



交銀國際
BOCOM International

行业剖析

互联网行业

2026年8月12日

从开源到前沿 Kimi K3 技术报告解析

基础模型行业系列(1)

“十五五”专题

分析师：郑民康

月之暗面旗下的Kimi K3在综合智能指数跻身全球前列，更重要的是，K3以低于闭源模型的推理成本实现了这个成绩。

K3的架构沿序列、深度、宽度三个维度扩展信息流，重塑成本与毛利率上限。同时，K3相对K2约2.5倍的整体扩展效率提升，是数据、scaling law研究、长上下文扩展三个维度数十个选择的叠加结果。

我们认为，Kimi K3的推出显著提升了市场对ARR增长的信心，月之暗面当前500亿美元的估值支撑，主要建立在其ARR有望在未来6-12个月达到约10亿美元的强劲增长逻辑之上。

2026 年 8 月 12 日

互联网行业-数据与智能

基础模型行业系列（1）：

从开源到前沿：Kimi K3 技术报告解析

郑民康

wallace.cheng@bocomgroup.com
(852) 3766 1810

- ④ **Kimi K3 - 具多项突破的 SOTA 开源模型：**月之暗面旗下的 Kimi K3，标志着开源大模型正式迈向 3 万亿级参数时代，并在原生多模态与百万级上下文 Agent 能力上达到前沿水平。在 Artificial Analysis 等第三方评测中，K3 的综合智能指数跻身全球前列，仅次于 Claude Fable 5/Opus 5 与 GPT-5.6 Sol 两大最强闭源模型。更重要的是 K3 以低于闭源模型的推理成本实现了这个成绩。对大模型行业而言，K3 再次拉近了开源与闭源的差距，国产大模型在全球第一梯队中的话语权随之提升。
- ④ **K3 模型三维架构重塑成本与毛利率上限：**K3 的架构沿序列、深度、宽度三个维度扩展信息流。**序列：**KDA 使显存占用不再随上下文线性膨胀，成本优势直接转化为单位经济模型与定价权。**深度：**AttnRes 消除逐层信息衰减，使同样 93 层产生更高有效利用率，贡献 2.5 倍 scaling efficiency 的一部分并加速模型迭代。**宽度：**Stable LatentMoE 以 896 专家中激活 16 个（稀疏度 56:1）实现知识容量与计算成本分离，以 1042 亿参数成本提供 2.8 亿参数能力。
- ④ **预训练：打破参数迷信，工程效率重塑成本。**我们认为在近 3 万亿参数规模上，真正的问题是如何用同样的钱买到更多智能。K3 相对 K2 约 2.5 倍的整体扩展效率提升，是数据、scaling law 研究、长上下文扩展三个维度的数十个选择的叠加结果。基础设施层面，自研 MoonEP 专家并行方案通过动态调度冗余专家，保证每块 GPU 在每步训练处理的 Token 数量绝对相等，消除了“快卡等慢卡”的算力气泡与闲置，提升集群利用率，大幅压缩训练成本。
- ④ **后训练：专家团队淬炼到训练场组成的护城河。**K3 的后训练遵循 SFT→RL→MOPD 三阶段流水线，三阶段的紧密耦合本身即构成工程壁垒的一部分。此外 K3 构建了统一白盒环境、自我演化的知识图谱出题工厂、以及七类高保真训练场。这部分是竞争对手最难复制的内容——不是“知道就能做到”，而是“做过才能做到”。
- ④ **Pre-IPO 融资估值反映了市场对 Kimi ARR 持续高增的预期。**近期月之暗面完成 F 轮融资 35 亿美元，投后估值较今年 5 月上升 75% 至 350 亿美元。以今年 6 月底 3 亿美元 ARR（年度经常性收入）测算，其对应 P/ARR 估值在 100 倍以上。目前市场预期月之暗面将在 8 月开始 Pre-IPO 轮融资，估值可能上升至 500 亿美元。我们认为，市场对 Kimi K3 的推出可以大幅提升 ARR 有高度信心，500 亿美元的估值应是建立在月之暗面的 ARR 在未来 6-12 个月内提升至约 10 亿美元的预期。我们认为在 2H26，公司是否可以及时补充算力、提高用户体验、在大模型迭代加快的情况下保持优势及 ARR，将是支撑更高估值的关键。对于潜在受益于月之暗面 K3 推出/估值提升的公司，可关注直接持股的招商局中国基金（133 HK/未评级），以及与月之暗面开展 Token 分成机制的中软国际（354 HK/未评级）。

此报告最后部分的分析师披露、商业关系披露和免责声明为报告的一部分，必须阅读。

下载本公司之研究报告，可从彭博搜寻 NH BCM 或 登录研究部网站 <https://research.bocomgroup.com>

更多免费行业报告

并购家

www.ipopo.net

目录

Kimi K3 - 具多项突破的 SOTA 开源模型.....4

模型架构深度解析 - 从技术看经济性.....7

注意力混合架构带来长上下文推理成本的压缩.....7

AttnRes：从“一维覆盖”到“二维检索”，信息不再随深度衰减.....10

Stable Latent MoE 专家分流成就万亿参数的稳定训练.....12

预训练层：打破参数迷信，工程效率重塑成本.....14

参数上 2.8 万亿是突破，但效率提升含金量更高.....14

MoonEP 系统带来算力成本实质性节约.....17

后训练：专家团队淬炼到训练场组成的护城河.....18

训练场与出题工厂：RL 真正的护城河.....19

Pre-IPO 融资估值反映对 Kimi ARR 持续高增的预期.....21

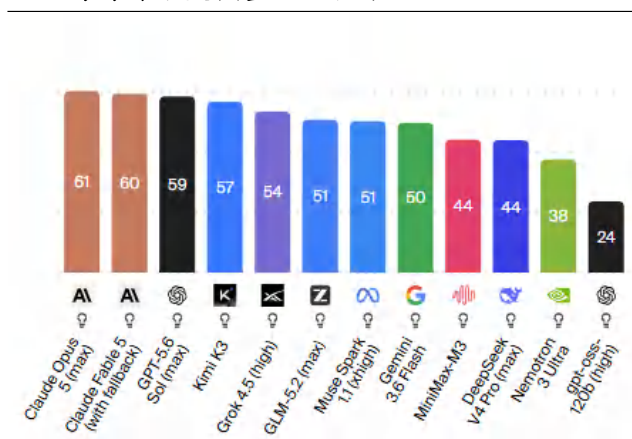
风险因素.....22

Kimi K3 - 具多项突破的 SOTA 开源模型

月之暗面旗下的基础模型 Kimi K3，标志着开源大模型正式迈向 3 万亿级参数时代，并在原生多模态与百万级上下文 Agent 能力上达到前沿水平。在 Artificial Analysis 等第三方评测中，K3 的综合智能指数跻身全球前列，仅次于 Claude Fable 5/Opus 5 与 GPT-5.6 Sol 两大领先闭源模型。更重要的是，K3 以远低于闭源模型的推理成本实现了这个成绩。我们认为 K3 的推出将加速 Agent 应用在企业端的落地。对大模型行业而言，K3 再次拉近了开源与闭源的差距，国产大模型在全球第一梯队中的话语权随之提升。K3 的优势包括：

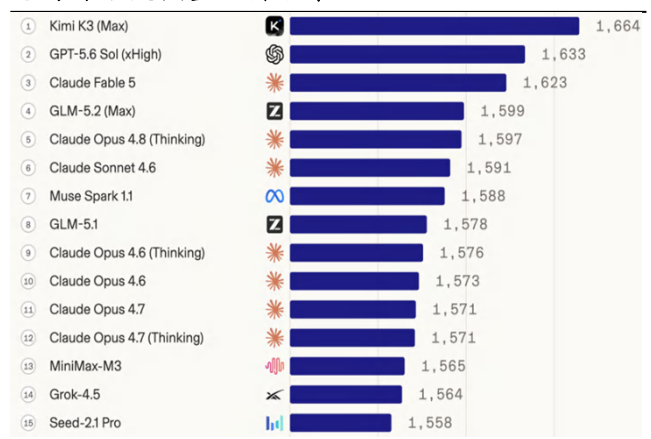
- ⊕ **前沿级的预训练底座：**K3 是一个总参数 2.8 万亿的原生多模态 MoE 模型，每个 Token 仅激活 1,042 亿参数，上下文窗口达 100 万 Token。KDA、AttnRes、Stable LatentMoE 三大架构创新，叠加改进的数据与训练配方，使整体扩展效率较上一代 Kimi K2 提升约 2.5 倍，换句话说，同样的算力预算，可以训练出明显更强的模型。
- ⊕ **跨多级推理强度的强化学习：**开发团队在通用任务、智能体（Agentic）和代码三大领域开展大规模强化学习，并为每个领域设置 low、high、max 三档推理强度，先训练出 9 个“领域专家”，再经多教师模式在线蒸馏，融合为一个统一的模型。成果是推理成本可按用户需要调节：简单问题可快速作答，深度思考后可完成复杂任务。
- ⊕ **支撑万亿参数与百万 Token 上下文的智能基础设施：**在架构层中，开发团队为新注意力机制 KDA 做了专门的算法—系统协同设计。同时，MoonEP 专家在预训练层的并行方案能在 896 个专家上实现完美负载均衡，内存高效训练方案令 2.8 万亿参数规模下算力利用率始终保持高位。至于强化学习层方面，同址部署（co-located）训练系统与可恢复沙盒，使百万 Token 级的长程 Agent 轨迹得以稳定、低成本地训练。

