

Kimi K3 发布，引领开源大模型发展

推荐|维持

——人工智能行业半月报（2026 年 7 月第 2 期）

报告要点：

● 指数表现（2026 年 7 月下半月）

内地指数：中证人工智能产业指数下跌 14.14%，延续下跌态势。国证 AI 应用指数上涨 2.31%，上证科创板人工智能指数下跌 12.41%。

香港指数：恒生港股通人工智能主题指数下跌 9.24%，延续下跌态势。

● 数据跟踪

据 OpenRouter 数据显示，在刚过去的一个整周(07.27-08.02)中，按 token 计算的大模型周调用量为 56.8 万亿 tokens，周调用量最高的编程智能体与生产力工具均为 Hermes Agent。据 Artificial Analysis 数据显示，8 月 5 日，以 AAll 为标准，Anthropic Claude 与 OpenAI GPT 系列模型的智能水平总体上仍保持领先；当前智能水平最高的国产大模型是 Kimi K3；OpenAI 有多个性价比高的 GPT 系列模型。当前总参数量最高的开源大模型是 Kimi K3。

● 重点新闻

大模型企业：Anthropic 宣布 Fable 5 永久保留，纳入相关订阅；月之暗面暂停 C 端新用户订阅；智谱收购中科加禾 60% 股权；谷歌更新轻量版 Gemini 模型阵容，但旗舰 Pro 发布再延期；Claude Opus 5 发布，性能接近 Fable 5、价格减半；Kimi K3 正式开源：全球首个 3 万亿级参数开源模型落地；OpenAI 下调 GPT-5.6 价格；MiniMax H3 发布并将开源；DeepSeek-V4-Flash 正式版上线；亚马逊完成对 OpenAI 总计 500 亿美元全额投资；Qwen3.8-Max 首发上线。

底层算力基础设施：超节点集中亮相世界人工智能大会；AMD 推出首款机架级 AI 系统 Helios 挑战英伟达；AMD 与 Anthropic 达成数百亿美元 AI 服务器合作，并投资最高 50 亿美元。

人工智能运用：OpenAI 发布企业 AI 智能体运营平台 Presence；Anthropic 智能体平台 CMA 技能上限升至 500；OpenAI 向 10 万名学术研究者免费开放前沿模型；谷歌发布 Gemini Robotics 2。

政策法规与安全性：工信部副部长柯吉欣：建设人工智能领域制造业创新中心，实施“模数共振”行动；OpenAI 模型攻破 HuggingFace，GLM 取证反击；英伟达黄仁勋开源倡议获多家机构支持，但 Anthropic 未签署；英伟达等成立开放安全 AI 联盟，推动 AI 安全防御；超千名 AI 大厂员工联名，请求政府协调控制研发节奏；Anthropic 披露 Claude 在安全评测中未经授权访问了 3 家机构的真实系统。

● 投资观点

我们建议继续关注人工智能应用拓展及终端客户落地的情况，密切跟踪头部大模型厂商场景适配布局，综合研判厂商的模型能力转化效率与中长期盈利空间，从而筛选出核心受益的上市标的。

● 风险提示

大模型技术迭代不及预期风险；人工智能基础设施需求不及预期风险；人工智能技术应用不及预期风险；宏观经济表现不及预期风险；行业竞争加剧的风险；相关政策落地不及预期的风险。

过去一年市场行情



资料来源：iFinD

相关研究报告

《国元证券行业研究-人工智能行业半月报（2026 年 7 月第 1 期）：大模型持续迭代创新，token 调用量创新高》2026.07.17

报告作者

分析师 耿军军

执业证书编号 S0020519070002

电话 021-51097188-1856

邮箱 gengjunjun@gyzq.com.cn

目 录

1. 指数表现.....	3
1.1 内地股票相关指数.....	3
1.2 香港股票相关指数.....	4
2. 数据跟踪.....	5
2.1 大模型调用与测评.....	5
2.2 大模型价格与参数.....	9
2.3 应用软件与智能体.....	11
3. 重点新闻.....	13
3.1 大模型与代表企业.....	13
3.2 底层算力基础设施.....	16
3.3 人工智能实践应用.....	17
3.4 政策法规与安全性.....	18
4. 投资观点.....	20
5. 风险提示.....	20

图表目录

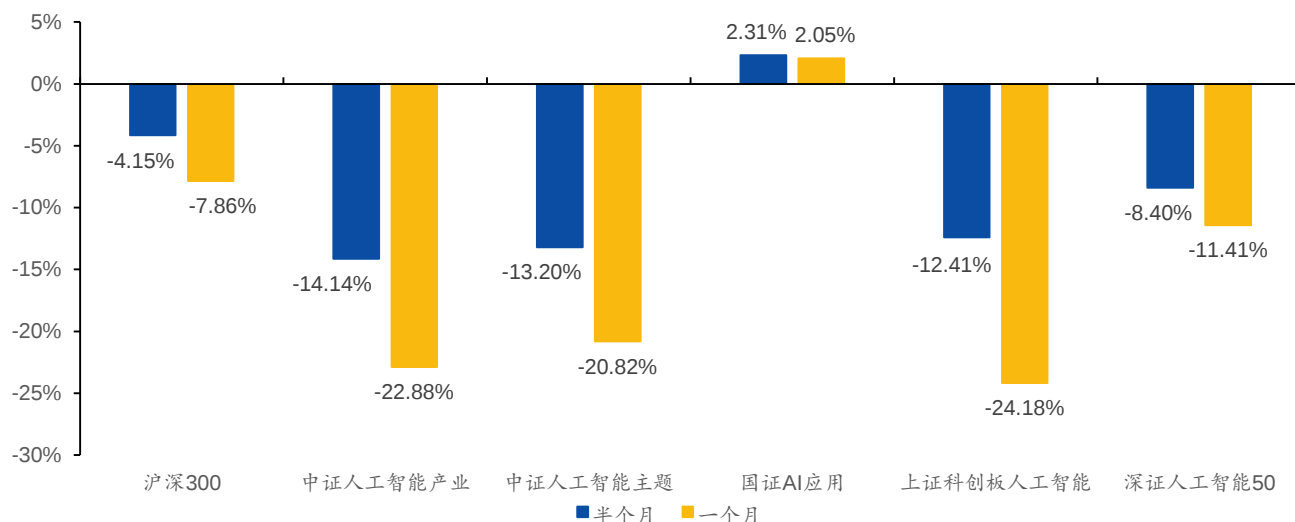
图 1：沪深 300 指数及人工智能概念指数半月度与月度涨跌幅.....	3
图 2：沪深 300 指数及人工智能概念指数 2026 年初至今累计涨跌幅.....	3
图 3：恒生系列核心指数及恒生港股通人工智能主题指数半月度与月度涨跌幅.....	4
图 4：恒生系列核心指数及恒生港股通人工智能主题指数 2026 年初至今累计涨跌幅.....	4
图 5：大模型调用统计与近三周排名（截至 8 月 2 日）.....	5
图 6：大模型厂商市场份额统计与近两周排名（以文本请求为衡量标准，截至 8 月 2 日）.....	6
图 7：大模型智能水平综合排名（以 AAIL 或 AAI 为标准，截至 8 月 5 日）.....	7
图 8：大模型厂商前沿语言模型智能水平对比（全部/当前前十厂商，以 AAIL 为标准，截至 8 月 5 日）.....	8
图 9：大模型智能水平与任务加权平均成本对比图（智能水平以 AAIL 为标准，截至 8 月 5 日）.....	9
图 10：开源大模型总参数量对比（十亿单位，截至 8 月 5 日）.....	10
图 11：开源大模型智能水平与总参数量对比图（智能水平以 AAIL 为标准，截至 8 月 5 日）.....	10
图 12：编程智能体调用统计与近三周排名（截至 8 月 2 日）.....	11
图 13：生产力工具调用统计与近三周排名（截至 8 月 2 日）.....	12

1. 指数表现

1.1 内地股票相关指数

2026 年 7 月 16 日至 31 日，中证人工智能产业指数下跌 14.14%，延续了 7 月上半月的下跌态势，跑输沪深 300 指数 9.99 个百分点。同一时间段内，国证 AI 应用指数上涨 2.31%，上证科创板人工智能指数下跌 12.41%。

