散在各功能中的权益整合为统一的额度池，会员自行分配使用； 2、额度消耗主要取决于处理任务的token消耗量； 3、额度以百分比形式呈现； 4、月度及年度会员的额度均按月刷新。
Moderato (效率升级；官方推荐)	79	1、Agent额度：2倍 2、Kimi Code：4倍额度 3、Kimi Chat 4、实验性功能：专业数据库；制作并发布带数据库的网站	
Allegretto (专业优选)	159	1、Agent额度：4倍 2、Kimi Code：20倍额度 3、Agent多任务并行 4、Kimi Chat 5、实验性功能：专业数据库；制作并发布带数据库的网站；Kimi Claw；Agent集群	
Allegro (全能尊享)	559	1、Agent额度：10倍 2、Kimi Code：60倍额度 3、Agent多任务并行 4、Kimi Chat 5、实验性功能：专业数据库；制作并发布带数据库的网站；Kimi Claw；Agent集群	

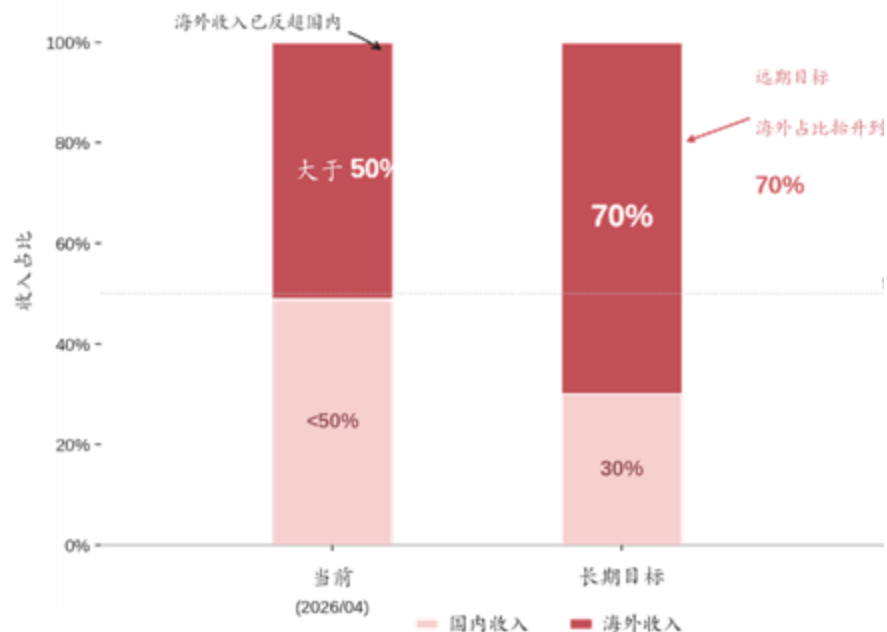
ARR快步提升，海外市场突破中

- 2026年以来Kimi的ARR快速增长，最新ARR预估已经超过2亿美金。在2025年7月K2开源之前，Kimi的商业化相对平缓。在K2.5发布后的26年3月份，ARR突破1亿美金，4月ARR进一步翻倍至2亿美金。
- 海外市场突破中。据杨植麟在2026年4月的内部信和投资人沟通：过去一年，Kimi海外及国内付费用户数月度环比增长超过170%，海外大模型API收入较9-11月实现4倍增长。