图表 1：Kimi K3 智能指数排名在国产模型中第一，与海外顶尖模型差距极小



资料来源：Artificial analysis，交银国际

图表 2：K3 在 Arena.ai WebDev 全栈开发超越其它海外顶尖模型，排名第一

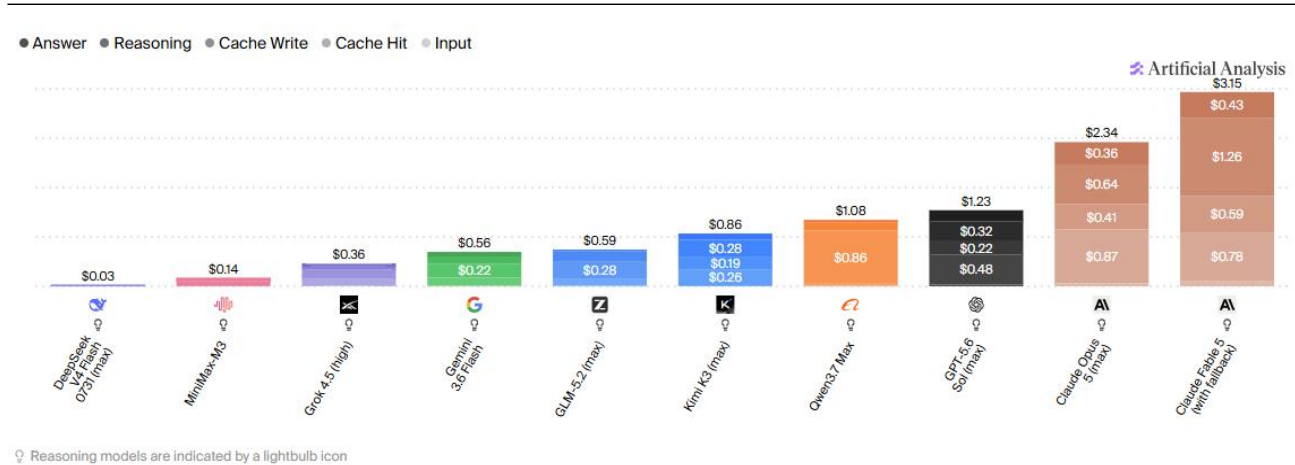


资料来源：Artificial analysis，交银国际

据 2026 年 8 月 3 日 Artificial Analysis 最新数据，Kimi K3（智能指数约 57、单任务成本 0.86 美元）是前十大 SOTA（State-of-the-Art）模型中较靠近最具吸引力象限的模型，且位于*帕累托前沿（Pareto Frontier）。K3 智能紧贴 GPT-5.6 Sol（智能指数约 59）与 Claude Opus 5/Fable 5（智能指数约 60），而测试任务所需的加权平均成本仅为 GPT-5.6 Sol（1.23 美元）的约 70%、Fable 5/Opus 5（3.15 美元）的 30-40%。其它开源同业的 GLM-5.2 与 DeepSeek V4 Flash 更便宜（0.59/0.03 美元），但智能指数上有所差距。

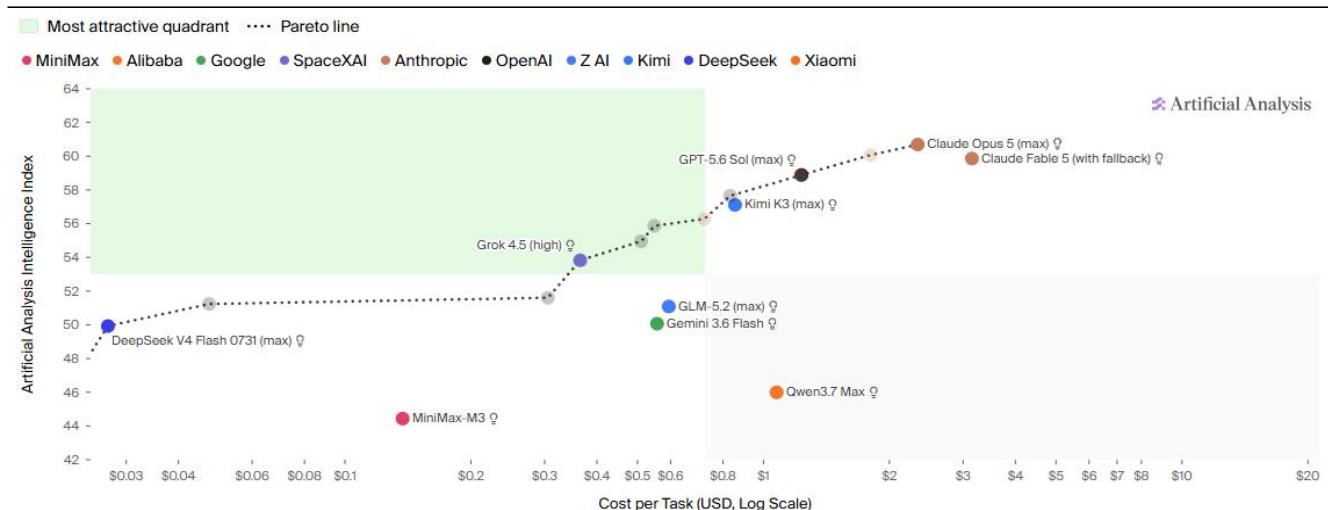
*帕累托前沿（Pareto Frontier）指的是一组最优解（权衡方案）的集合。它的核心逻辑是：当同时优化 A 和 B 两个目标时，任何一方性能的提升，必然以牺牲另一方为代价。这些最佳权衡点连成的曲线或边界，就是帕累托前沿。

图表 3：完成一个标准智能测试任务所需的加权平均成本（以美元计）



资料来源：Artificial analysis，交银国际 数据截至 8 月 3 日

图表 4：传统 Transformer 与 KDA 架构在处理长上下文的对比



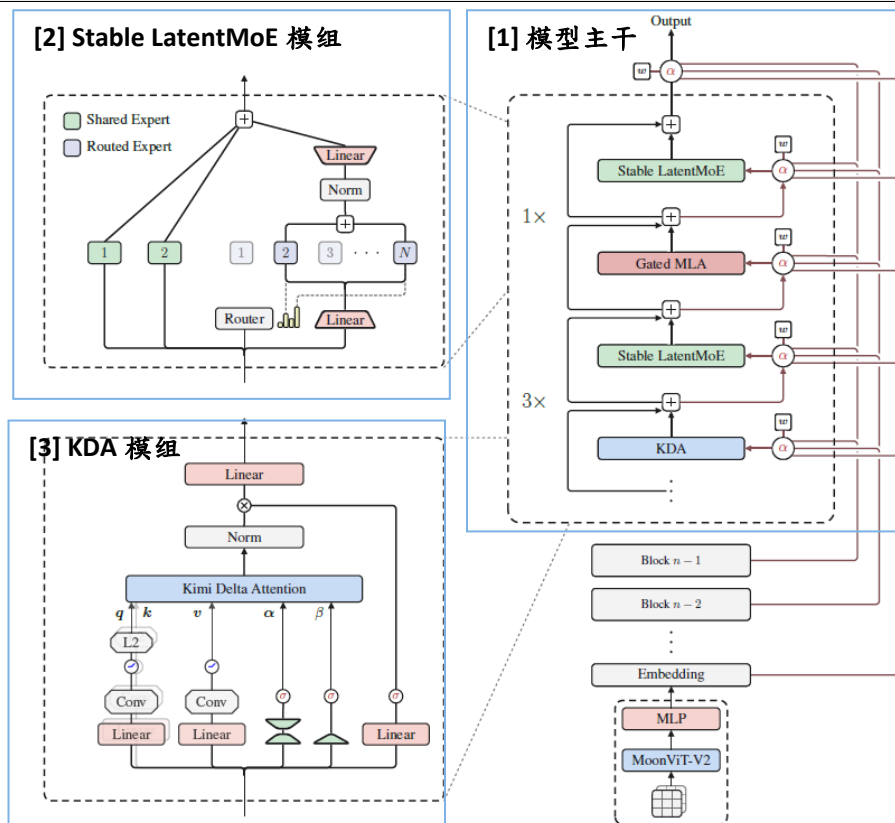
资料来源：Artificial analysis，交银国际 数据截至 8 月 3 日