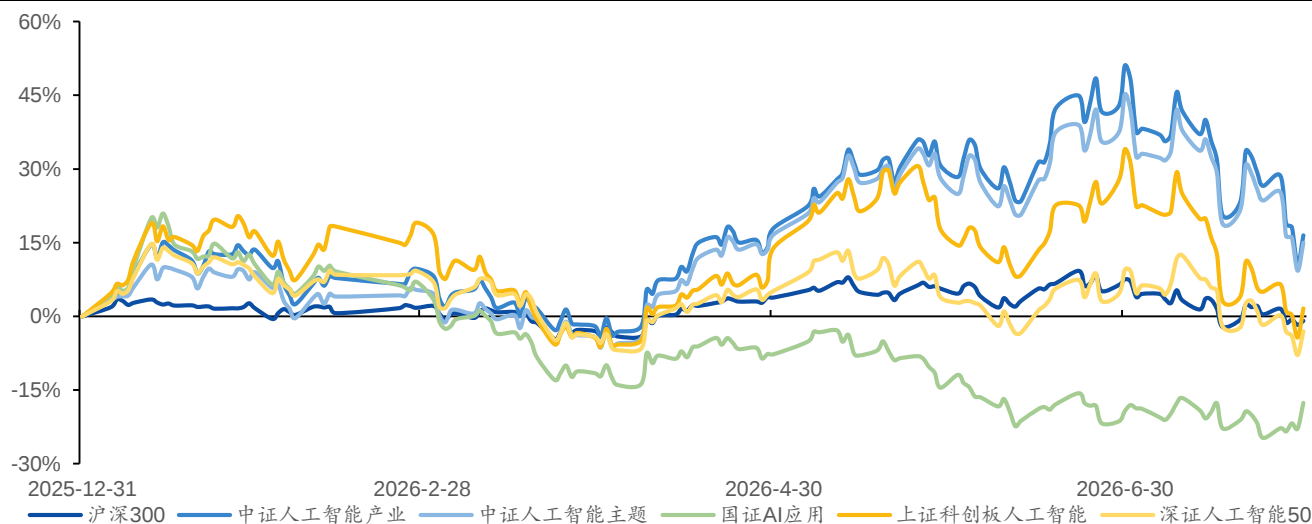
图 1：沪深 300 指数及人工智能概念指数半月度与月度涨跌幅



资料来源：iFinD，国元证券研究所

今年以来，中证人工智能产业指数累计上涨 16.51%，跑赢沪深 300 指数 17.41 个百分点。国证 AI 应用指数累计下跌 17.63%，上证科创板人工智能指数累计上涨 1.62%。

图 2：沪深 300 指数及人工智能概念指数 2026 年初至今累计涨跌幅

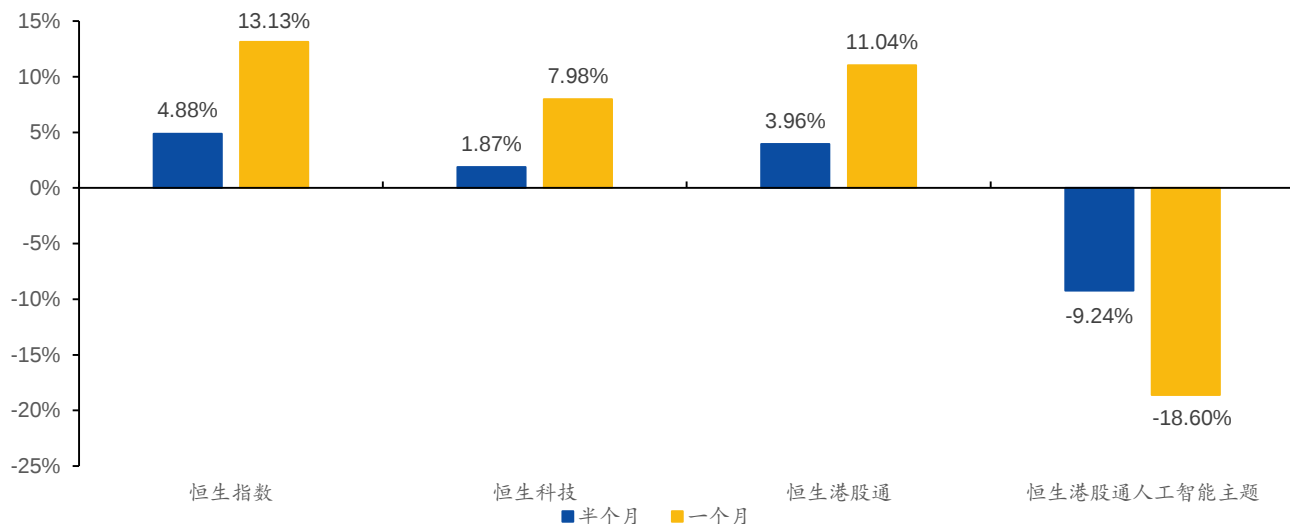


资料来源：iFinD，国元证券研究所

1.2 香港股票相关指数

2026 年 7 月 16 日至 31 日，恒生港股通人工智能主题指数下跌 9.24%，延续了 7 月上半月的下跌态势，跑输恒生指数 14.12 个百分点。同一时间段内，恒生科技指数上涨 1.87%，恒生港股通指数上涨 3.96%。

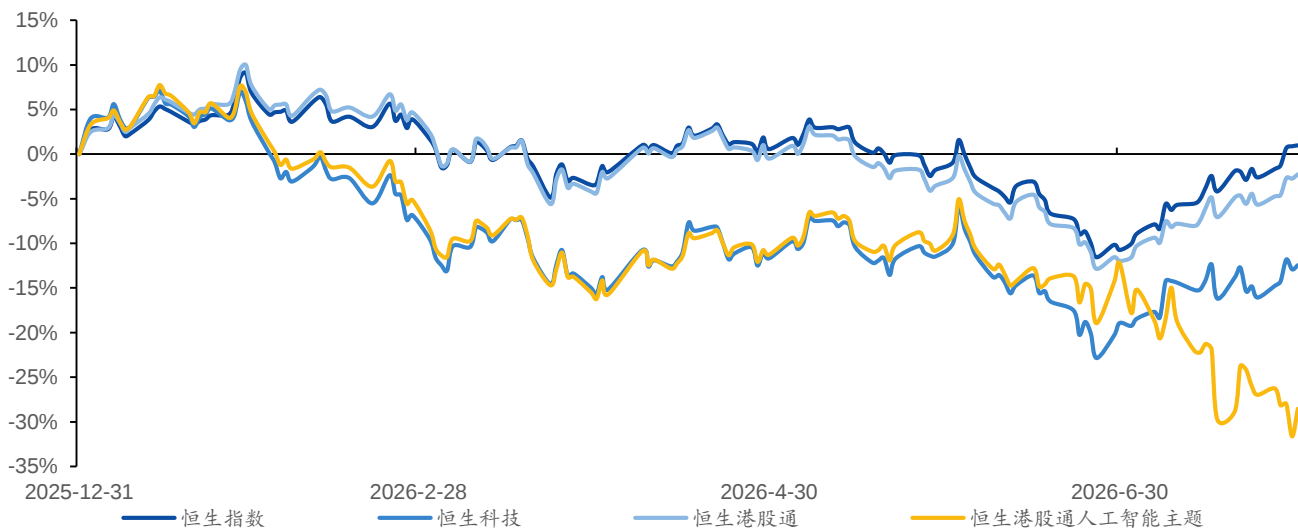
图 3：恒生系列核心指数及恒生港股通人工智能主题指数半月度与月度涨跌幅



资料来源：iFinD，国元证券研究所

今年以来，恒生港股通人工智能主题指数累计下跌 28.54%，跑输恒生指数 29.53 个百分点。恒生科技指数累计下跌 12.45%，恒生港股通指数累计下跌 2.27%。