图表：Kimi历史以来的ARR情况（美元）



图表：目前kimi海外收入超过50%，中长期目标是70%

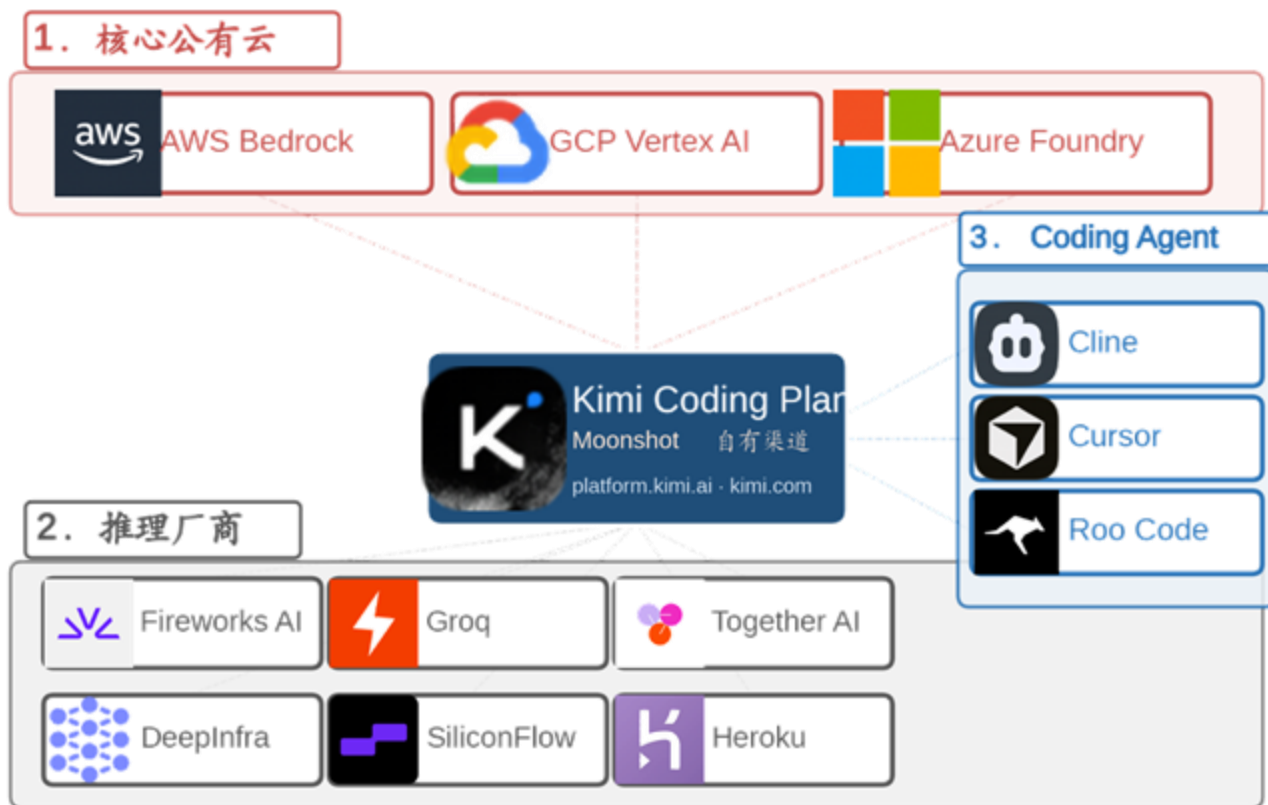


海外渠道：核心公有云、推理优化平台、Coding Agent平台

除自有平台外，Kimi在海外还布局了三类外部渠道，触达不同客户：

- ✓ **主流公有云，B端企业客户的规模化触达：**已入驻AWS Bedrock、GCP Vertex AI、Azure Foundry，借助云厂商的企业客户网络，实现对中大型企业的规模化触达。
- ✓ **推理优化平台，对延迟和成本敏感的规模化开发者：**接入Groq、Together AI、Fireworks AI、SiliconFlow等主流推理平台，可以覆盖对延迟/成本敏感的开发群体。
- ✓ **Coding Agent，日常普通个人开发者：**集成Cursor、Cline、Roo Code等AI编程工具，直接嵌入开发者日常工作流，获取高频高粘性用户。