参考图表 5，Kimi K3 模型架构可以分为 3 个主体：

- 1) **右上是模型主干**：展开的 block 按“三层 KDA 加一层 Gated MLA”的 3:1 节奏排布，每层注意力后接一个 Stable LatentMoE 前馈网络，层层堆叠、并以一层全局 MLA 收尾，共 93 层；图表右缘的 w 方框、 α 圆圈，以及连向 Embedding 与各前层 block 的红褐色连线，正是 AttnRes——让每一层能回头“取阅”前面任何楼层的摘要，而非沿传话链逐层衰减。
- 2) **左上方 Stable LatentMoE 模组**：展示包含 896 个路由专家与 2 个共享专家的稀疏前馈网络架构，路由机制针对每个 Token 仅动态激活 16 个路由专家，从而在超大规模参数下实现极高的计算稀疏度与效率。
- 3) **左下方 KDA 模组**：具备通道级遗忘门与有界衰减机制的递归注意力，其通过维护固定大小的循环状态取代了随序列长度线性膨胀的 KV Cache，构成了实现百万级上下文恒定显存推理的架构基础。

总括来说，主干上的 KDA、AttnRes 与 MoE，即论文所谓“沿序列、深度、宽度三维扩展信息流”。此外，右下方的 MoonViT-V2 视觉路径则在这一主干的输入端注入原生多模态。

图表 5：Kimi K3 模型架构图



资料来源：月之暗面，交银国际

模型架构深度解析 – 从技术看经济性

核心观点：在 AI 大模型的投资分析中，“架构”这个词往往让非技术背景的投资期望而却步。但我们认为在目前大模型公司的竞争格局下，特别是盈利基础暂时偏薄弱的阶段，架构选择并非是纯技术问题，而是更偏重于商业模式及经济性的问题。例如：注意力机制的选择决定了推理成本曲线，稀疏度的设计决定了训练资本效率，多模态的整合方式决定了产品形态的天花板。理解架构能更好地判断一家大模型公司的成本结构是否可持续、技术护城河是否可复制、产品路线图是否可信。

K3 的架构设计同时在三个维度上扩展了信息在模型中流动的带宽：让模型读得更长（序列维度）、想得更深（深度维度）、知道得更多（宽度维度）。以下各节将逐一拆解这三个维度的具体实现，并解释每一个设计选择背后的经济逻辑。

图表 6：K3 模型架构设计亮点

维度	机制	作用
序列长度	KDA + Gated MLA（3:1 混合）	百万 Token 高效处理
网络深度	Attention Residuals（Block AttnRes）	跨层选择性信息检索
模型宽度	Stable LatentMoE（896 专家/16 激活专家）	稀疏通道混合

资料来源：月之暗面，交银国际

注意力混合架构带来长上下文推理成本的压缩

Kimi K3 采用的 KDA（Kimi Delta Attention）机制，除了提升模型智能水平外，亦改变了长上下文推理的成本结构，使显存占用不再随上下文长度线性膨胀。在 Agentic 应用需要频繁处理数十万 Token 级别的情景下，在架构层面的成本优势，将直接转化为单位经济模型和定价权，从而提升大模型公司的护城河。

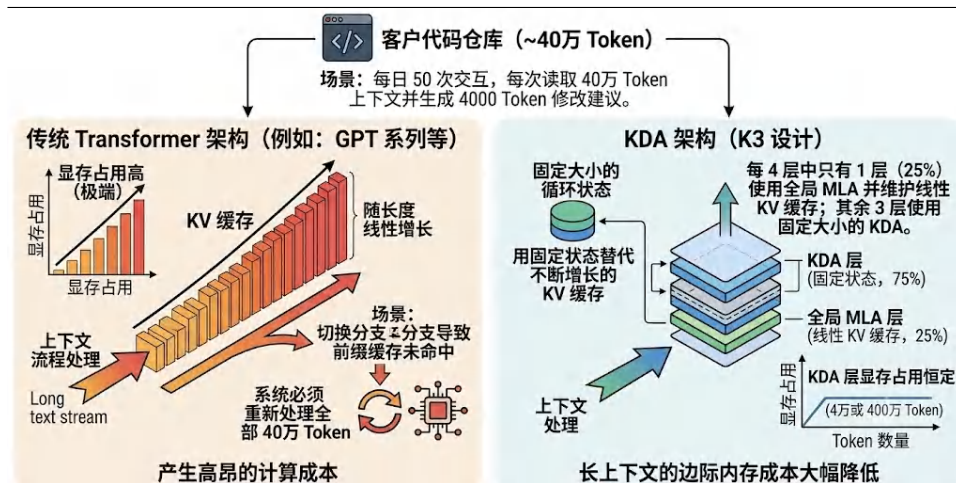
我们以一个典型的企业场景来说明。假设客户其中一个代码仓库包含约 40 万 Token，每次对话模型需读取这 40 万 Token 上下文，并生成 4 千 Token 的修改建议，每天交互 50 次。在传统的 Transformer 架构（如 GPT 系列等）中，处理文本的核心机制是“全局注意力”，意思是模型每读入一个新词，都需要将这个词与之前的所有词进行关联比对。虽然近期模型通过 KV Cache 技术避免了重复计算，但模型必须为之前所有的词维护一份随文本长度线性增长的缓存数据。上下文越长，占用的显存就越大，导致传统架构在处理长文本时成本高昂。

Kimi K3 采用 3:1 的 KDA 与 MLA 混合架构，将长上下文推理的内存占用从按长度线性增长，转变为次线性固定，重构了长文本处理的成本曲线。KDA 放弃了不停累积的笔记式记忆，转用固定大小的循环工作记忆。新信息写入时，会通过遗忘门按比例衰减旧信息。然而，因为固定记忆存在细节流失的局限，K3 并未完全抛弃全局注意力，而是每隔 3 层 KDA 保留 1 层 Gated MLA（Gated Multi-head Latent Attention，由 DeepSeek 开创的高效全局注意力）。在总计 93 层的

网络中，仅有 24 层维护随长度增长的 KV Cache，其余 69 层内存恒定。这意味着当上下文从 1 万 Token 扩展至 100 万 Token 时，K3 的内存膨胀主要由 Gated MLA 的层（25%）决定，长上下文的边际内存成本被大幅压缩。

用上述例子说明，无论上下文是 4 万还是 400 万 Token，因 KDA 层的显存占用固定不变，长上下文的边际内存成本得到了有效压缩。这种固定状态的架构不仅缓解了大模型对显存容量的压力，还提升了有限显存下的前缀缓存命中率。

图表 7：传统 Transformer 与 KDA 架构在处理长上下文的对比



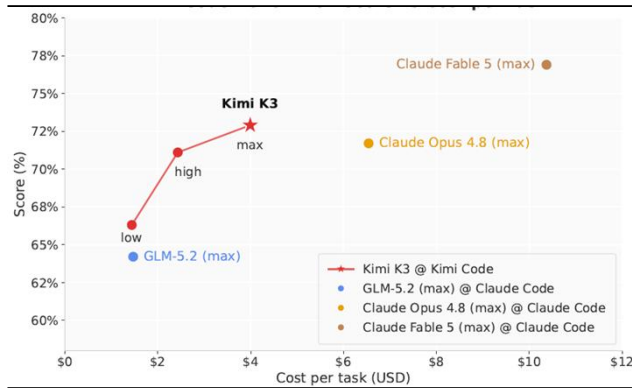
资料来源：月之暗面，交银国际

更重要的是，KDA 层对显存占用的固定，能优化“投机解码”的效率。投机解码是大模型加速生成文本的常用技术，模型在正式输出前会先“抢跑”预测几个词，如果预测错了就需要推翻重来（即状态回滚）。在传统架构下，处理超长文本时一旦预测出错，撤销已生成的冗长记忆缓存代价极高。而 KDA 的固定记忆机制，使模型在“纠错回滚”时只需处理极小的数据量，大幅降低了这种试错带来的算力浪费，从而有效提升了长文本生成的速度。

由于 KDA 将大部分的缓存状态固化，需要缓存的数据量得到有效减少，从而在有限的显存空间内提高了有效前缀的命中率。这些系统工程上的优势叠加 MXFP4 量化，最终在 BrowseComp 测试（图表 9）中得到了兑现。Kimi K3 以 91.2% 的准确率和单任务 2.03 美元的成本，显著低于同类闭源模型（如 Claude Fable 5 准确率 88.0%，成本超 20 美元）。这种倍量级的成本差距，其根源正是 KDA 架构带来的长上下文推理经济性。

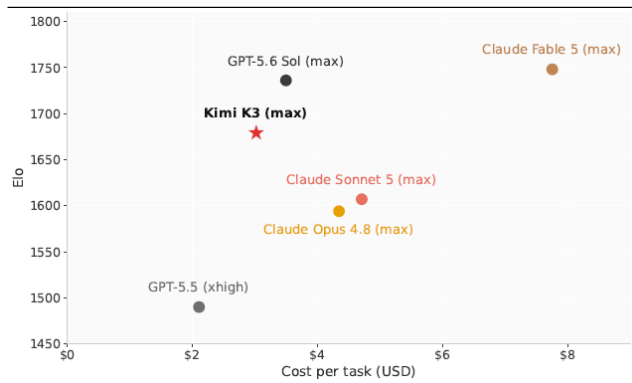
我们认为 KDA 与 MLA 混合架构为 K3 提供了结构性成本优势，有望改善 API 服务的毛利率并赋予其更强的价格竞争力。同时，大模型公司的估值其中的重要因素也建基在单位经济上。如果 K3 的底层架构能够有效压低推理算力成本，在同等 API 定价下，K3 的毛利率或有条件优于采用传统架构的竞品。