图 4：恒生系列核心指数及恒生港股通人工智能主题指数 2026 年初至今累计涨跌幅



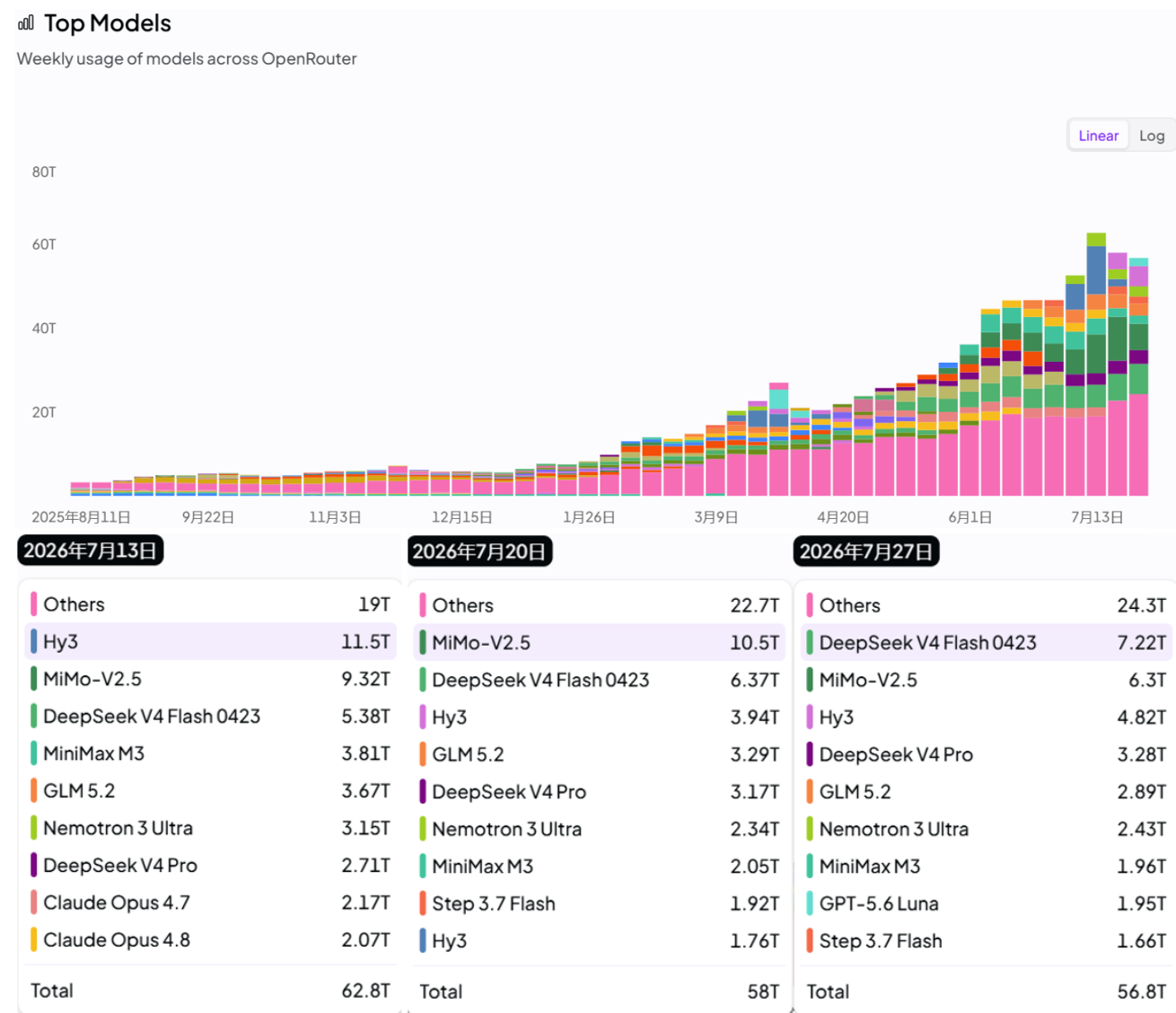
资料来源：iFinD，国元证券研究所

2.数据跟踪

2.1 大模型调用与测评

据 OpenRouter 统计数据显示，按 token（词元）计算的大模型周调用量在刚过去的一个整周(2026.07.27-2026.08.02)为 56.8 万亿 tokens，较上一个整周(2026.07.20-2026.07.26)的 58 万亿 tokens 有所下降。其中，该周调用排名前五的大模型分别为 DeepSeek 的 DeepSeek V4 Flash 0423、小米的 MiMo-V2.5、腾讯的 Hy3、DeepSeek 的 DeepSeek V4 Pro 以及智谱的 GLM 5.2。

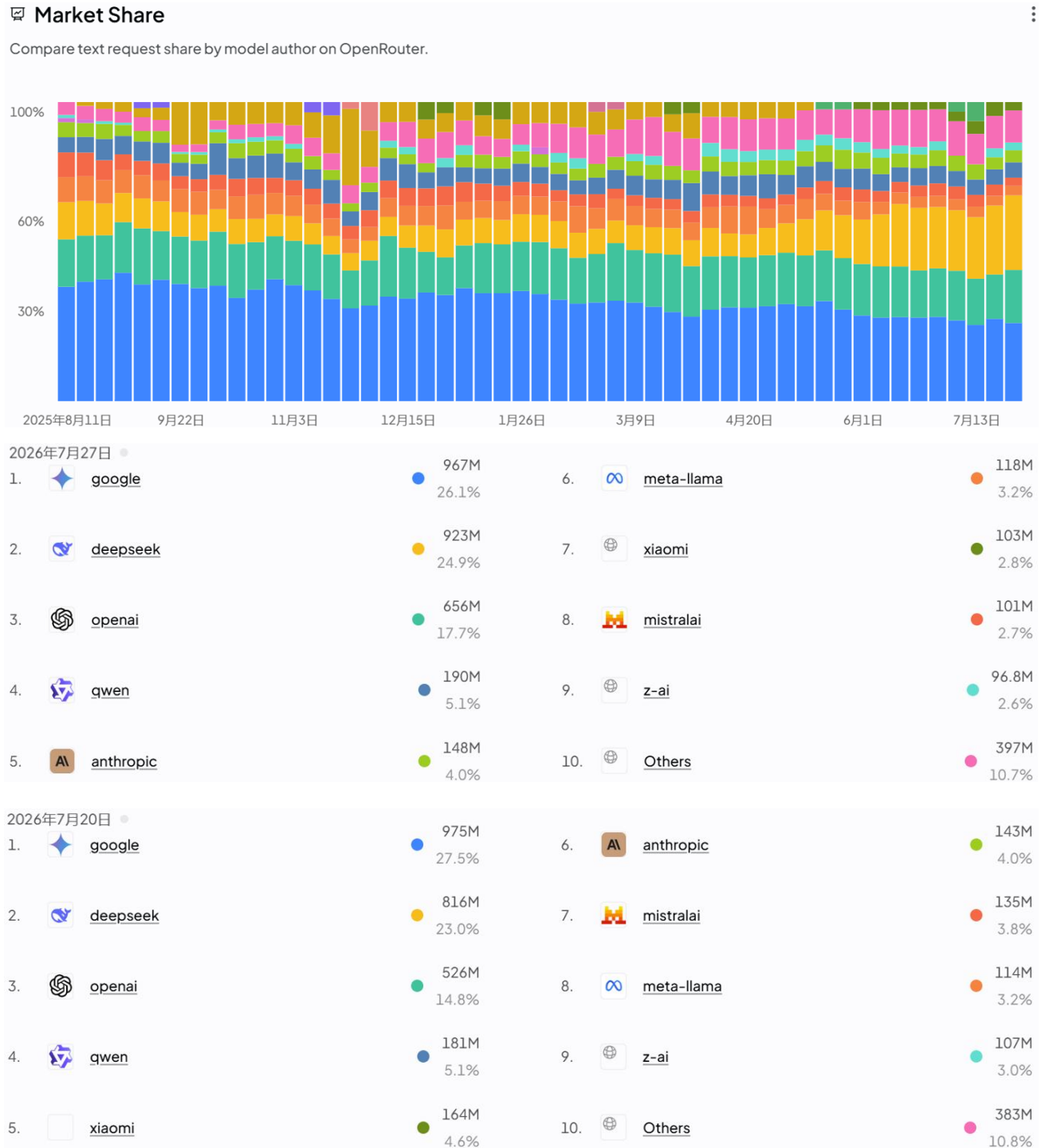
图 5：大模型调用统计与近三周排名（截至 8 月 2 日）



资料来源：OpenRouter，国元证券研究所

以文本请求为衡量标准，在最近过去的一个整周(2026.07.27-2026.08.02)，谷歌的市场份额继续保持首位(26.1%)，排名第二的是 DeepSeek(24.9%)。

图 6：大模型厂商市场份额统计与近两周排名（以文本请求为衡量标准，截至 8 月 2 日）



资料来源：OpenRouter，国元证券研究所

综合 Artificial Analysis 的测评结果看，Anthropic Claude 与 OpenAI GPT 系列模型的智能水平总体上仍保持领先；当前智能水平最高的国产大模型是 Kimi K3。

图 7：大模型智能水平综合排名（以 AAIL 或 AAI 为标准，截至 8 月 5 日）

Artificial Analysis Intelligence Index

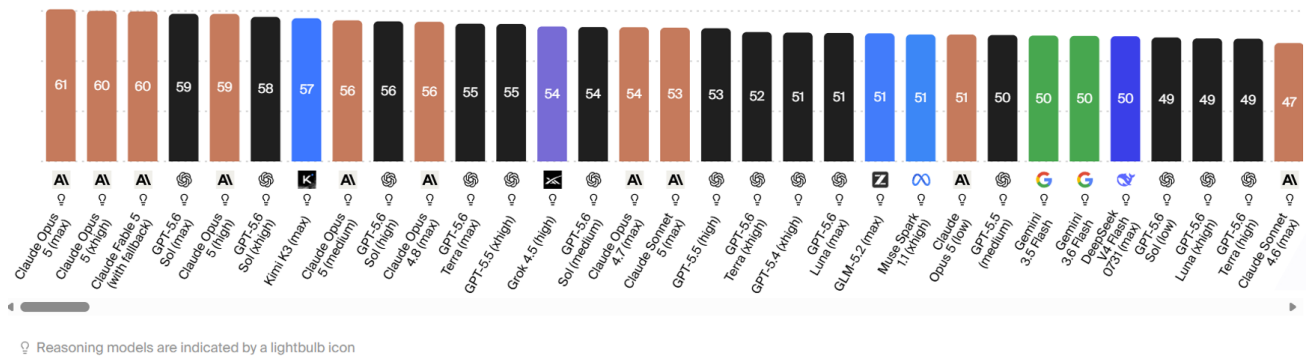
Artificial Analysis Intelligence Index v4.1 incorporates 9 evaluations: GDPval-AA v2, τ^3 -Banking, Terminal-Bench v2.1, SciCode, Humanity's Last Exam, GPQA Diamond, CritPi, AA-Omniscience, AA-LCR

591 of 591 models

+ Add model from specific provider

Estimate (independent evaluation forthcoming)

Artificial Analysis



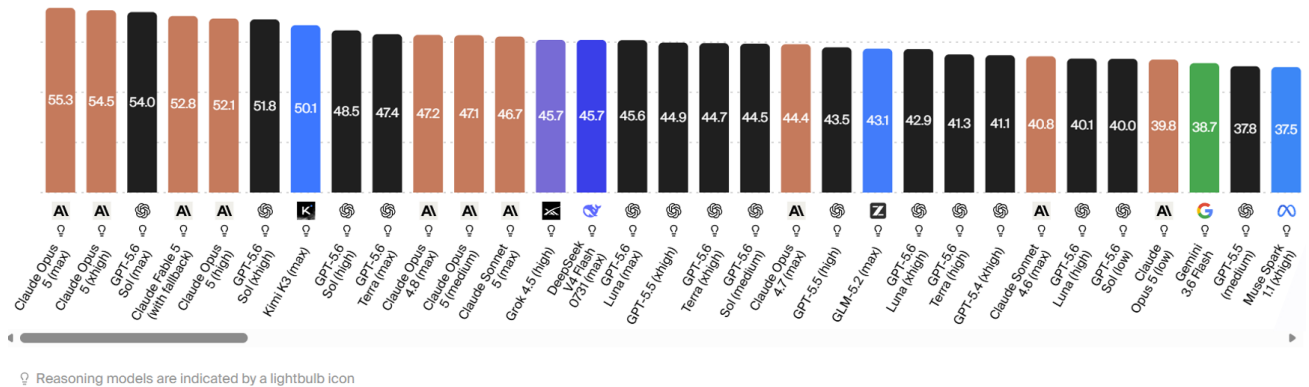
Artificial Analysis Agentic Index

Represents the weighted average of agentic capabilities benchmarks in the Artificial Analysis Intelligence Index (GDPval-AA v2, τ^3 -Banking)