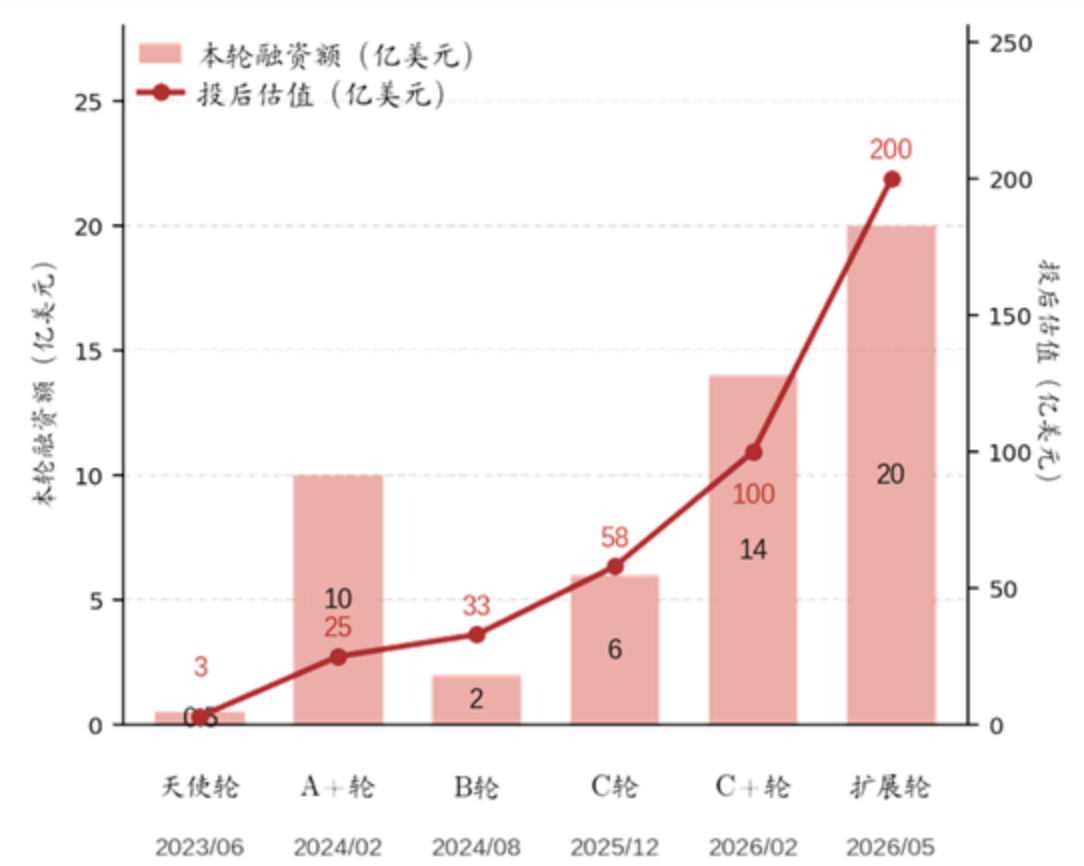
图表：Kimi在海外的渠道主要是核心公有云/推理厂商/coding agent



融资与估值：国产模型累计融资额TOP，最新估值200亿美金

- **Kimi**累积融资额超过**50亿美金**，估值攀升至**200亿美金**。2026年5月，Kimi完成20亿美元扩展轮融资，投后估值达200亿美元，刷新国产大模型单轮融资纪录。股东汇聚了阿里、腾讯、美团、红杉、IDG等核心产业资本。
- 横向来看，国产大模型整体均获得一二级资本的大力支持。国产大模型中，目前Kimi累计融资额位于第一梯队，累计融资额超过50亿美金。除此之外，阶跃星辰累积融资约25亿美元、MiniMax与智谱AI已相继登陆港交所完成二级市场融资。

图表：月之暗面融资历程与估值情况



图表：国产模型累计融资额统计

公司名称	最新一轮时间	累计融资额	核心股东/主要产业投资人
月之暗面 (Kimi)	2026/05 (扩展轮)	~52亿美元	阿里巴巴、腾讯、美团、高瓴创投、IDG
阶跃星辰	2026/05 (新轮进行中)	~25亿美元	上海国资、中国人寿股权、华勤技术、龙旗科技
MiniMax	2026/01 (港股IPO)	~23亿美元 (含IPO)	阿里巴巴 (12.5%)、米哈游 (5.9%)、腾讯 (2.4%)、红杉中国
智谱AI	2026/01 (港股IPO)	~18亿美元 (含IPO)	美团、蚂蚁集团、腾讯、小米、顺为资本、北京/地方国资
零一万物	2024年 (B轮)	~10亿美元	阿里云、创新工场 (李开复)
百川智能	2024/07 (A轮)	~7亿美元	阿里巴巴、腾讯、小米、北上深国资产业基金
DeepSeek	暂无外部融资	—	幻方科技 (母公司) 自有资金

风险提示

- 版权保护力度不及预期，知识产权未划分明确的风险，消费习惯难以改变的风险，关联公司治理风险，生成式AI技术发展不及预期的风险，产品研发难度大的风险，产品上线延期的风险，营销买量成本上升风险，人力成本上升的风险，政策监管的风险，商业化能力不及预期的风险，AI产品和服务的用户增长和付费率不及预期的风险，AI产品和服务降本提效效果不及预期的风险，新客户拓展不及预期的风险，生成式AI带来的数据安全风险，全球经济环境恶化风险，国内外互联网行业反垄断及监管风险，地缘政治导致客户出海困难的风险，竞争加剧风险，头部广告主预算缩减风险，品牌客户拓展不及预期的风险，人才流失风险。

分析师介绍

杨艾莉：中信建投证券研究发展部全球AI创新及互联网首席分析师，中国人民大学传播学硕士，曾任职于百度、新浪，担任商业分析师、战略分析师。2015年起，分别任职于中银国际证券、广发证券，担任传媒与互联网分析师、资深分析师。2019年4月加入中信建投证券研究发展部担任传媒互联网首席分析师。曾荣获2024年新浪财经金麒麟评选传媒行业菁英分析师第一名；2024年21世纪金牌分析师评选传媒互联网行业第三名；2024年金牌分析师评选传媒行业第四名；2024年证券时报·新财富杂志最佳分析师评选入围。

杨晓玮：中信建投证券研究发展部全球AI创新及互联网联席首席分析师，上海交通大学金融硕士，主要研究大模型、AI应用、互联网平台、游戏内容及潮流玩具等。

马晓婷：中信建投证券研究发展部全球AI创新及互联网分析师，上海财经大学学士，上海社会科学院硕士，主要研究广告营销、AI应用及其相关标的。

评级说明

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

分析师声明

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

法律主体说明

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

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

一般性声明

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

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

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

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

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

中信建投证券研究发展部

北京
朝阳区景辉街16号院1号楼18层

电话：(8610) 56135088
联系人：李祉瑶
邮箱：lizhiyao@csc.com.cn

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

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

中信建投（国际）

香港
中环交易广场2期18楼

电话：(852) 3465-5600
联系人：刘泓麟
邮箱：charleneliu@csci.hk