



计算机行业研究

买入（维持评级）

行业周报
证券研究报告

计算机组

分析师：李可夫（执业 S1130525120009） 分析师：刘高畅（执业 S1130525120005）
 likefu@gjzq.com.cn liugaochang@gjzq.com.cn

国产模型持续闪耀

Kimi K3 发布，多重架构实现底层创新

2026 年 7 月 16 日月之暗面在 WAIC 前夕推出 Kimi K3，是全球首个开源的近 3 万亿参数大模型，拥有 2.8 万亿参数、100 万 token 超长上下文，原生支持视觉理解，完整权重将于 7 月 27 日前公开。模型搭载多项原创架构技术：自研 KDA 混合线性注意力可大幅降低长文本缓存消耗、提升长序列解码速度；Block 注意力残差以极小算力开销优化训练效率；配合 896 专家仅激活 16 个的高稀疏 MoE、MXFP4/MXFP8 量化感知训练，整体扩展效率相较 K2 提升 2.5 倍，攻克万亿参数模型稳定训练的技术难点。

综合性能跻身全球第一梯队，大幅收窄与闭源顶尖模型差距

Kimi K3 综合实力显著拉近与 GPT-5.6 Sol、Claude Fable 5 等海外闭源旗舰的差距，改写此前“开源落后闭源半年”的行业现状。在编程、通用智能体、视觉智能体各类权威基准测试中多项成绩名列前茅，Arena.AI 前端开发 WebDev 榜单直接超越 Claude Fable 5 登顶；仅整体综合上限仍小幅不及海外头部闭源模型，国产开源大模型正式迈入全球前沿阵营。

定价迈入高端前沿区间，商业化与估值同步高速增长

Kimi K3 API 定价对标海外顶级前沿模型，定价标准为每百万 token 输入缓存命中 2 元、未命中 20 元，输出 100 元；依托 Mooncake 分离式推理架构，编程场景缓存命中率超 90%，大幅降低用户实际调用成本。商业化层面，截至 6 月中旬月之暗面 ARR 突破 3 亿美元，API 收入占总收入七成以上；企业新一轮融资投前估值达 315 亿美元。同期国内 DeepSeek 也启动大额融资、商业化收入持续走高，行业形成技术迭代、API 增收、估值抬升的正向循环，国产头部大模型正式从低价获客转向靠硬核能力兑现价值。

投资建议

相关标的：

海外算力：东山精密、工业富联、胜宏科技、江海股份、中际旭创、中钨高新、蓝思科技、东阳光、火炬电子、三环集团、欧科亿、天孚通信、鼎泰高科、新易盛、领益智造、兆易创新、鹏鼎控股、唯科科技、天岳先进、大普微、源杰科技、麦格米特、景旺电子、英维克、京东方等；英特尔、SK 海力士、Lumentum、闪迪、铠侠、美光。

国内算力：寒武纪、海光信息、东阳光、中芯国际、华虹半导体、禾盛新材、摩尔线程、天数智芯、沐曦股份、壁仞科技、协创数据、华丰科技、杰华特、京基智农、华达科技、中科曙光、利通电子、浪潮信息、利扬芯片、胜蓝股份、华勤技术、国科微、中国长城、晶科科技、罗曼股份、盈峰环境、芯原股份、晶科科技、亿田智能、豫能控股、星环科技、鸿日达、盛视科技、神州数码、润泽科技、大位科技、润建股份、奥飞数据、瑞晟智能、科华数据、潍柴重机、欧陆通、杰创智能、奥尼电子。

大模型&云：智谱、MiniMax、阿里巴巴、腾讯控股、金山云、百度集团、优刻得、首都在线、网宿科技、云赛智联、青云科技等。

风险提示

行业竞争加剧的风险；技术研发进度不及预期的风险；特定行业下游资本开支周期性波动的风险。



内容目录

Kimi K3 发布：全球首个开源三万亿级模型	3
1.1 WAIC 前夕发布，刷新开源模型规模上限	3
架构创新：KDA 与注意力残差，扩展效率较 K2 提升约 2.5 倍	3
2.1 KDA 混合线性注意力：面向百万 token 长序列	3
2.2 注意力残差：训练效率提升	3
2.3 高稀疏 MoE 与量化感知训练，支撑 2.8 万亿参数稳定训练	3
评测逼近闭源前沿，商业化步入 frontier 定价区间	4
3.1 基准表现：多项智能体与编程基准登顶，接近海外闭源旗舰	4
3.2 定价大幅提升，ARR 突破 3 亿美元	5
3.3 商业化与估值同步前进	5
风险提示	5