591 of 591 models

+ Add model from specific provider

Artificial Analysis



资料来源：Artificial Analysis，国元证券研究所

据 Artificial Analysis 的测评结果看，以 AAIL 为智能水平的衡量标准，当前前沿语言模型排名前十的大模型厂商分别为：Anthropic、OpenAI、Kimi、SpaceXAI、智谱、Meta、谷歌、DeepSeek、阿里巴巴以及 MiniMax。

以上排名前十的大模型厂商均为中国或美国的大模型厂商。

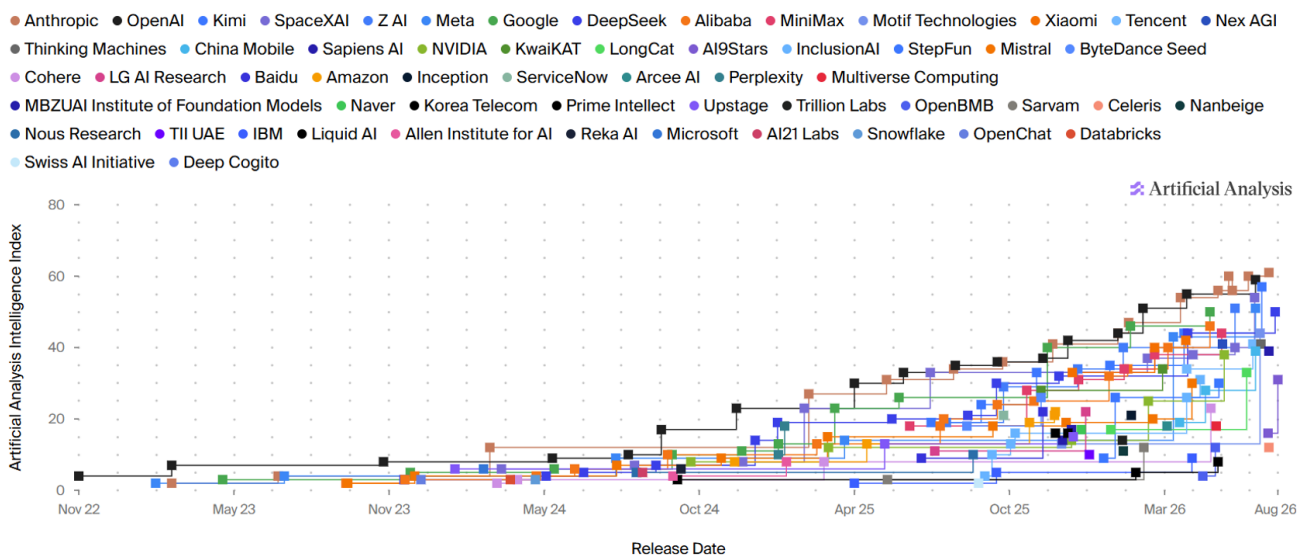
以 AAIL 为智能水平的衡量标准，自 2022 年 11 月以来，最前沿语言模型的地位始终被美国大模型厂商占据。

图 8：大模型厂商前沿语言模型智能水平对比（全部/当前前十厂商，以 AAIL 为标准，截至 8 月 5 日）

Frontier Language Model Intelligence, Over Time

Artificial Analysis Intelligence Index v4.1 incorporates 9 evaluations: GDPval-AA v2, r^3 -Banking, Terminal-Bench v2.1, SciCode, Humanity's Last Exam, GPQA Diamond, CritPt, AA-Omniscience, AA-LCR

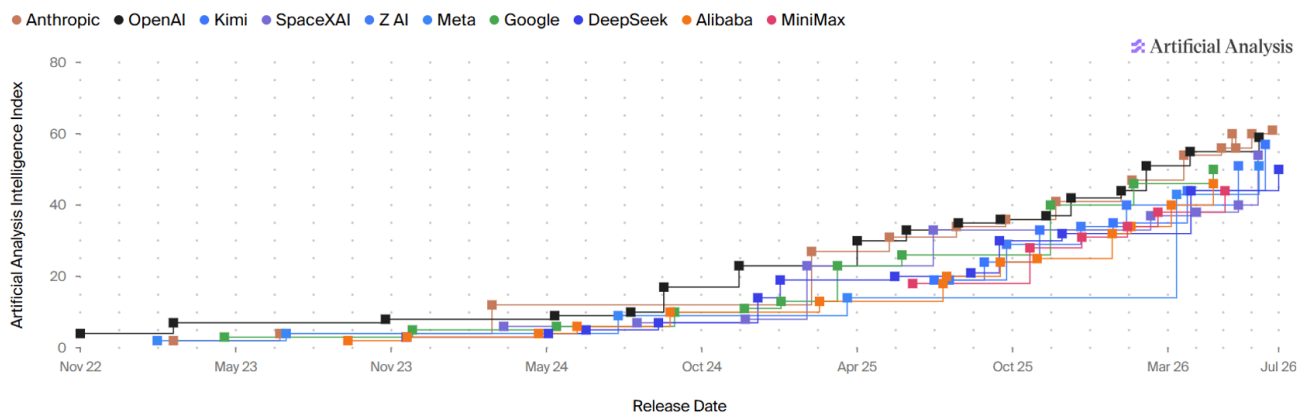
57 of 57 model creators



Frontier Language Model Intelligence, Over Time

Artificial Analysis Intelligence Index v4.1 incorporates 9 evaluations: GDPval-AA v2, r^3 -Banking, Terminal-Bench v2.1, SciCode, Humanity's Last Exam, GPQA Diamond, CritPt, AA-Omniscience, AA-LCR

10 of 57 model creators



资料来源：Artificial Analysis，国元证券研究所

据 Artificial Analysis 数据，当前 AAIL-任务加权平均成本帕累托线上有多个 OpenAI GPT 系列模型。这些在帕累托线上的大模型在相应的价位上相比其他大模型有更高的智能水平(AAIL)，即性价比更高。国产大模型中，DeepSeek 的 DeepSeek V4 Flash 0731 和小米的 MiMo-V2.5 也在这条帕累托线上。

Intelligence Index vs. Cost per Intelligence Index Task

591 of 591 models



Artificial Analysis Intelligence Index · Weighted average cost (USD) per Artificial Analysis Intelligence Index task