图表 8：Kimi Code Bench 2.0 测试: 评估模型作为在真实软件开发场景中，端到端解决跨语言、多技术栈工程问题的实战编程能力与算力成本效益



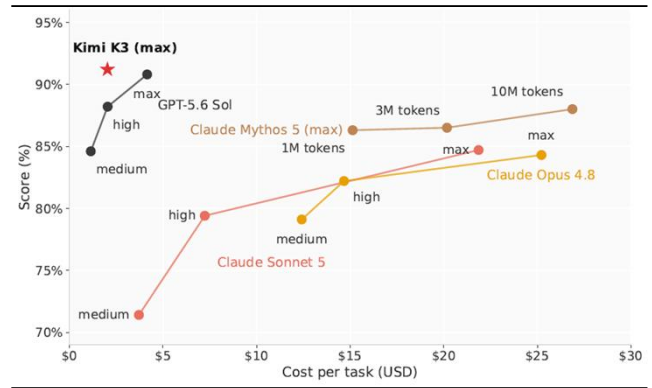
资料来源：月之暗面，交银国际

图表 10：GDPval-AA v2 测试: 衡量模型在执行具有现实经济价值的专业知识工作时，其通用智能水平与实际生产力转化表现



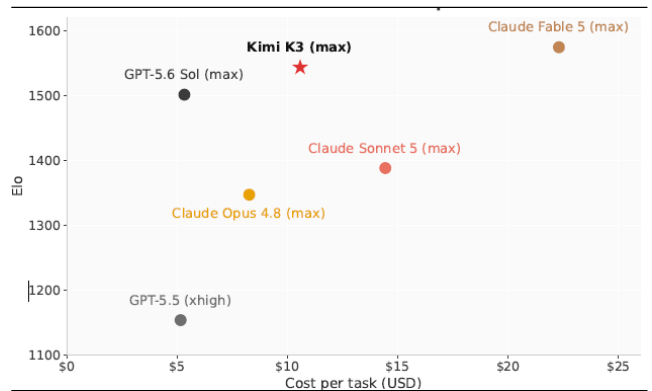
资料来源：月之暗面，交银国际

图表 9：BrowseComp 测试: 考核模型作为自主浏览代理，在复杂开放网络环境中进行多步搜索、信息甄别与深度研究的综合能力



资料来源：月之暗面，交银国际

图表 11：AA-Briefcase 测试: 专门检验模型在复杂知识型白领工作场景下，端到端执行多步骤专业知识任务并输出可用交付物的智能体能力



资料来源：月之暗面，交银国际

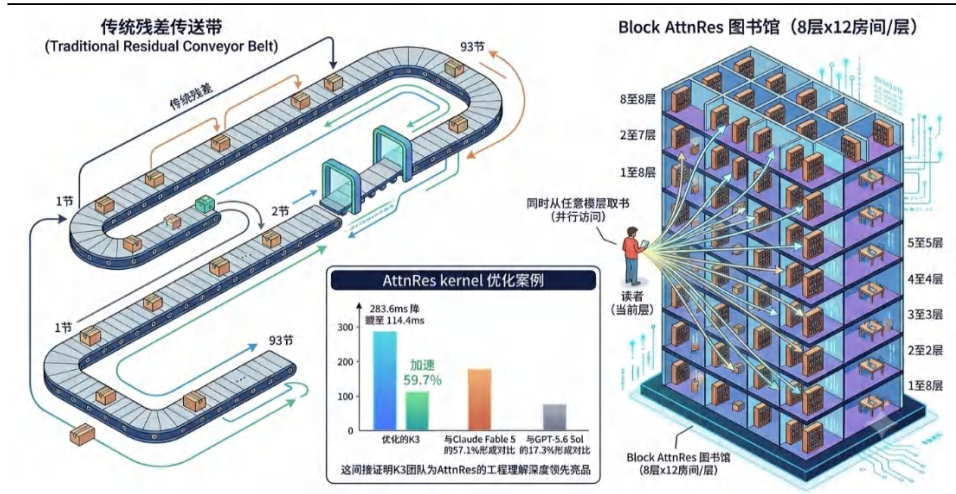
AttnRes：从“一维覆盖”到“二维检索”，信息不再随深度衰减

如果用最直观的几何视角来理解，传统的 ResNet 残差连接是一维空间里的覆盖式写入，信息像接力棒一样在 93 层的单行道上逐层传递并逐渐衰减，模型被迫用更深的网络来“补偿”遗忘。而 K3 的 AttnRes 则将其升维到了二维空间的索引式写入，每一层都将自己的原始输出作为原件存入，后续的任何一层都可以跨越深度，直接按需要取阅前序任何一层的原件。这种从“一维”到“二维”的拓扑学跃迁，有效消除了深度网络中的信息瓶颈，这将内存消耗从 $O(93d)$ 压缩到 $O(9d)$ ，使得这个设计在工程上可行。

自 2015 年 ResNet 提出以来，深度网络都普遍使用逐层传递的方式。用几何视角看，所有信息只能沿着深度这一维空间顺序写入，并无法回头查阅。这可以比喻为一条 93 人的传话链。第一个人（代表第 1 层）看到原始文本，形成初步理解，然后把自己的批注（代表残差更新）加到原话，再传给第二个人，又加上自己的批注传给第三个人，如此类推。而因为原话经 92 次传递后的篇幅仍固定（状态维度不变），所以后来新批注加进来越多，越早的细节就被挤得越模糊，到第 93 个人看到的原始细节早已面目全非。这就是为什么深度网络需要堆叠的层数非常多，原因不是每一层都需要大量计算，而是因信息在传递中衰减，需要额外的层来找回/重建在逐层传递中被挤掉及扭曲的早期信息。

AttnRes 的做法是打破旧有的传话链，换成像图书馆的架构。用几何视角看，就是把信息的写入从一维单行道升级为二维空间：在原有的“深度”轴之外，增加了第二条“检索”轴，每一层处理完信息后不再去覆盖那一句话，而是把自己的原始输出永久存档在对应层号的“书架”上。在形成第 93 层最终判断时，可以直接走到“书架”前，翻阅浅层保留的第一手原始档案、中层整理的分类读书笔记、或是撰写的深度研究报告——不需要再听中间层的口耳相传。

图表 12：传统 Transformer 与 KDA 架构在处理长上下文的对比



资料来源：月之暗面，交银国际

当然，如果真让每一层存储和访问所有 93 层的完整输出，内存开销将是巨大的（实际上是 $O(93d)$ ，其中 d 是隐藏维度 7,168）。AttnRes 原始论文提出了一个务实的折中：Block AttnRes。将 93 层分成 8 个“层块”，每个层块约 12 层。

每层只需要访问 8 个层块的“摘要”加上本层块内的累积，内存可从 $O(93d)$ 降到 $O(9d)$ ，达到大幅度的压缩。

K3 的具体配置是 8 个层块，前 7 个层块各 12 层，最后一个层块 9 层，加上 embedding 层共 9 个“信息源”。每个层块的“摘要”是该层块内所有层输出的求和，成为一个简单但有效的压缩。AttnRes 的注意力权重是数据依赖的，对于不同的 Token，模型从不同层块中取阅的比例不同。这代表着对于简单 Token（如“的”、“是”），模型可能主要从最后几个层块取信息（等效于较浅的计算）。如果是对于复杂 Token（如专业推理的关键步骤），模型可能均匀地从所有层块取信息（等效于完整的 93 层深度）。

AttnRes 与 KDA 和 MLA 的交互也值得注意。在 K3 的架构中，每个层块内部按“3 层 KDA+1 层 MLA”的 3:1 节奏排布，一个完整的 12 层层块含 9 层 KDA 和 3 层 MLA（每层注意力后接对应的 MoE 前馈网络）。AttnRes 的“楼层摘要”因此同时包含了 KDA 的递归状态信息和 MLA 的全局注意力信息。这使得 AttnRes 的“图书馆”中，每个书架上的“书”既有“直觉判断”（KDA），也有“精确查阅”（MLA），读者（当前层）可以按需选择。

AttnRes 对大模型公司的影响主要通过两点：1) 部分贡献在“2.5 的 scaling efficiency”：通过消除逐层信息衰减，同样的 93 层产生更高的有效信息利用率，达到同等性能所需的训练 FLOPs 更少；2) 为未来的“动态深度推理”优化预留了架构接口：如果推理引擎能够根据注意力权重动态，跳过“未被使用”的层块，K3 的有效推理成本可能显著低于“93 层×每层计算量”的简单估算。这是一个尚未实现但架构上已预留的优化空间。

以上两点对月之暗面增加了迭代及定价权的优势：1) 加速产品迭代周期在“同等算力预算”的约束下，更高的效率允许团队缩短单次训练周期，更快地推出下一代模型（如 K4）。在 AI 行业，抢先 3 个月发布往往意味着抢占开发者心智和首批企业客户的预算，有望直接转化为早期的收入先发优势；2) 提高 API 业务的毛利率上限：推理算力是 API 服务最核心的营业成本，而当前行业的成本核算均基于“名义深度”假设。但 AttnRes 的注意力权重是数据依赖的，复杂 Token 分散于全部层块，有效推理深度因输入而异。若未来推理引擎能基于注意力权重实装动态深度推理，单 Token 的实际 FLOPs 将显著下降。在 API 终端定价不变的情况下，这一边际成本的下降将直接推高公司毛利率上限。