图表目录

图表 1：AttnRes 示意图	3
图表 2：Kimi K3 与闭源前沿模型官方基准对比	4
图表 3：Arena.AI WebDev 榜单	5



Kimi K3 发布：全球首个开源三万亿级模型

1.1 WAIC 前夕发布，刷新开源模型规模上限

据新华社报道，2026 世界人工智能大会暨人工智能全球治理高级别会议即将举行之际，月之暗面于 7 月 16 日发布新一代模型 Kimi K3，参数规模达 2.8 万亿，为目前全球参数最大的开源模型。Kimi K3 是月之暗面迄今能力最强的模型，基于 KDA 混合线性注意力（Kimi Delta Attention）与注意力残差构建，原生支持视觉理解，拥有 100 万 token 上下文窗口。

官方称 Kimi K3 完整模型权重将于 2026 年 7 月 27 日前发布，架构、训练与评测的进一步细节将随 Kimi K3 技术报告一同公布。

架构创新：KDA 与注意力残差，扩展效率较 K2 提升约 2.5 倍

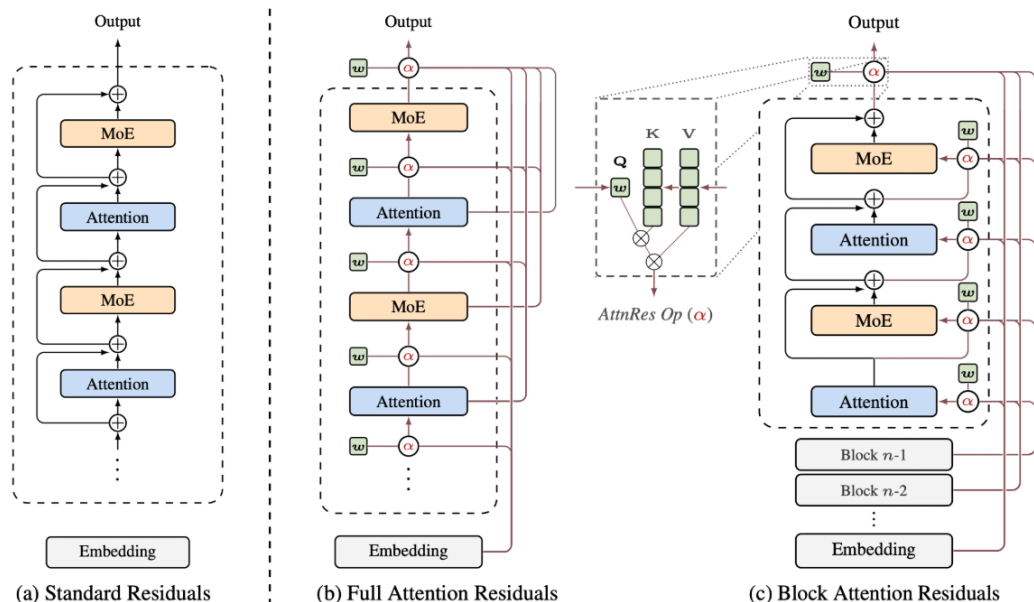
2.1 KDA 混合线性注意力：面向百万 token 长序列

Kimi Linear 是一种混合线性注意力架构，在各种场景下（包括长、短和强化学习（RL）扩展机制）均优于传统的全注意力方法。其核心是 Kimi Delta Attention (KDA)——Gated DeltaNet 的改进版本，引入了一种更高效的门控机制来优化有限状态 RNN 内存的使用。Kimi Linear 实现了卓越的性能、优异的硬件效率，尤其是在处理长上下文任务时。它可将大型键值缓存的需求降低高达 75%，并将 1m 上下文的解码吞吐量提升高达 6 倍。

2.2 注意力残差：训练效率提升

官方研究称，标准残差连接以固定权重累加所有层输出，随深度增加会稀释各层贡献并导致隐状态幅值无界增长；kimi 采用了替代品 Attention Residuals (AttnRes)。工程化版本 Block AttnRes 以约 8 个块即可恢复 Full AttnRes 的大部分收益，开销微小。Block AttnRes 可达到用 1.25 倍算力训练的基线模型的 loss 水平