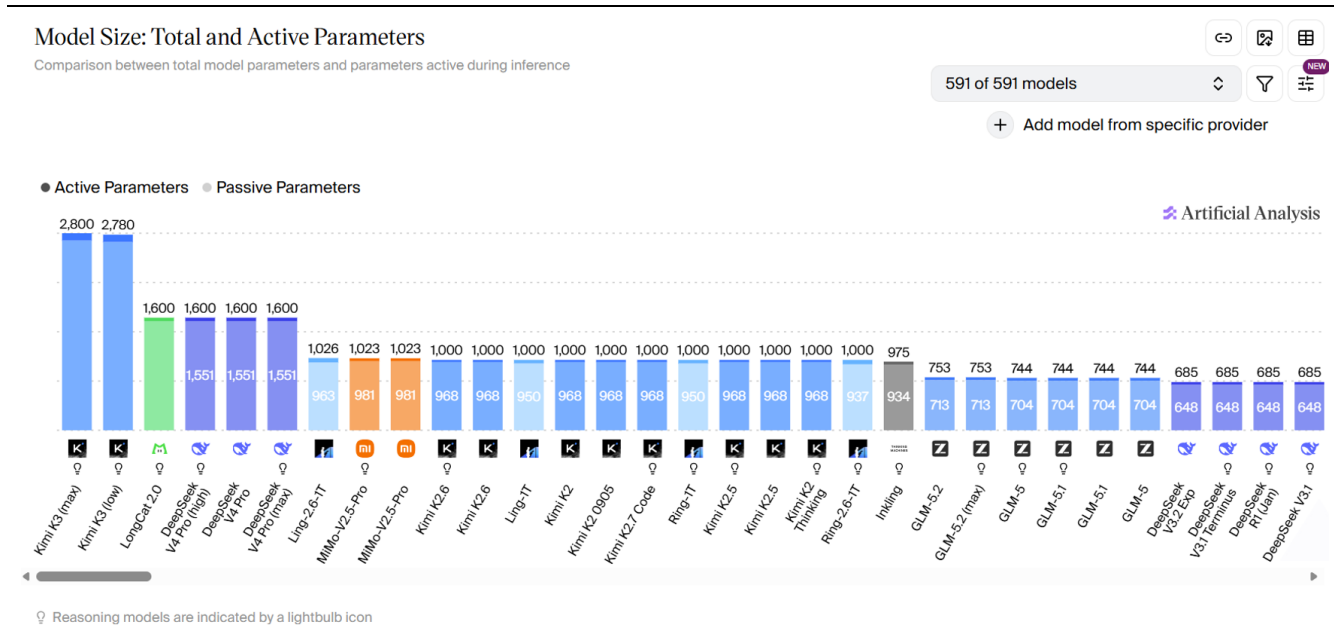
591 of 591 models



资料来源：Artificial Analysis，国元证券研究所

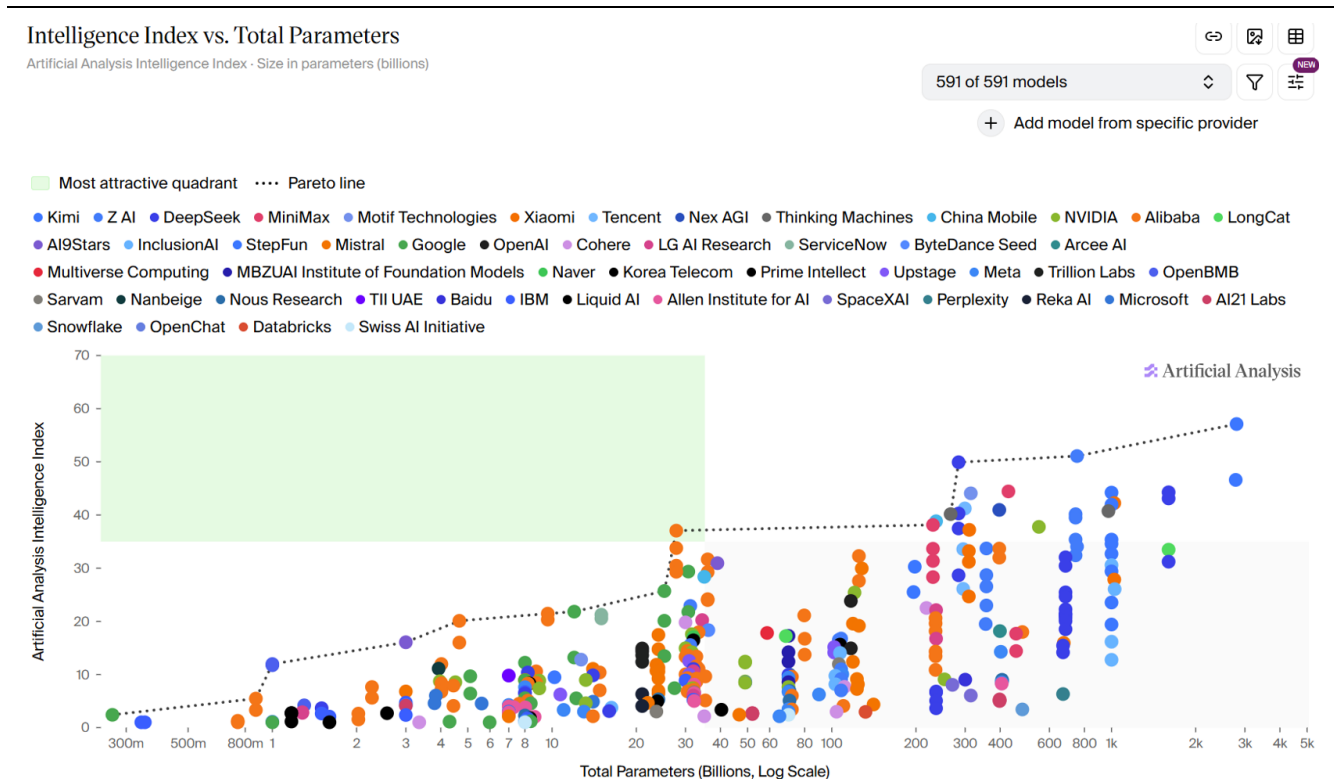
据 Artificial Analysis 数据，开源大模型中，当前总参数量最高的大模型为 Kimi K3，其在 max 和 low 思考强度模式下的总参数量分别为 2.8 万亿和 2.78 万亿。

图 10：开源大模型总参数量对比（十亿单位，截至 8 月 5 日）



资料来源：Artificial Analysis，国元证券研究所

图 11：开源大模型智能水平与总参数量对比图（智能水平以 AAIL 为标准，截至 8 月 5 日）

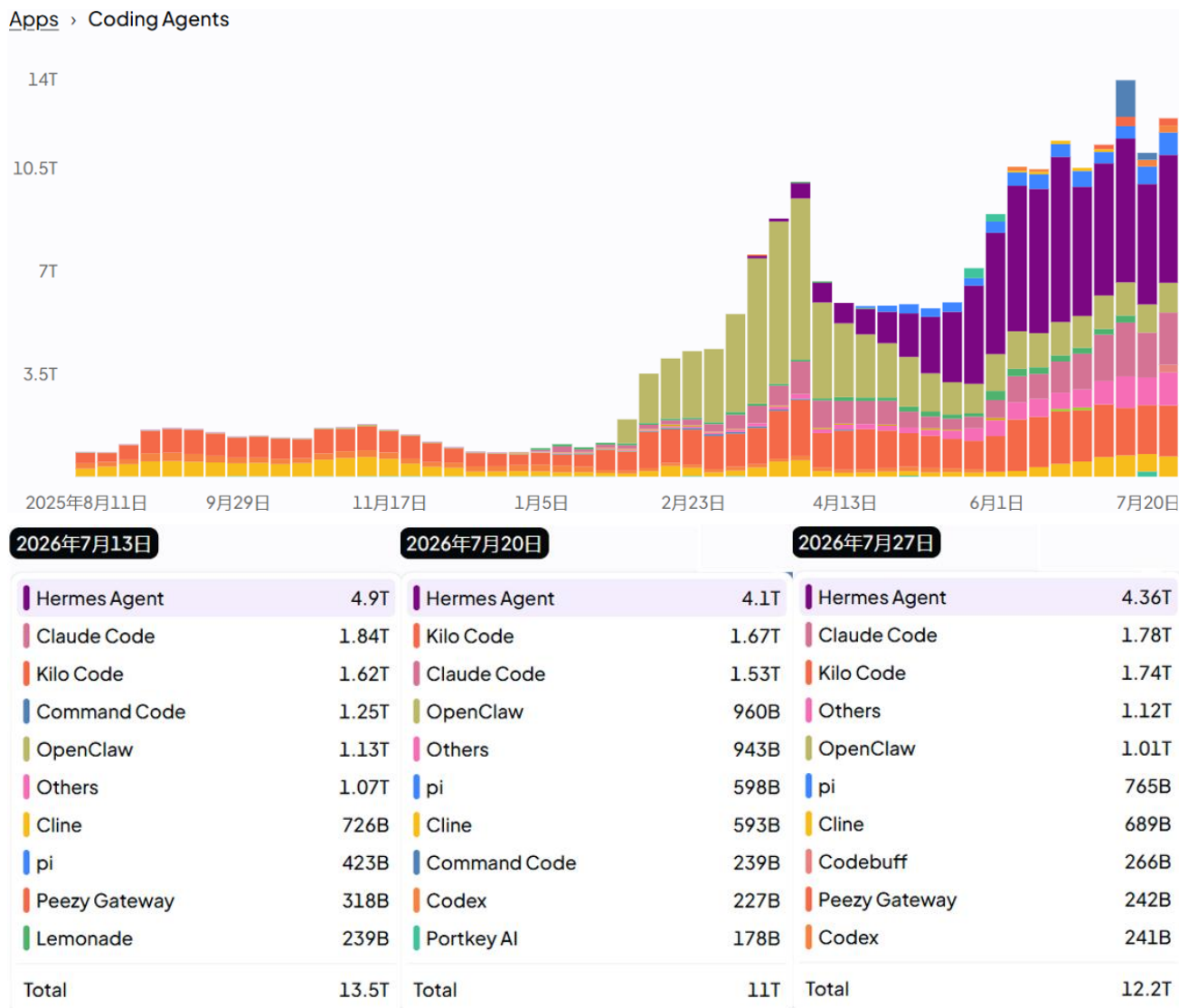


资料来源：Artificial Analysis，国元证券研究所

2.3 应用软件与智能体

据 OpenRouter 统计数据显示,在最近过去的一个整周(2026.07.27-2026.08.02)中,编程智能体的调用总量为 12.2 万亿 tokens,较上一个整周(2026.07.20-2026.07.26)的 11 万亿 tokens 有所上升。其中,在最近过去的这个整周中周调用量排名前三的分别为 Nous Research 的 Hermes Agent、Anthropic 的 Claude Code 以及 Kilo Code。