Stable Latent MoE 专家分流成就万亿参数的稳定训练

K3 拥有 896 个“专家”模块（路由专家），每个 Token 只激活其中 16 个，稀疏度达到 56:1。这是当前公开模型中专家数量最多、稀疏度最高的设计。但极端稀疏带来了两个工程难题：激活值爆炸和负载不均衡。K3 通过三个原创组件解决了这两个问题，使得 896 个专家的极端稀疏设计在 2.8 万亿参数规模下稳定训练成为可能。

MoE（Mixture of Experts，混合专家）机制目前已在不少大模型中使用，核心在于解耦模型的总参数量与单次推理的计算量。在处理每个输入 Token 时，路由网络会动态计算其与各专家的匹配度，并仅将其分配给得分最高的少数几个专家进行前馈计算与加权聚合。通过这种稀疏激活策略，模型能够在大幅扩充总参数规模以提升知识容量的同时，有效控制单次前向传播的算力开销。

专家数目的提升，直观的结论是代表可以对专家更精细的分工，例如 Python 代码、法律文本、数学公式等任务都有专门处理的专家。K3 的 56:1 稀疏度机制下，模型即使拥有 2.8 万亿参数的知识容量，但每个 Token 只“咨询”其中约 1,042 亿参数，即是可以较少的算力成本，去获取近 3 万亿参数级别的知识储备。但把稀疏度推到极端，在 896 个专家叠加 2.8 万亿参数的情况下，随即引出两个工程挑战：

- ⊕ **挑战一：数值溢出导致训练崩溃。** 模型在处理数据时，信息需经过连续四步的矩阵运算，任何微小的数值波动都会在这条计算链中被逐级放大。在超大规模参数下，这种放大效应容易导致数据超出硬件的计算上限，直接引发训练崩溃。

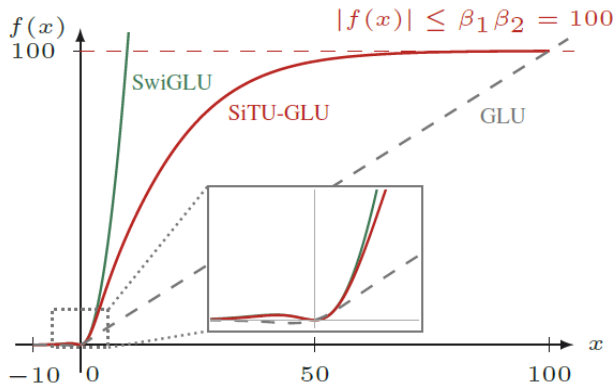
K3 的解法：引入 SiTU-GLU (Sigmoid Tanh UnitGated Linear Units) 激活函数。传统的计算通道像一条没有限速的高速公路，数值可以无限加速，或有失控的风险。SiTU-GLU 则在通道两侧加装了“弹性护栏”，当数值逼近上限时，护栏会平滑地将其压回安全区间，既防止了溢出崩溃，又没有损失模型处理常规数据的精准度。

- ⊕ **挑战二：负载倾斜导致算力闲置。** 如果某些热门的专家被过度集中调用，它们所在的 GPU 就会严重过载，而“冷门”专家则无人问津。传统做法是加设一个“惩罚项”来强行摊派任务，就像对拥堵路段收罚款，“惩罚项”减分相对调用程度难以一概定论，而且容易干扰模型主任务的正常学习。

K3 的解法：采用无辅助损失的分位数均衡（Quantile Balancing）。K3 彻底抛弃了“罚款”思路，直接统计并计算出“要让每条路的车流量绝对平均，该收多少过路费”，随后一步到位地设定该费率。这种方法无需人工调参，能在几步计算内让 896 个专家的工作量达到绝对均衡，消除了算力空耗。

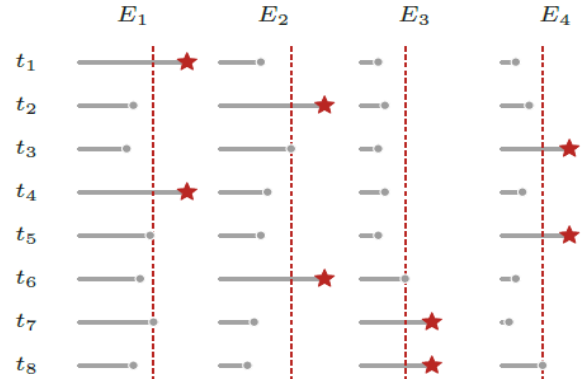
896 专家/16 专家激活的设计是 K3 “知识容量”到“计算成本”分离的核心。K3 的推理计算成本等效于一个 1040 亿参数的稠密模型，但其知识容量和 benchmark 表现接近甚至超越 2000 亿+参数的稠密模型。对于 API 定价而言，其结构性优势在于 K3 可以以“1042 亿模型的成本”提供“2.8 万亿模型的能力”。同时，以 QB 的无损失均衡特性，训练过程中没有“为了均衡而浪费”的计算，即所有 FLOPs 都用于提升模型能力，而非维持负载均衡。这直接贡献了 2.5 倍的 scaling efficiency。

图表 13：Kimi K3 的 SiTU-GLU 激活函数：在正常数值范围内它与 SwiGLU 几乎完全一样，但数值过大时会平滑地封顶在 100



资料来源：月之暗面，交银国际

图表 14：Kimi K3 的分位数均衡机制，仅凭一次前向传播，就算出每个专家的新偏置，核心动作是把让每个专家恰好接到 q 个 token



资料来源：月之暗面，交银国际

预训练层：打破参数迷信，工程效率重塑成本

核心观点：市场对预训练的理解一般停留在参数总量，用了多少 GPU 的卡群等等实体数字。但在 2.8 万亿参数的规模上，真正的问题是如何用同样资本开支，能使 K3 比 K2 提升智能。K3 相对 K2 约 2.5 倍的整体扩展效率提升，是数据、scaling law、长上下文扩展三个维度数十个选择的叠加结果。

K3 的预训练语料覆盖四大文本域（Web 文本、代码、数学、知识）加一个大规模视觉语料。过滤三件套——规则启发式、分类器质量打分、去重——已是行业标准做法，不构成差异化。值得展开的是 rephrasing：团队对知识和数学语料用风格与视角多样化的提示进行重写，采用分块自回归生成，并对照源文档做忠实性验证。用比喻说：这不是让学生背诵课本原文，而是让不同的老师用不同的口吻把同一个知识讲给学生听——学生学到的是知识本身，而不是某本书的措辞。忠实性验证则是防止“改写者歪曲原意”的质检环节。这道工序的直接效果是降低模型对特定文本风格的过拟合，提升知识迁移能力。

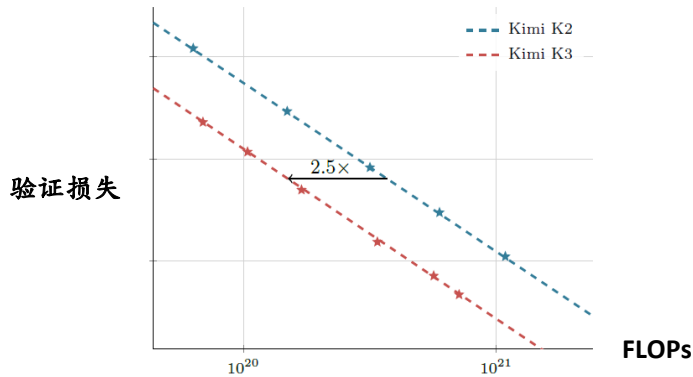
程序化多模态数据是另一个关键选择。团队大幅扩展了“代码片段与其渲染视觉相耦合”的数据，覆盖 SVG、3D 资产、网页、游戏、CAD 图纸等领域特定格式。这与“给图片配文字说明”的传统做法有本质区别——模型看到的不是“一张图加一句描述”，而是“一份图纸和按这份图纸建出来的成品”，代码与视觉之间的因果对应关系天然成立。这正是第三章所述“写代码→看渲染截图→改代码”闭环能力的侧面解释，也是 Table 4 中 K3 在 3D/WebGL/Shader 任务上对 Claude Opus 4.8 取得+59.1%净胜率的底层来源。此外，坐标监督同时提供绝对与归一化两种格式，使定位能力对不同分辨率保持稳健。

参数上 2.8 万亿是突破，但效率提升含金量更高

我们认为市场对 Kimi K3 的第一观感，在于 K3 是 2.8 万亿参数的全球最大开源模型，并认为参数量最大就是 SOTA（State of the art）模型。同时，市场仍将参数量与模型能力及研发投入直接画等号，但若仅关注规模扩张，将忽视 Kimi K3 工程效率在算力竞赛中的决定性作用。