图表1: AttnRes 示意图



来源：kimi 官方 github，国金证券研究所

2.3 高稀疏 MoE 与量化感知训练，支撑 2.8 万亿参数稳定训练

官方博客显示，K3 进一步扩大了 MoE 稀疏度，结合 Stable LatentMoE 框架后，模型可以在 896 个专家中高效激活 16 个。

叠加训练方法和数据配方的优化，这些结构性改进使 Kimi K3 相比 K2 的整体扩展效率提升约 2.5 倍，能更有效地把算力转化为能力。

Kimi K3 从 SFT 阶段开始应用量化感知训练，使用 MXFP4 权重和 MXFP8 激活函数以实



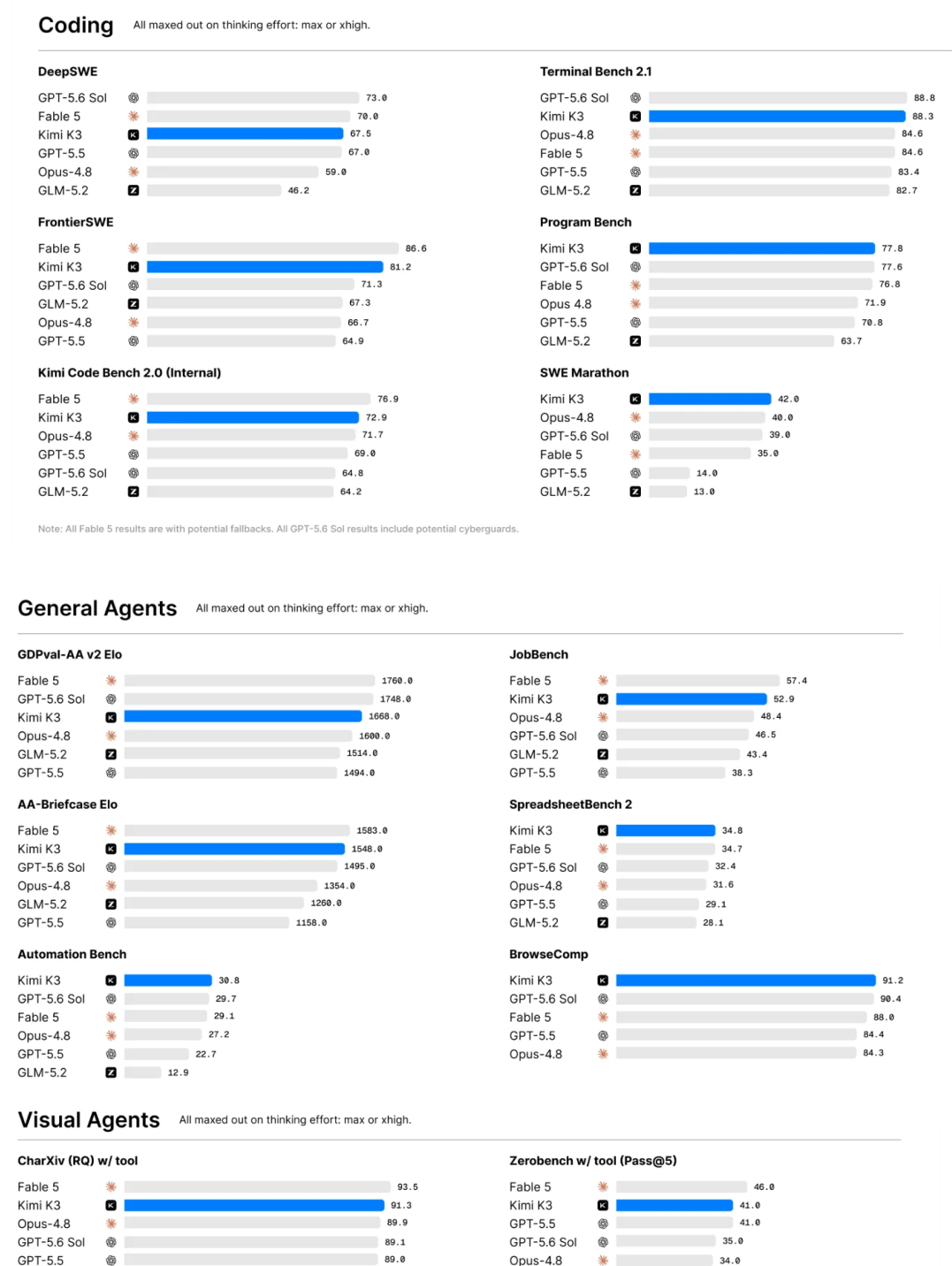
现广泛的硬件兼容性。为了防止专家不平衡导致大规模专家并行训练的吞吐量下降，引入了一种完全平衡的专家并行训练方法。

评测逼近闭源前沿，商业化步入 frontier 定价区间

3.1 基准表现：多项智能体与编程基准登顶，接近海外闭源旗舰

官方承认，Kimi K3 整体表现仍与最强的闭源模型 Claude Fable 5 和 GPT-5.6 Sol 存在差距。

图表2: Kimi K3 与闭源前沿模型官方基准对比



来源：月之暗面官方博客，国金证券研究所

在 Arena AI 代码竞技场的 WebDev 榜单上 Kimi K3 超越 Claude Fable 4 登顶榜首。



图表3: Arena. AI WebDev 榜单

Code Arena | WebDev Overall

View overall rankings across AI models on front-end web development tasks, including agentic coding workflows that require multi-step reasoning and tool use.