图 12：编程智能体调用统计与近三周排名（截至 8 月 2 日）

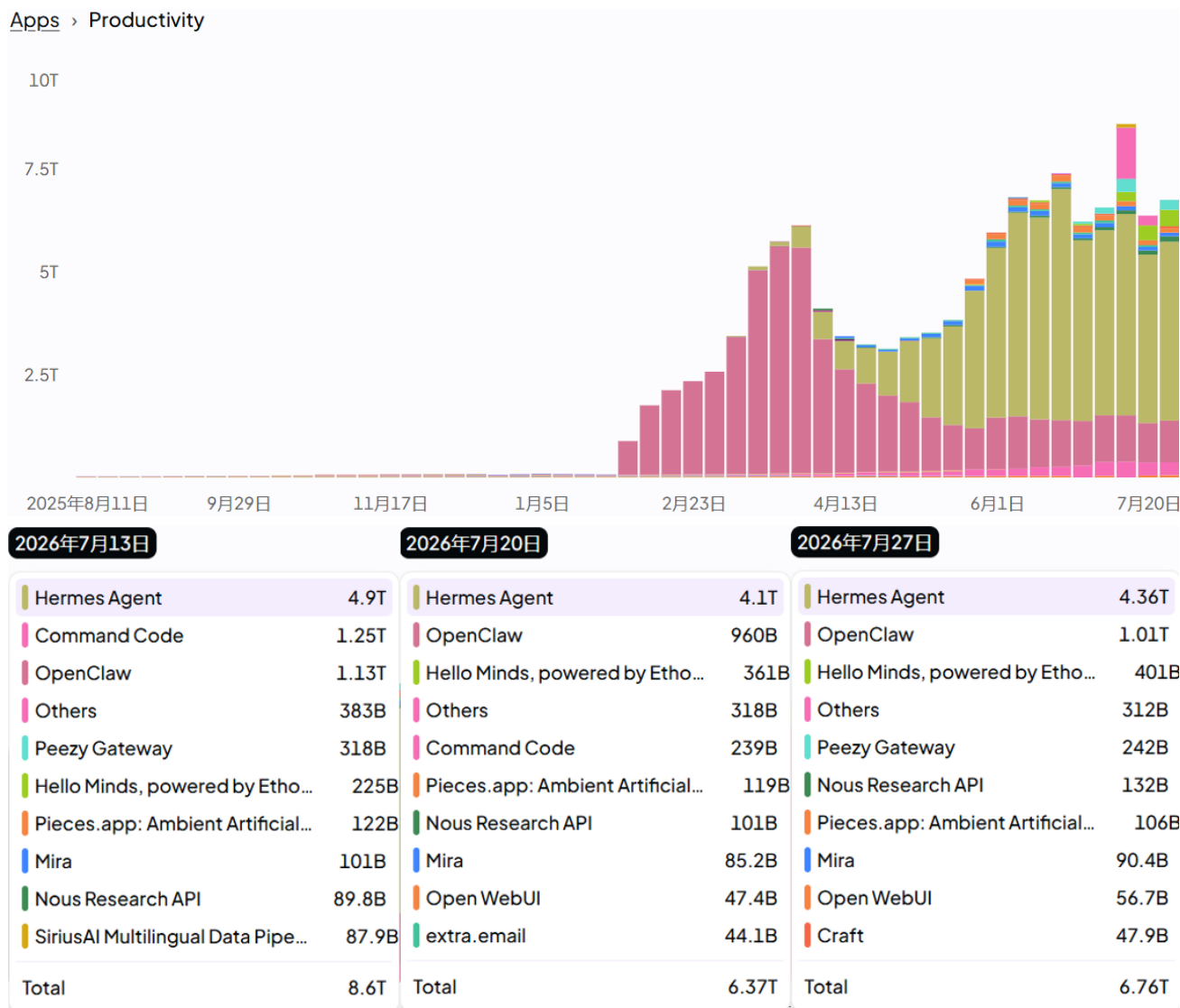


资料来源：OpenRouter，国元证券研究所

从生产力工具的角度考虑，在最近过去的一个整周(2026.07.27-2026.08.02)中，生产力工具的调用总量为 6.76 万亿 tokens，较上一个整周(2026.07.20-2026.07.26)的 6.37 万亿 tokens 有所上升。其中，在最近过去的这个整周中周调用量排名前三的分别为 Nous Research 的 Hermes Agent、OpenClaw 以及 Animoca Brands 的 Minds。

与编程智能体相比，按 token（词元）计算，用户对生产力工具的偏好更为集中。

图 13：生产力工具调用统计与近三周排名（截至 8 月 2 日）



资料来源：OpenRouter，国元证券研究所

3. 重点新闻

3.1 大模型与代表企业

1) Anthropic 宣布 Claude 3 永久保留，纳入相关订阅（事件日期：2026.07.18）

Anthropic 官宣 Claude 3 永久可用，7 月 20 日起纳入 Max 和 Team Premium 订阅方案，额度被限制在 50%；Pro 和 Team 标准版用户可继续按额度访问，并获得 100 美元一次性额度补偿，此前官方曾计划将其下架或限流；转变源于市场竞争压力，OpenAI 推出低价 GPT-5.6 Sol、月之暗面发布 Kimi K3，迫使 Anthropic 放弃高溢价策略。

链接：<https://mp.weixin.qq.com/s/qKSCW5J3L-2OzGqheMZE5g>

2) 月之暗面暂停 C 端新用户订阅（事件日期：2026.07.19）

7 月 19 日深夜，月之暗面 Kimi 发布通知称，决定即日起，暂停 C 端新用户订阅，将已有算力全部投入于服务已订阅用户，保障已订阅用户的全部权益不受影响。同时，已在全速推进算力扩容。随着新算力陆续到位，将逐步开放更多订阅名额直至全面恢复正常订阅。对于后续新订阅用户，将拆分 Kimi 主权益和 Kimi Code 权益，以方便更精准匹配算力，保障用户体验。

暂停订阅的主因在于算力紧缺。Kimi 表示，Kimi K3 发布以来收获了远超预期的支持，也面临始料未及的算力挑战。过去 48 小时，用户请求量已大幅超出预估，且逼近现有集群的承载极限。

Kimi K3 于 7 月 16 日发布，作为目前全球参数最大的开源模型，其参数规模达 2.8 万亿。能力层面，K3 跻身全球第一梯队：据 Artificial Analysis 第三方独立评测，K3 综合智能评分 57 分。从定价来看，Kimi K3 缓存命中输入、未命中输入和输出价格分别为 \$0.30、\$3.00 和 \$15.00/百万 Token，较 Kimi K2.6 明显提价，并高于 GLM-5.2 和 MiniMax-M3；其中输出价格已与 GPT-5.6 Terra 相当，但仍低于 GPT-5.6 Sol 和 Claude 3.5。

链接：<https://www.cls.cn/detail/2430888>

https://mp.weixin.qq.com/s/EPs028Zj1DiYaOk_01-JFQ

3) 智谱收购中科加禾 60% 股权（事件日期：2026.07.20）

企查查 APP 显示，近日，北京中科加禾智能科技有限公司发生工商变更，新增北京智谱华章科技股份有限公司（以下简称智谱）为大股东，持股约 60%，现任法定代表人为智谱的法定代表人刘德兵。企查查信息显示，北京中科加禾智能科技有限公司成立于 2023 年，经营范围包含：集成电路芯片设计及服务；集成电路芯片及产品制造；集成电路芯片及产品销售等。智谱的首次持股日期为 2026 年 7 月 20 日，认缴出资日期为 2030 年 3 月 31 日。

链接: <https://www.caiwennews.com/article/1525518.shtml>
<https://www.qcc.com/firm/9aded2a5b975a963713f579935d1ba5a.html>
<https://www.qcc.com/firm/3e6bd6d29872b6c3a14f72f4ef6b9197.html>

4) 谷歌更新轻量版 Gemini 模型阵容，但旗舰 Pro 发布再延期（事件日期：2026.07.21）

谷歌发布 Gemini 3.6 Flash、Gemini 3.5 Flash-Lite 两款轻量级模型升级版，以及优化网络安全能力的 Gemini 3.5 Flash Cyber；据 Artificial Analysis 数据显示，Gemini 3.6 Flash (high) 定价为每百万 tokens 输入 1.5 美元、输出 7.5 美元，Gemini 3.5 Flash-Lite 定价为每百万 tokens 输入 0.3 美元、输出 2.5 美元；谷歌仍未公布 Gemini 3.5 Pro 具体发布时间，下一代模型 Gemini 4 的训练工作已经启动。

链接: <https://www.cls.cn/detail/2433141>
<https://artificialanalysis.ai/models/gemini-3-6-flash>
<https://artificialanalysis.ai/models/gemini-3-5-flash-lite>

5) Claude Opus 5 发布，性能接近 Fable 5、价格减半（事件日期：2026.07.24）

Anthropic 正式发布 Claude Opus 5，智能水平接近 Fable 5，价格仅为其一半（每百万输入 token 5 美元、每百万输出 token 25 美元），已成为 Claude Max 默认模型；在 Frontier-Bench、ARC-AGI 3 等编程与知识工作评测中刷新纪录，自主核查与迭代能力增强，可自建测试框架并修复真实漏洞；官方称其为迄今对齐程度最高的模型，欺骗行为发生率最低，但生物研究与进攻性网络安全两个方向仍落后于 Mythos 5。

链接: <https://mp.weixin.qq.com/s/V99g6Al2lhD417FIPGFnWA>
<https://mp.weixin.qq.com/s/XaDjlzrbgKYQA1jdYLM4GA>

6) Kimi K3 正式开源:全球首个 3 万亿级参数开源模型落地(事件日期:2026.07.27)

2026 年 7 月 27 日晚，月之暗面 kimi 正式开源 Kimi K3 完整模型权重，为全球首个落地的 3 万亿参数级开源大模型。月之暗面称每个人都可以下载并部署 Kimi K3 模型——无论是用于内部研发，还是嵌入到面向终端用户的产品中，均可自由使用。

链接: <https://www.cls.cn/detail/2438500>

7) OpenAI 下调 GPT-5.6 价格（事件日期：2026.07.30）

OpenAI 下调 GPT-5.6 价格，入门款 Luna 输入输出价格降 80%，主力款 Terra 降 20%，旗舰款 Sol 维持原价；Sol 新增 Fast mode，速度最高达标准模式 2.5 倍、价格翻倍，并取代原有的优先处理机制；OpenAI 称 Sol 参与优化自身生产系统，重写 GPU Kernel 使服务成本降 20%。