K3 的技术报告指出，通过 Kimi Delta Attention、Attention Residuals 及 Stable Latent MoE 等架构创新，K3 在整体缩放效率上较 K2 提升了约 2.5 倍（图表 15）。这表示在相同验证损失的情况下，K3 所需的计算量仅为 K2 的 40%。换个角度来说，在同等算力预算下，K3 能够合理地支撑更庞大的参数规模（总参数较 K2 增长 167%），路由专家数增长 133%下形成更高复杂度的专家网络，以及上下文窗口较 K2 扩大 8 倍至 1 百万（图表 16）。我们认为这并非由简单堆算力所产生的结果，而是透过架构与系统工程，对计算资源进行极限压缩的成果，表明开发团队具备将固定 GPU 预算转化为更高模型性能的工程能力。

图表 15：K2 vs K3 scaling law 曲线



资料来源：月之暗面，交银国际

我们认为缩放效率的提升，主要影响是减低了对未来大模型迭代的研发资本开支预期，有助于延长公司的现金跑道，同时适度降低稀释性融资的风险。在一般的现金流估算中，投资者较习惯按照参数数量的线性增长，来推导算力成本线性上升，以及AI企业的未来资本开支。然而，Kimi K3的2.5倍效率提升或者能打破这个线性逻辑思维的惯性。如果该效率曲线能够持续，企业在研发下一代模型时，即使参数规模再翻倍，所需的算力投入增幅将低于参数规模的线性增幅。AI企业如果成功做到，将直接降低了模型迭代对现金流的消耗压力，同时在亏损阶段，现金消耗速度能有所减慢，使企业的生存周期延长。

图表 16：K2 与 K3 的模型建构比较

项目	Kimi K2	Kimi K3	变化
架构 (Architecture)	MoE (混合专家模型)	MoE	-
层数 (Layers)	61	93	↑ 52%
总参数量 (Total Parameters)	1.04 万亿	2.78 万亿	↑ 167%
激活参数量 (Activated Parameters)	326 亿	1,042 亿	↑ 220%
隐藏层维度 (Hidden Dimension)	7,168	7,168	=
潜在 MoE 维度 (Latent MoE Dimension)	-	3584 (0.5x)	-
每个专家的 MoE 隐藏层维度	2,048	3,072	↑ 50%
路由专家数 (Routed Experts)	384	896	↑ 133%
每个 Token 激活的专家数	8	16	↑ 100%
共享专家数 (Shared Experts)	1	2	↑ 100%
注意力头数 (Attention Heads)	64	96	↑ 50%
密集层数量 (Number of Dense Layers)	1	1	=
词表大小 (Vocabulary Size)	160K	160K	=
训练上下文长度 (Training Context Length)	128K	1M (100 万)	8x
注意力机制 (Attention Mechanism)	MLA*	混合 KDA-MLA	-
激活函数 (Activation Function)	SwiGLU	SiTU-GLU	-
注意力层组成 (Attention-Layer Composition)	61 MLA	69 KDA + 24 MLA	-
MTP 层数量 (Number of MTP Layers) *	1 层	1 层	无变化
ViT 总参数量*	-	4.01 亿	-
ViT 层数 (ViT Layers)	-	27 layers	-
ViT 补丁大小 (Patch Size of ViT)	-	14	-
ViT 注意力头数 (Attention Heads of ViT)	-	12	-

资料来源：月之暗面，交银国际 * ViT: Vision Transformer, MLA: Multi-Head Latent Attention, MTP: Multi-Token Prediction

MoonEP 系统带来算力成本实质性节约

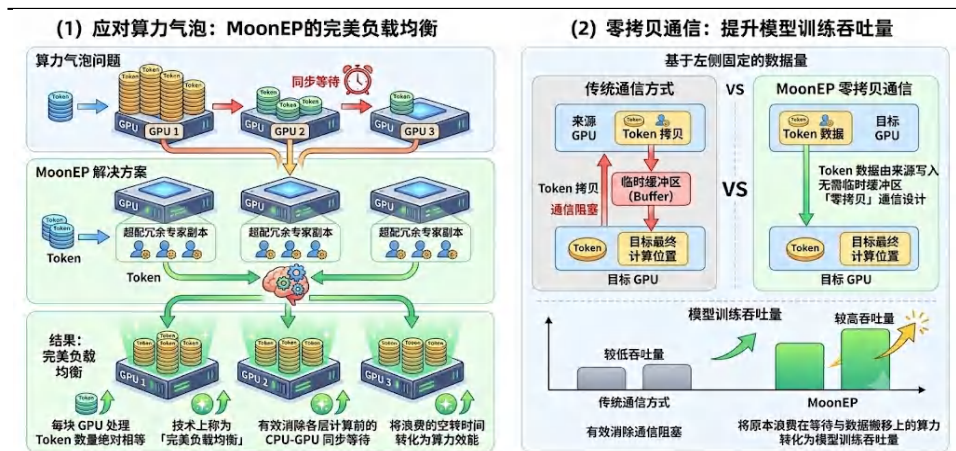
月之暗面自研的 MoonEP 可理解成预训练阶段的基础设施，为解决超大规模专家网络的算力调度难题。通过动态调度冗余专家，保证每块 GPU 在每步训练处理的 Token 数量绝对相等，消除了传统架构中“快卡等慢卡”导致的算力气泡与闲置，将 GPU 集群的利用率推向物理极限，大幅压缩训练成本。

Kimi K3 的 MoonEP 是创新的专家并行（Expert Parallelism）设定。在 K3 的预训练阶段，MoonEP 作用于底层混合专家（MoE）网络的调度层，是一个专为大模型定制的算力负载均衡系统。K3 拥有 896 个专家，但每次计算只激活 16 个，极易导致部分 GPU 过载而部分闲置。传统 MoE 路由受制于输入序列的高度随机性，经常导致严重的专家负载倾斜。在千卡集群的强制同步机制下，整体计算速度被迫等待最慢的 GPU，形成大规模空转的“算力气泡”。

应对算力气泡，MoonEP 的解决方案是在每块 GPU 上超配预留特定数量的冗余专家副本，只要冗余阈值满足特定条件，通过动态调度这些冗余专家，可以确保每块 GPU 在每一步训练中，接收并处理的 Token 数量绝对相等。这种在技术论文中称为“完美负载均衡”，有效消除各层计算前的 CPU-GPU 同步等待，将原本浪费的空转时间转化为算力效能。同时，固定的数据量使 MoonEP 能够实现「零拷贝」通信，即 Token 数据由来源直接写入目标的最终计算位置，无需经过任何临时缓冲区。“零拷贝”通信设计有效消除通信阻塞，同时将原本浪费在等待与数据搬移上的算力，转化为模型训练吞吐量。

对大模型公司而言，MoonEP 解决的是训练成本问题，直接降低了近 3 万亿参数规模下的现金消耗速率。因为在这个规模下，训练一次模型动用的 GPU 卡群集需要消耗上百万 GPU 小时。10%至 30%的利用率提升，意味着同等硬件条件下训练周期缩短数周，或同等时间下所需采购或租赁的 GPU 数量大幅减少。对于尚未盈利的 AI 企业，这直接减慢了现金消耗速度，延长了融资周期。

图表 17：MoonEP 的技术原理



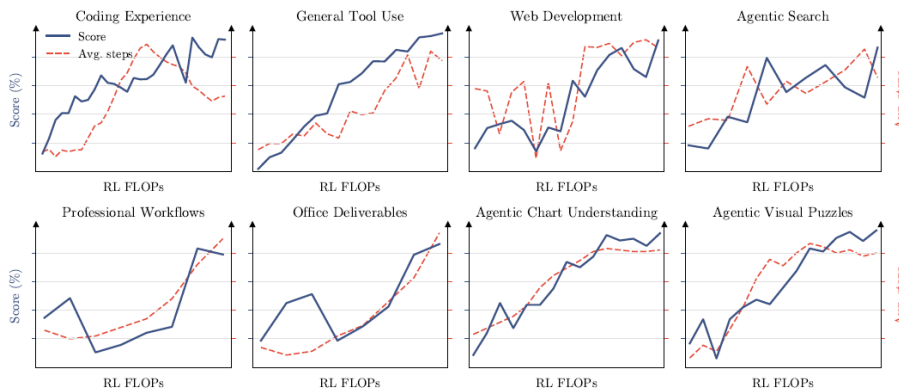
资料来源：月之暗面，交银国际

后训练：专家团队淬炼到训练场组成的护城河

核心观点：K3 的后训练遵循 Supervised Fine-Tuning (SFT) → Reinforced Learning (RL) → Multi-Teacher On-Policy Distillation (MOPD) 三阶段流水线：SFT 建立高质量冷启动策略，并引入量化感知训练，RL 在三大任务领域与三档推理强度上交叉训练 9 个专家模型，MOPD 以多教师在线策略蒸馏将九个专家整合为统一模型，形成推理强度可调的产品形态。三个阶段紧密结合后构成工程壁垒的一部分。