Jul 16, 2026 483,895 votes 99 models

Hide Filters Rank by Models Labs

Rank	Rank Spread	Model	Score	Votes	Price \$/M	Context
1	1 → 1	kimi-k3 Moonshot · Proprietary	1679 +17/-17 Preliminary	1,757	\$3 / \$15	1M
2	2 → 3	claude-fable-5 Anthropic · Proprietary	1631 +13/-13	2,505	\$10 / \$50	1M
3	2 → 3	gpt-5.6-sol-xhigh (codex-harness) OpenAI · Proprietary	1618 +13/-13	2,542	\$5 / \$30	1.1M
4	4 → 4	glm-5.2 (max) Zai · MIT	1587 +10/-10	4,722	\$1.40 / \$4.40	1M
5	5 → 9	claude-opus-4-8-thinking Anthropic · Proprietary	1562 +9/-9	7,309	\$5 / \$25	1M
6	5 → 11	grok-4.5 SpaceXAI · Proprietary	1558 +13/-13	2,214	\$2 / \$6	500K
7	5 → 10	claude-opus-4-7-thinking Anthropic · Proprietary	1558 +7/-7	10,534	\$5 / \$25	1M
8	5 → 11	claude-opus-4-7 Anthropic · Proprietary	1555 +7/-7	9,976	\$5 / \$25	1M
9	7 → 14	claude-opus-4-6-thinking Anthropic · Proprietary	1542 +6/-6	12,919	\$5 / \$25	1M
10	5 → 15	claude-sonnet-5-high Anthropic · Proprietary	1542 +12/-12	2,959	\$2 / \$10	1M
11	6 → 17	muse-spark-1.1 Meta · Proprietary	1538 +14/-14 Preliminary	2,083	\$1.25 / \$4.25	N/A

View all

来源: arena.ai 国金证券研究所

3.2 定价大幅提升, ARR 突破 3 亿美元

Kimi API 开放平台 kimi-k3 人民币定价为每百万 token 输入 2 元 (命中缓存)、20 元 (未命中缓存), 输出 100 元。

据《科创板日报》独家报道, 截至 6 月中旬, Kimi ARR (年度经常性收入) 突破 3 亿美元, 收入增长主要来自模型迭代带动的开发者使用和 API 收入提升。

我们认为, K3 定价较此前 Kimi 系列明显提升, 一方面反映 2.8 万亿参数规模以及长上下文、深度推理与复杂任务执行带来的算力成本, 另一方面也显示国产头部模型正从以性价比优势获客, 转向依靠模型能力提升实现 API 价值量增长。

3.3 商业化与估值同步前进

据《科创板日报》6 月 30 日独家报道, 月之暗面上一轮 200 亿美元估值融资已于近日完成交割, 新一轮融资已经启动, 投前估值涨至 315 亿美元。

据路透社 7 月 15 日报道, DeepSeek 计划以约 5000 亿元人民币 (约 740 亿美元) 估值启动新一轮融资, 融资规模最高 500 亿元人民币, 并筹备在内地上市; The Information 援引三名知情人士报道, DeepSeek 近期年化收入达到 4 亿至 5 亿美元, 主要来自企业和开发者付费调用其云端模型 API。

我们认为, 国内模型能力提升、ARR 增长与估值中枢齐头并进, 看好国内模型持续能力进步与商业化加速。

风险提示

行业竞争加剧的风险:

众多新兴玩家参与到市场竞争之中, 若市场竞争进一步加剧, 竞争优势偏弱的企业或面临出清, 某些中低端品类的毛利率或受到一定程度影响。

技术研发进度不及预期的风险:

技术开发需投入大量资源, 如果相关厂商新品研发进程不及预期, 表现层面将呈现出投入产出在较长时期的滞后特征。

特定行业下游资本开支周期性波动的风险:

部分公司下游资本开支波动与行业周期性相关性较强, 或在个别年份对于上游厂商的营收



表现产生扰动。

行业投资评级的说明：

买入：预期未来 3—6 个月内该行业上涨幅度超过大盘在 15%以上；

增持：预期未来 3—6 个月内该行业上涨幅度超过大盘在 5%—15%；

中性：预期未来 3—6 个月内该行业变动幅度相对大盘在 -5%—5%；

减持：预期未来 3—6 个月内该行业下跌幅度超过大盘在 5%以上。



特别声明：

国金证券股份有限公司经中国证券监督管理委员会批准，已具备证券投资咨询业务资格。

本报告版权归“国金证券股份有限公司”（以下简称“国金证券”）所有，未经事先书面授权，任何机构和个人均不得以任何方式对本报告的任何部分制作任何形式的复制、转发、转载、引用、修改、仿制、刊发，或以任何侵犯本公司版权的其他方式使用。经过书面授权的引用、刊发，需注明出处为“国金证券股份有限公司”，且不得对本报告进行任何有悖原意的删节和修改。

本报告的产生基于国金证券及其研究人员认为可信的公开资料或实地调研资料，但国金证券及其研究人员对这些信息的准确性和完整性不作任何保证。本报告反映撰写研究人员的不同设想、见解及分析方法，故本报告所载观点可能与其他类似研究报告的观点及市场实际情况不一致，国金证券不对使用本报告所包含的材料产生的任何直接或间接损失或与此有关的其他任何损失承担任何责任。且本报告中的资料、意见、预测均反映报告初次公开发布时的判断，在不作事先通知的情况下，可能会随时调整，亦可因使用不同假设和标准、采用不同观点和分析方法而与国金证券其它业务部门、单位或附属机构在制作类似的其他材料时所给出的意见不同或者相反。

本报告仅为参考之用，在任何地区均不应被视为买卖任何证券、金融工具的要约或要约邀请。本报告提及的任何证券或金融工具均可能含有重大的风险，可能不易变卖以及不适合所有投资者。本报告所提及的证券或金融工具的价格、价值及收益可能会受汇率影响而波动。过往的业绩并不能代表未来的表现。

客户应当考虑到国金证券存在可能影响本报告客观性的利益冲突，而不应视本报告为作出投资决策的唯一因素。证券研究报告是用于服务具备专业知识的投资者和投资顾问的专业产品，使用时必须经专业人士进行解读。国金证券建议获取报告人员应考虑本报告的任何意见或建议是否符合其特定状况，以及（若有必要）咨询独立投资顾问。报告本身、报告中的信息或所表达意见也不构成投资、法律、会计或税务的最终操作建议，国金证券不就报告中的内容对最终操作建议做出任何担保，在任何时候均不构成对任何人的个人推荐。

在法律允许的情况下，国金证券的关联机构可能会持有报告中涉及的公司所发行的证券并进行交易，并可能为这些公司正在提供或争取提供多种金融服务。

本报告并非意图发送、发布给在当地法律或监管规则下不允许向其发送、发布该研究报告的人员。国金证券并不因收件人收到本报告而视其为国金证券的客户。本报告对于收件人而言属高度机密，只有符合条件的收件人才能使用。根据《证券期货投资者适当性管理办法》，本报告仅供国金证券股份有限公司客户中风险评级高于 C3 级（含 C3 级）的投资者使用；本报告所包含的观点及建议并未考虑个别客户的特殊状况、目标或需要，不应被视为对特定客户关于特定证券或金融工具的建议或策略。对于本报告中提及的任何证券或金融工具，本报告的收件人须保持自身的独立判断。使用国金证券研究报告进行投资，遭受任何损失，国金证券不承担相关法律责任。

若国金证券以外的任何机构或个人发送本报告，则由该机构或个人为此发送行为承担全部责任。本报告不构成国金证券向发送本报告机构或个人的收件人提供投资建议，国金证券不为此承担任何责任。

此报告仅限于中国境内使用。国金证券版权所有，保留一切权利。

上海

电话：021-80234211

邮箱：researchsh@gjzq.com.cn

邮编：201204

地址：上海浦东新区芳甸路 1088 号

紫竹国际大厦 5 楼

北京

电话：010-85950438

邮箱：researchbj@gjzq.com.cn

邮编：100005

地址：北京市东城区建国内大街 26 号

新闻大厦 8 层南侧

深圳

电话：0755-86695353

邮箱：researchsz@gjzq.com.cn

邮编：518000

地址：深圳市福田区金田路 2028 号皇岗商务中心

18 楼 1806



【小程序】
国金证券研究服务



【公众号】
国金证券研究