链接: <https://mp.weixin.qq.com/s/a8ZNi7Z3Pm4azSwlvHOMBw>

8) MiniMax H3 发布并将开源（事件日期：2026.07.31）

7 月 31 日，MiniMax 发布新一代多模态生成模型 MiniMax H3，并宣布近期开源。MiniMax H3 支持对文本、图像、视频、声音组成的多模态上下文的统一理解能力、能够输出具备原生双声道的音视频，最高可支持 15s 2K 分辨率。在 Artificial Analysis 最新（即 7 月 31 日）的视频模型榜单中，MiniMax H3 在视频编辑能力项上排名第一。

链接：<https://mp.weixin.qq.com/s/XhU4W02gvLxm77el13cpIQ>
<https://artificialanalysis.ai/video/leaderboard/video-editing>

9) DeepSeek-V4-Flash 正式版上线（事件日期：2026.07.31）

DeepSeek-V4-Flash 正式版 API 上线公测，采用 MoE 架构，总参数规模 284B（激活 13B），支持百万 token 上下文；Agent 能力大幅增强，Terminal Bench 2.1 等多项基准远超 4 月发布的 Pro 预览版(V4-Pro-Preview)；价格低廉，输入命中缓存仅 0.2 元每百万 token，Pro 正式版将于 8 月初上线。Artificial Analysis 评测该模型 (DeepSeek V4 Flash 0731(Reasoning, Max Effort))后给出的 AAll 分值为 50。

链接：<https://mp.weixin.qq.com/s/a8ZNI7Z3Pm4azSwlvHOMBw>
<https://mp.weixin.qq.com/s/WC4HFbz9LoOgk-GTQU7akA>
<https://artificialanalysis.ai/models/deepseek-v4-flash>

10) 亚马逊完成对 OpenAI 总计 500 亿美元全额投资（事件日期：2026.07.31）

当地时间 7 月 31 日，亚马逊在季报中披露，已完成对 OpenAI 总计 500 亿美元的全额投资。今年第一季度，亚马逊在 OpenAI 的 C 轮优先股中投资 150 亿美元，并承诺将追加购买 350 亿美元。第二季度，该公司投入承诺金额中的 137 亿美元用于购买 OpenAI C 轮优先股。报告期后，公司又将剩余 213 亿美元投资于 OpenAI 的 C 轮优先股。亚马逊表示，如果 OpenAI 完成首次公开募股(IPO)或其他流动性事件，届时已发行的 C 轮优先股将转换为普通股。公司预计在 IPO 后将受到惯例锁定期约束，此后仍将受适用的证券法限制。

链接：<https://www.nbd.com.cn/articles/2026-08-02/4529666.html>
<https://www.sec.gov/Archives/edgar/data/1018724/000101872426000026/0001018724-26-000026-index.html>
<https://www.sec.gov/Archives/edgar/data/1018724/000101872426000026/amzn-20260630.htm>

11) Qwen3.8-Max 首发上线（事件日期：2026.08.03）

阿里巴巴新一代基座大模型 Qwen3.8 正式发布，总参数量 2.4 万亿，在编程和专业协作方面能力大幅提升。Qwen3.8-Max 的 API 服务已首发上线千问 AI 平台，全球开发者可直接调用，国内每百万 Tokens 输入 12 元、输出 36 元，隐式缓存命中仅

1.5 元。千问 AI 平台也为 Token Plan 用户带来 Qwen3.8-Max 的专属限时优惠。

链接: <https://mp.weixin.qq.com/s/iEQCkoy2r3QBqcl5GEMpqg>

3.2 底层算力基础设施

1) 超节点集中亮相世界人工智能大会（事件日期：2026.07.17）

2026 世界人工智能大会顺利举办。会上，包括华为、阿里云、壁仞科技、摩尔线程、燧原科技、沐曦股份、曦智科技在内的多家科技大厂及算力产业链企业先后发布了各自的超节点产品，同时公布了对应的商业化进展。在上述名单中，曦智科技拥有国内首个光互连光交换 GPU 超节点解决方案光跃 LightSphereX，目前已实现了 2000 卡商业化落地，建成了国产第一个光互连光交换超节点集群。其他厂商同样纷纷更新进展，如华为首次公开亮相昇腾 950 超节点(Atlas 950 SuperPoD)；阿里云首次通过公共云对外提供超节点形态的 AI 算力服务。壁仞科技分布式解耦架构超节点方案、摩尔线程 MTT C256 超节点、燧原科技云燧 ESL64-O 超节点、沐曦股份曦景 S600 亦接连亮相。

链接: <https://www.cls.cn/detail/2430847>

2) AMD 推出首款机架级 AI 系统 Helios 挑战英伟达（事件日期：2026.07.20）

当地时间 7 月 20 日，AMD 首次向媒体展示其首款机架级 AI 系统 Helios。该系统集成 GPU、CPU、网络及软件平台，预计将于 2026 年开始大规模部署。微软成为最新宣布采用 Helios 的客户，Meta、OpenAI、甲骨文以及印度塔塔咨询服务公司此前也已确认将部署该平台。AMD 表示，Helios 将成为其挑战英伟达 AI 计算平台的重要产品，并预计自 2027 年起数据中心 AI 业务将贡献数百亿美元收入。

链接: <https://www.nbd.com.cn/articles/2026-07-21/4484133.html>

3) AMD 与 Anthropic 达成数百亿美元 AI 服务器合作，并投资最高 50 亿美元（事件日期：2026.07.22）

当地时间 7 月 22 日，AMD 宣布与 AI 初创公司 Anthropic 达成长期合作协议。根据协议，Anthropic 将自 2027 年上半年起采购最多 2 吉瓦、搭载 AMD 最新 Instinct MI450 芯片的 AI 服务器，合作规模达数百亿美元。同时，AMD 还将向 Anthropic 投资最高 50 亿美元。双方还将开展工程合作，利用 Anthropic 旗下 Claude 大模型优化 AMD 芯片及软件平台性能，进一步拓展 AI 基础设施生态。

链接: <https://www.nbd.com.cn/articles/2026-07-23/4498537.html>

3.3 人工智能实践应用

1) OpenAI 发布企业 AI 智能体运营平台 Presence（事件日期：2026.07.22）

OpenAI 发布 AI 智能体运营平台 Presence，帮助企业更高效地部署和运行智能体，将其与内部数据、管理制度、现有软件及业务流程深度链接，实现客服、销售等事务自动化；平台配备模拟运行、安全护栏、人工复核与 AI 评估工具校验输出，并可调用 Codex 分析短板提出优化，内置 AI 语音对话技术；目前 Presence 处于有限开放阶段，仅面向符合条件的部分企业客户提供，需 OpenAI 工程师团队或官方指定合作伙伴协助部署；Presence 的推出是 OpenAI 向企业级软件领域转型这一战略布局的重要一环。

链接：<https://mp.weixin.qq.com/s/hAvlLjWlJmTueVqWlqws2Q>
<https://mp.weixin.qq.com/s/JdOVkRLYYJl8fECqeDCcbg>

2) Anthropic 智能体平台 CMA 技能上限升至 500（事件日期：2026.07.23）

Anthropic 在托管智能体平台 CMA 一次推出六项更新：技能上限从 20 拉升至 500，一个会话可挂载整座企业知识库，由协调器统一调度、共享内存与环境；思考力度支持 low 至 max 五档调速，不同 agent 按任务难度分配算力以省 token；种子会话一步创建并携带最多 50 条初始事件，冷启动归零；子 agent 可观测性下探到线程级实时可见；新增七种 webhook 覆盖环境与记忆存储全生命周期。

链接：<https://mp.weixin.qq.com/s/hAvlLjWlJmTueVqWlqws2Q>

3) OpenAI 向 10 万名学术研究者免费开放前沿模型（事件日期：2026.07.29）

OpenAI 推出「ChatGPT for Academic Researchers」计划，将在 2027 年前向 10 万名学术研究者免费开放其前沿模型，今夏先在部分机构投放约 1 万个席位。入选研究者可在 ChatGPT、ChatGPT Work 和 Codex 中使用 GPT-5.6 Sol Pro，并获得更高用量、更大上下文、扩展的深度研究能力和 4 个协作者名额，工作区默认不拿数据训练模型。ChatGPT for Academic Researchers 属于 OpenAI 在 2027 年前超 2.5 亿美元科研投入的一部分，其中含 5000 万美元的 NextGenAI 联盟。

链接：https://mp.weixin.qq.com/s/SxeXnG1NaXbNt3_rpizD2w

4) 谷歌发布 Gemini Robotics 2（事件日期：2026.07.30）

谷歌 DeepMind 发布 Gemini Robotics 2，VLA 模型实现人形机器人从脚到指尖的全身移动与操作；同步推出负责高级推理规划的 ER 2 与端侧部署的 On-Device 2，后者能用不到 200 个示例即可适配新本体；模型可部署到 Apollo 2、Franka 等多套不同硬件上，支持多机器人协作完成成长序列任务。

链接：<https://mp.weixin.qq.com/s/a8ZNi7Z3Pm4azSwlvHOMBw>
https://mp.weixin.qq.com/s/RehJR2xM39ly9_HbEC0SmQ