- ④ **阶段一，SFT：建立冷启动策略，量化感知从第一天介入。** SFT 负责打好基础，用原模型生成任务数据并加人工审核，格式通用，模型学起来不费劲。更进一步的是开始时就按“压缩版”标准训练，保证模型上线后的实际表现和训练时一模一样，效能不会打折扣。
- ④ **阶段二，RL：9 专家矩阵，与两个工程解法。** K3 把通用任务、通用智能体、编码智能体三大领域，与 low/high/max 三档推理强度交叉，训练出 9 个专家模型。K3 以两个机制化解长程 RL 的主要痛点：1) **针对长尾延迟采用 partial rollout**：不等最慢的，一部分轨迹跑完就先开始学习，没跑完的暂停、下轮接着跑。数据过时也不怕，靠“小步更新”稳住训练；2) **针对 token 效率采用 Reasoning Effort RL**：每题先设预算，超支直接判为不合格，先放开训出深思档，再逐步收紧训出快答档。
- ④ **阶段三，MOPD：在线策略蒸馏将 9 个专家整合为统一模型。** 9 个专家最终经蒸馏合并成一个模型。奖励不是算在九个专家模型的标准答案上，而是算在 K3（统一模型）先自行采样，写出一份完整回答，再由对应任务领域的专家模型进行打分。产出的模型带“强度调节”，简单问题快答，复杂任务深思，使推理成本变成用户按需可调的变量，这正是后训练直接塑造商业模式的案例。

图表 18：RL 算力每上一台阶，八项评测得分与工具步数同升，验证 RL 是第二条 scaling 轴



资料来源：月之暗面，交银国际

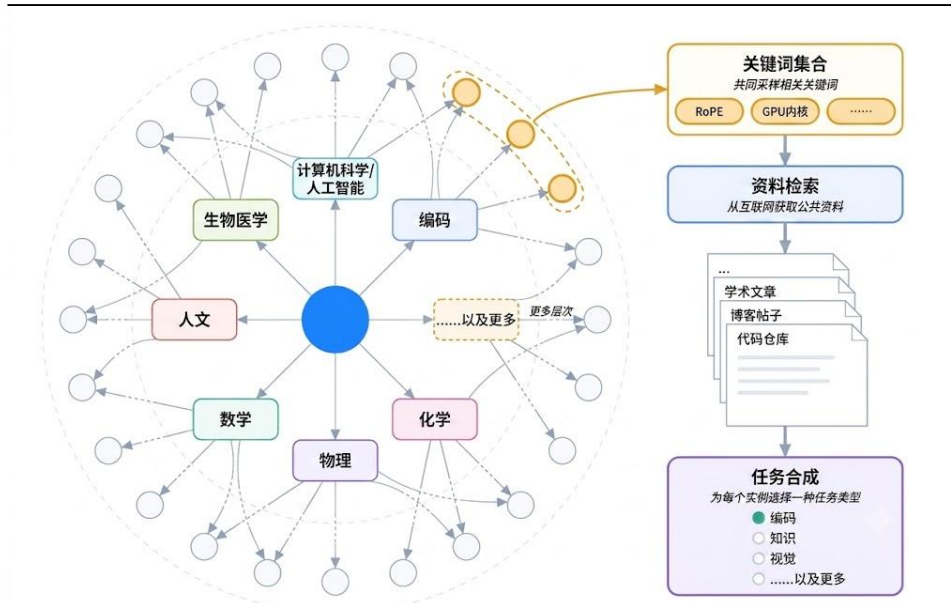
训练场与出题工厂：RL 真正的护城河

RL 的效果上限取决于环境的丰富度、保真度和可验证性。K3 构建了统一白盒环境（防止过拟合单一工具框架）、自我演化的知识图谱出题工厂、以及七类高保真训练场。这部分是竞争对手最难复制的内容，因为它不是知道就能做到，而是做过才能做到。

如果模型只在一种 agent 框架（比如固定的工具接口、系统提示、上下文管理策略）里训练，它可能会过度拟合（overfitting）到这套“考试格式”而非有真实的推理能力。K3 把 agent harness 抽象为可配置、可组合的模块（工具接口、系统提示、上下文管理、技能、记忆、子 agent），通过配置即可实例化的 Kimi Code、Claude Code、Codex、OpenClaw、Hermes 等主流框架，并在训练中动态切换。情况可以驾校作比喻，不让学员只练一款车及相同路段，而是频繁换车款及路段，以提升真实的驾驶能力。

知识图谱出题工厂解决的是题目从何而来的问题。如图表 19 所示，一张由 agent 持续扩张的层级化知识图谱（从计算机科学/人工智能、化学、物理等大学科到细分概念），按不同粒度采样节点组成关键词集，驱动互联网真实材料（论文、博客、代码仓库）的检索，再由合成 agent 产出各类任务。细分粒度采样能深挖专业知识，跨域采样保证广度，真实材料保证任务质量，形成由训练、探索、做题多样、复利增长的数据资产。

图表 19：K3 利用一个自进化的“知识图谱”，从互联网上自动检索真实材料，进而源源不断地合成高质量、不重复的复杂考题，打破了依赖人类出题的数据瓶颈，大幅降低了后训练成本。



资料来源：月之暗面，交银国际

2026 年 8 月 12 日

互联网行业-数据与智能

以上七类训练场各自针对一种真实工作能力。可验证搜索与专业工作（投行、数据分析、法律实务，数十到数百步完成交付物）。例如：

- 1) **带 Python 解释器的视觉推理**：裁剪、放大、变换图像并验证中间结果。
- 2) **内核优化任务**：奖励及考核执行的正确性与性能，匹配专家实现得 0.5 分，逼近硬件 roofoffline 趋向 1 分，并配有持续升级的反作弊系统，惩罚 CUDA graph 重放、输入缓存、降精度等黑客策略，相当于考场里不断更新的监考规则。
- 3) **个人助手任务**：在模拟的 Gmail、Notion、Slack、Canvas 中跨多个模拟日处理数十个相互依赖事件，单次 rollout 可达数千次工具调用、数百万上下文 Token。
- 4) **自主执行任务**：只给目标和验证接口，不给参考流程，奖励基于验证器对最终环境状态的评估而非 agent 的自我汇报。
- 5) **Web 开发任务**：功能检查加模型裁判双轨评分，如出现项目构建失败、带错运行、伪造而非实现等情况，交付时的奖励直接归零。

这些环境的共同特征是可验证与保持高度真实性。一方面能保证奖励信号真实，模型无法靠自我汇报骗取得分，同时亦保证训练分布贴近真实部署分布。对投资者而言，这张知识图谱与 K3 的 5100 万个沙盒实例，共同构成了 K3 的实测数据库，虽然权重可以开源，但这些资产无法随权重下载。

Pre-IPO 融资估值反映对 Kimi ARR 持续高增的预期

目前月之暗面刚完成 F 轮融资 35 亿美元，投后估值较今年 5 月上升 75% 至 350 亿美元。以今年 6 月底 3 亿美元 ARR（年度经常性收入）测算，其对应 P/ARR 估值在 100 倍以上。目前市场预期月之暗面将在 8 月开始 Pre-IPO 轮融资，估值可能上升至 500 亿美元。我们认为，市场对 Kimi K3 的推出可以大幅提升 ARR 有高度信心，500 亿美元的估值应是建立在月之暗面的 ARR 可以在未来 6-12 个月内提升至约 10 亿美元的预期。我们认为在 2H26 期间，公司是否可以及时补充算力、提高用户体验、在大模型迭代加快的情况下保持优势及 ARR，将是支撑更高估值的关键。

在港上市公司中，潜在受益于月之暗面 K3 推出/估值提升的公司包括：

- ④ **招商局中国基金（133 HK/未评级）**：曾参与月之暗面 A 轮融资，基金采用审慎估值（或含前期成本计价），2025 年底账面值为 1676 万美元，如计及 2026 年至今的估值提升，2026 年 6 月底该项投资的账面值已上升至 5500 万美元，占公司每股净资产净值（38.8 港元）的 5.7%。招商局中国基金或仍受惠月之暗面后续估值提升的红利。
- ④ **中软国际（354 HK/未评级）**：与月之暗面签署“登月计划”，正式向高专业价值 Token 运营商转型。双方合作包括底层、应用、算力等方面：1）**技术与产品底座**：将中软国际的 AllMeta 企业智能操作系统与月之暗面的 Kimi K2.7 Code 及 K3 模型深度绑定；2）**商业模式创新**：建立“Token 分成机制”，对合作落地中产生的 Kimi 大模型 Token 消耗收益及衍生收入按比例分成；3）**算力协同**：中软国际自建行业算力中心将优先满足月之暗面的算力需求，形成“从算力供给到企业应用”的一体化服务。其中，我们认为 Token 分成机制是可以快速在下半年贡献公司整体收入的利好。

图表 20：月之暗面过去主要融资轮次比较

项目	轮次	融资额(美元)	投后估值(美元)	主要投资者
2023 年 4 月	天使轮	超 2 亿	3 亿	红杉中国、真格基金、今日资本
2023 年 7 月	A1 轮	未披露	未披露	美团龙珠
2024 年 2 月	A+轮	超 10 亿	25 亿	阿里巴巴、红杉、美团、小红书
2024 年 8 月	B 轮	超 3 亿	33 亿	腾讯、高榕
2025 年 12 月	C 轮	5 亿	43 亿	IDG 资本、腾讯
2026 年 1 月-2 月	C+轮及延伸	合计 19 亿	100 亿→180 亿	阿里巴巴、腾讯、IDG 资本、美团龙珠、中国移动
2026 年 5 月	D 轮	约 20 亿	约 200 亿	美团龙珠、中移动、源峰资本
2026 年 6 月	E 轮	超 20 亿	约 315 亿	多地国资、IDG 资本
2026 年 7 月	F 轮	超 35 亿	350 亿	中芯聚源、深创投
2026 年 8 月 (预估)	G 轮/Pre-IPO	未披露	超 500 亿(目标)	未披露

资料来源：公开资料整理，交银国际

风险因素

- ④ **研究级推理能力的差距仍可能被闭源模型抛离。**K3 在编码、智能体执行、Web 开发等可直接变现的赛道领先，但在研究级任务上落后闭源 SOTA，例如 HLE-Full 与 CritPt 等。尽管当前的差距尚未冲击其商业化进程，但如果闭源同业的更新模型在科学研究与原创推理上进一步拉开，K3 的 SOTA 地位仍将受挑战。
- ④ **开源后，先发优势可能随时间快速衰减。**K3 真正的竞争壁垒，如近 3 万亿参数下的 2.5 倍训练效率、5,100 万沙箱与知识图谱出题工厂等等，无法随权重公开下载，但这些优势能否维持仍取决于时间。若同业在短期内复现工程体系，或以蒸馏方法推出成本更低的可比模型，K3 先发优势将快速衰减。
- ④ **算力资源仍是对迭代及 ARR 增长的最大约束。**K3 的下一代迭代与 ARR 的跃升，仍依赖大规模高端算力补充。如果在 2H26 算力获取不及预期，模型迭代速度与推理服务体验将同时受限，直接冲击公司估值。
- ④ **网络安全能力既是入场券，也可能因此被加强监管。**K3 能发现 Linux 内核的 16 个此前未知漏洞，这既是大模型在网络安全领域的潜在切入点，也意味着大模型公司有可能引起监管机构及政府的关注，以及后续潜在推出更严格的出口管制、部署限制与安全审查。这可能直接限制高模型能力版本的发布节奏或 API 服务地域范围。
- ④ **价格竞争可能压缩全行业利润空间。**在相若的智能水平下，K3 跟 SOTA 闭源模型仍有相对成本优势，但如果开源同业在新版模型智能水平提升后仍主动降价，或者闭源模型选择降价守住市场份额，行业 API 价格水平可能系统性下移，K3 的成本优势将被单价下降部分减低，并影响毛利率改善路径。

交銀國際

香港中环德辅道中 68 号万宜大厦 9 楼
总机: (852) 3766 1899 传真: (852) 2107 4662

评级定义

分析员个股评级定义：

买入：预期个股未来12个月的总回报**高于**相关行业。

中性：预期个股未来12个月的总回报与相关行业**一致**。

沽出：预期个股未来12个月的总回报**低于**相关行业

无评级：对于个股未来12个月的总回报与相关行业的比较，分析员**并无确信观点**。

分析员行业评级定义：

领先：分析员预期所覆盖行业未来12个月的表现相对于大盘标杆指数**具吸引力**。

同步：分析员预期所覆盖行业未来12个月的表现与大盘标杆指数**一致**。

落后：分析员预期所覆盖行业未来12个月的表现相对于大盘标杆指数**不具吸引力**。

香港市场的标杆指数为**恒生综合指数**，A股市场的标杆指数为**MSCI 中国A股指数**，美国上市中概股的标杆指数为**标普美国中概股50（美元）指数**

2026年8月12日

互联网行业-数据与智能

分析员披露

本研究报告之作者，兹作以下声明：i) 发表于本报告之观点准确地反映关于他们个人对所提及的证券或其发行者之观点；及ii) 他们之薪酬与发表于报告上之建议/观点并无直接或间接关系；iii) 对于提及的证券或其发行者，他们并无接收到可影响他们的建议的内幕消息/非公开股价敏感消息。

本报告之作者进一步确认：i) 他们及他们之相关有联系者【按香港证券及期货监察委员会之操守准则的相关定义】并没有于发表本报告之30个日历日前交易或买卖本报告内涉及其所评论的任何公司的证券；ii) 他们及他们之相关有联系者并没有担任本报告内涉及其评论的任何公司的高级人员（包括就房地产基金而言，担任该房地产基金的管理公司的高级人员；及就任何其他实体而言，在该实体中担任负责管理该等公司的高级人员或其同级人员）；iii) 他们及他们之相关有联系者并没有拥有于本报告内涉及其评论的任何公司的证券之任何财务利益。根据证监会持牌人或注册人操守准则第16.2段，“有联系者”指：i) 分析员的配偶、亲生或领养的未成年子女，或未成年继子女；ii) 某信托的受托人，而分析员、其配偶、其亲生或领养的未成年子女或其未成年继子女是该信托的受益人或酌情对象；或iii) 惯于或有义务按照分析员的指示或指令行事的另一人。

有关商务关系及财务权益之披露

交银国际证券有限公司及/或其有关联公司在过去十二个月内与交通银行股份有限公司、国联证券股份有限公司、交银国际控股有限公司、四川能投发展股份有限公司、光年控股有限公司、武汉有机控股有限公司、Sincere Watch (Hong Kong) Limited、滴普科技股份有限公司、山东快驴科技发展有限公司、拔康视云制药有限公司、富卫集团有限公司、宜搜科技控股有限公司、广州银诺医药集团股份有限公司、劲方医药科技（上海）股份有限公司、长风药业股份有限公司、武汉艾米森生命科技股份有限公司、上海海达科技发展有限公司、上海森亿医疗科技股份有限公司、协创数据技术股份有限公司、上海宝济药业股份有限公司、深圳迅策科技股份有限公司、北京智谱华章科技股份有限公司、天九共享智慧企业服务股份有限公司、红星冷链（湖南）股份有限公司、爱芯元智半导体股份有限公司、牧原食品股份有限公司、国民技术股份有限公司、福信富通科技股份有限公司、宁波菲仕技术股份有限公司、智慧互通科技股份有限公司、深圳市兆威机电股份有限公司、南京埃斯顿自动化股份有限公司、上海易景信息科技股份有限公司、广东云徙智能科技股份有限公司、驭势科技（北京）股份有限公司、北京深演智能科技股份有限公司及Yangtuo Holdings Inc.有投资银行业务关系。

交银国际证券有限公司及/或其集团公司现持有东方证券股份有限公司及光大证券股份有限公司的已发行股本逾1%。

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），表示并保证其根据下述的条件下有权获得本报告，并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。

本报告为高度机密，并且只以非公开形式供交银国际证券的客户阅览。本报告只在基于能被保密的情况下提供给阁下。未经交银国际证券事先以书面同意，本报告及其中所载的资料不得以任何形式(i)复制、复印或储存，或者(ii) 直接或者间接分发或者转交予任何其它人作任何用途。

交银国际证券、其附属公司、关联公司、董事、关联方及/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、或对其有兴趣。此外，交银国际证券、其附属公司及关联公司可能与本报告内所述或有关的公司不时进行业务往来，或为其担任市场庄家，或被委任替其证券进行承销，或可能以委托人身份替客户买入或沽售其证券，或可能为其担当或争取担当并提供投资银行、顾问、包销、融资或其它服务，或替其从其它实体寻求同类型之服务。投资者在阅读本报告时，应该留意任何或所有上述的情况，均可能导致真正或潜在的利益冲突。

本报告内的资料来自交银国际证券在报告发行时相信为正确及可靠的来源，惟本报告并非旨在包含投资者所需要的所有信息，并可能受递送延误、阻碍或拦截等因子所影响。交银国际证券不明示或暗示地保证或表示任何该等数据或意见的足够性、准确性、完整性、可靠性或公平性。因此，交银国际证券及其集团或有关的成员均不会就由于任何第三方在依赖本报告的内容时所作的行为而导致的任何类型的损失（包括但不限于任何直接的、间接的、随之而发生的损失）而负上任何责任。

本报告只为一般性提供数据之性质，旨在供交银国际证券之客户作一般阅览之用，而非考虑任何某特定收取者的特定投资目标、财务状况或任何特别需要。本报告内的任何资料或意见均不构成或被视为集团的任何成员作出提议、建议或征求购入或出售任何证券、有关投资或其它金融证券。

本报告之观点、推荐、建议和意见均不一定反映交银国际证券或其集团的立场，亦可在没有提供通知的情况下随时更改，交银国际证券亦无责任提供任何有关资料或意见之更新。

交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，在参与有关报告内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。本报告的发送对象不包括身处中国内地的投资人。如知悉收取或发送本报告有可能构成当地法律、法则或其他规定之违反，本报告的收取者承诺尽快通知交银国际证券。

本免责声明以中英文书写，两种文本具同等效力。若两种文本有矛盾之处，则应以英文版本为准。

交银国际证券有限公司是交通银行股份有限公司的附属公司。



机构销售团队



@bocomgroup.com

张靖

(852) 3710 3330

jing.zhang

张家尔

(852) 3710 3206

William.Zhang

www.bocomgroup.com