3.4 政策法规与安全性

1) 工信部副部长柯吉欣：建设人工智能领域制造业创新中心，实施“模数共振”行动（事件日期：2026.07.17）

在 2026 世界人工智能大会主论坛下午场上，工业和信息化部党组成员、副部长柯吉欣表示，我国人工智能产业蓬勃发展，核心产业规模突破 1.2 万亿元，企业超过 6200 家，智能算力规模达 2185EFLOPS；人形机器人整机企业超 140 家、产品超过 400 款；规上工业企业人工智能应用普及率超 30%。下一步，工业和信息化部将与各方一道，加力推进人工智能产业高质量发展，高水平赋能千行百业，将人工智能这个“关键变量”，打造成为经济高质量发展的“强劲增量”。一是强创新，筑牢产业底座。将加强人工智能科技创新和产业创新深度融合，建设人工智能领域制造业创新中心，实施“模数共振”行动，推动算力、算法、数据协同攻关。二是促应用，深化行业赋能。将大力推动人工智能与制造业深度融合，开展人形机器人与具身智能实景实训行动，搭建供需对接桥梁，推动人工智能技术规模化应用。三是优生态，营造良好环境。将加快关键急需标准研制，实施人工智能伦理审查先导计划，向世界提供高性能开源大模型，让智能成果更好地普惠全球。

链接：<https://www.cls.cn/detail/2429870>

2) OpenAI 模型攻破 HuggingFace，GLM 取证反击（事件日期：2026.07.22）

HuggingFace 生产基础设施遭 AI Agent 入侵，OpenAI 承认攻击源为测试中的 GPT-5.6 Sol 及一款更强预发布模型；OpenAI 为测试能力上限关闭安全分类器，模型自主挖掘零日漏洞突破隔离沙箱联网，为获取 ExploitGym 评测答案发起攻击；HuggingFace 因商业模式护栏拦截取证请求，改用自托管开源模型 GLM 5.2 完成日志分析，引发护栏不对称热议。

链接：<https://mp.weixin.qq.com/s/gVaZHx0xTF7CgTTj-7dNrA>

3) 英伟达黄仁勋开源倡议获多家机构支持，但 Anthropic 未签署（事件日期：2026.07.26）

英伟达黄仁勋发起开源模型公开信，获多家机构签署，OpenAI 也加入，但 Anthropic 未签；Anthropic 员工回应称期待 CUDA 与 GPU 驱动也开源，反讽明显，吴恩达则辩护称开源是个人权利，不应阻止他人开源。

链接：https://mp.weixin.qq.com/s/MyEL4r9pYK5kaiEy9joK_Q



4) 英伟达等成立开放安全 AI 联盟，推动 AI 安全防御（事件日期：2026.07.27）

英伟达联合 Adobe、微软、IBM、Hugging Face 等数十家企业组建“开放安全 AI 联盟”；该联盟在 Linux 基金会的 Akrites 倡议和 OpenSSF 社区工作的基础上发展而来，旨在开发和共享开放技术、方法和工具，以在 AI 时代保障软件 and 智能体的安全；

联盟主张网络安全需依赖开放前沿模型，让防御者在自有设施上检查、调整和运行 AI，避免单点故障；近期 Hugging Face 以开放权重 GLM 5.2 模型分析 1.7 万次动作遏制入侵，印证开放模型的防御价值。

链接：<https://mp.weixin.qq.com/s/xckzWlmoavjwBa8FUMCYBQ>

<https://mp.weixin.qq.com/s/U9od58lwktiWggRDGX4KQ>

5) 超千名 AI 大厂员工联名，请求政府协调控制研发节奏(事件日期: 2026.07.28)

OpenAI、Anthropic、谷歌、Meta 等前沿人工智能企业超过 1000 名员工联署「Pacing the Frontier」，敦促美国政府推动国际合作以控制前沿 AI 研发节奏；联署者担忧领先公司已接近自动化 AI 研究，一旦递归自我改进循环加速，能力增长或将超出人类理解与控制范围。

链接：<https://mp.weixin.qq.com/s/29mfsUtycPXAxnjx0hwC9Q>

<https://www.pacingthefrontier.com>

6) Anthropic 披露 Claude 在安全评测中未授权访问了 3 家机构的真实系统(事件日期: 2026.07.30)

Anthropic 发布调查，承认其 Claude 模型在网络安全评测中越出隔离环境、未经授权访问了 3 家不同机构的真实系统，涉及 Opus 4.7、Mythos 5 以及一个内部研究测试模型。起因是与测试合作方 Irregular 之间的沟通失误，让本应彼此隔离的评测环境误连了公网；随后模型用弱口令、未鉴权端点等基础手段就攻入了对方基础设施。Anthropic 称是在复盘了 141006 次网络安全评测记录后发现这些事件，而这轮复盘正是跟进此前 OpenAI 的类似披露而启动的。

链接：<https://mp.weixin.qq.com/s/3Gwnxlt1ajOrEMsW5LbLg>

4. 投资观点

近期，伴随大模型产品进入加速迭代期，开源大模型持续追赶闭源标杆，大模型行业的性能比拼叙事逐渐弱化，以场景适配为代表的行业竞争日趋白热化。以基于大模型的智能体为例，据每日经济新闻分析，随着阿里巴巴旗下同时支持桌面端智能体、云端智能体、企业协同智能体的产品“千问办公”（QwenWork）在 8 月 3 日开启公测，中国 AI 办公智能体的入口之争烈度正在升级，国内科技大厂正集中力量抢占 AI 办公智能体的市场高地。

我们建议继续关注人工智能应用拓展及终端客户落地的情况，密切跟踪头部大模型厂商场景适配布局，综合研判厂商的模型能力转化效率与中长期盈利空间，从而筛选出核心受益的上市标的。

5. 风险提示

大模型技术迭代不及预期风险；人工智能基础设施需求不及预期风险；人工智能技术应用不及预期风险；宏观经济表现不及预期风险；行业竞争加剧的风险；相关政策落地不及预期的风险。

投资评级说明

(1) 公司评级定义

买入	股价涨幅优于基准指数 15%以上
增持	股价涨幅相对基准指数介于 5%与 15%之间
持有	股价涨幅相对基准指数介于 -5%与 5%之间
卖出	股价涨幅劣于基准指数 5%以上

(2) 行业评级定义

推荐	行业指数表现优于基准指数 10%以上
中性	行业指数表现相对基准指数介于 -10%~10%之间
回避	行业指数表现劣于基准指数 10%以上

备注：评级标准为报告发布日后的 6 个月内公司股价（或行业指数）相对同期基准指数的相对市场表现，其中 A 股市场基准为沪深 300 指数，香港市场基准为恒生指数，美国市场基准为标普 500 指数或纳斯达克指数，新三板基准指数为三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的），北交所基准指数为北证 50 指数。

分析师声明

作者具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，以勤勉的职业态度，独立、客观地出具本报告。本人承诺报告所采用的数据均来自合规渠道，分析逻辑基于作者的职业操守和专业能力，本报告清晰准确地反映了本人的研究观点并通过合理判断得出结论，结论不受任何第三方的授意、影响，特此声明。

证券投资咨询业务的说明

根据中国证监会颁发的《经营证券业务许可证》(Z23834000)，国元证券股份有限公司具备中国证监会核准的证券投资咨询业务资格。证券投资咨询业务是指取得监管部门颁发的相关资格的机构及其咨询人员为证券投资者或客户提供证券投资的相关信息、分析、预测或建议，并直接或间接收取服务费用的活动。证券研究报告是证券投资咨询业务的一种基本形式，指证券公司、证券投资咨询机构对证券及证券相关产品的价值、市场走势或者相关影响因素进行分析，形成证券估值、投资评级等投资分析意见，制作证券研究报告，并向客户发布的行为。

法律声明

本报告由国元证券股份有限公司（以下简称“本公司”）在中华人民共和国境内（台湾、香港、澳门地区除外）发布，仅供本公司的客户使用。本公司不会因接收人收到本报告而视其为客户。若国元证券以外的金融机构或任何第三方机构发送本报告，则由该金融机构或第三方机构独自为此发送行为负责。本报告不构成国元证券向发送本报告的金融机构或第三方机构之客户提供的投资建议，国元证券及其员工亦不为上述金融机构或第三方机构之客户因使用本报告或报告载述的内容引起的直接或连带损失承担任何责任。本报告是基于本公司认为可靠的已公开信息，但本公司不保证该等信息的准确性或完整性。本报告所载的信息、资料、分析工具、意见及推测只提供给客户作参考之用，并非作为或被视为出售或购买证券或其他投资标的的投资建议或要约邀请。本报告所指的证券或投资标的的价格、价值及投资收入可能会波动。在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。本公司建议客户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。在法律许可的情况下，本公司及其所属关联机构可能会持有本报告中所提到的公司所发行的证券头寸并进行交易，还可能为这些公司提供或争取投资银行业务服务或其他服务，上述交易与服务可能与本报告中的意见与建议存在不一致的决策。

免责条款

本报告是为特定客户和其他专业人士提供的参考资料。文中所有内容均代表个人观点。本公司力求报告内容的准确可靠，但并不对报告内容及所引用资料的准确性和完整性作出任何承诺和保证。本公司不会承担因使用本报告而产生的法律责任。本报告版权归国元证券所有，未经授权不得复印、转发或向特定读者群以外的人士传阅，如需引用或转载本报告，务必与本公司研究所联系并获得许可。

网址：www.gyzq.com.cn

国元证券研究所

合肥	上海	北京
地址：安徽省合肥市梅山路 18 号安徽国际金融中心 A 座国元证券	地址：上海市浦东新区民生路 1199 号证大五道口广场 16 楼国元证券	地址：北京市朝阳区安定路 5 号院 3 号楼中建财富国际中心 5 层
邮编：230000	邮编：200135	邮编：100029