

板块持续调整，前沿大模型快速迭代

推荐|维持

——人工智能行业半月报（2026 年 9 月第 1 期）

报告要点：

● 指数表现（2026 年 9 月上半月）

内地指数：中证人工智能产业指数下跌 6.58%，跌幅有所增加。国证 AI 应用指数下跌 7.12%，上证科创板人工智能指数下跌 9.39%。

香港指数：恒生港股通人工智能主题指数下跌 14.22%，跌幅有所增加。

● 数据跟踪

据 OpenRouter 数据显示，在刚过去的一个整周(09.07-09.13)中，按 token 计算的大模型周调用量为 127 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，9 月 17 日，以 AAll 为标准，Anthropic Claude 与 OpenAI GPT 系列模型的智能水平总体上领先；当前智能水平最高的国产大模型是 Qwen3.8 Max (0902)；OpenAI 有个性性价比高的 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 发布 Fable 5.1 和 Mythos 5.1；谷歌发布 Gemini 3.8 Flash；Meta 发布新旗舰模型 Muse Spark 1.3；OpenAI 发布 GPT-6 Astra；OpenAI 发布 Images 2.5；DeepSeek 发布 V4.1 Flash；Kimi 发布 K2.8 Preview；智谱公告配售新 H 股，并建议同步发行零息可换股债券。

底层算力基础设施：三星公布下一代存储器愿景；英伟达 129 亿美元收购 Hugging Face；阿斯麦与台积电、三星与英特尔推进新一代 EUV 光刻技术应用；Arm 发布新一代 AI 原生计算平台；谷歌在芬兰投资 130 亿欧元用于 AI 基础设施；美光推出 512GB DDR5 模组。

人工智能运用：英伟达发布 DLSS 5；腾讯 WorkBuddy 开放平台上线，开放 Agent 底座；千问办公推出多人工作台；Meta 推出个人 AI 智能体 Muse；谷歌 DeepMind 发布 AlphaGenome Atlas，覆盖 90 亿种变异；宇树公开具身智能模型 UnifoLM-WLA-1.0；OpenAI 发布四条产品线，其中 Agents API 开放公测。

政策法规与安全性：十部门：推动普惠算力赋能中小企业发展，降低中小企业用算成本；最高法发布《关于依法审理涉人工智能纠纷案件的意见》；工信部：打造高效协同的算力设施；工信部印发《“人工智能+软件”专项行动实施方案》。

● 投资观点

我们建议继续围绕人工智能应用端与基础设施端布局：应用端，重点关注已初步验证商业模式，或在中长期具备较清晰商业闭环前景的上市公司；基础设施端，优先关注能有效承接算力需求增长，产品与服务迭代有序推进，且盈利提升能见度较高的上市公司。

● 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 8 月第 1 期）：大模型加速迭代，AI 应用百花齐放》2026.08.19

《国元证券行业研究-人工智能行业半月报（2026 年 7 月第 2 期）：Kimi K3 发布，引领开源大模型发展》2026.08.05

《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.....	18
3.4 政策法规与安全性.....	20
4. 投资观点.....	22
5. 风险提示.....	22

图表目录

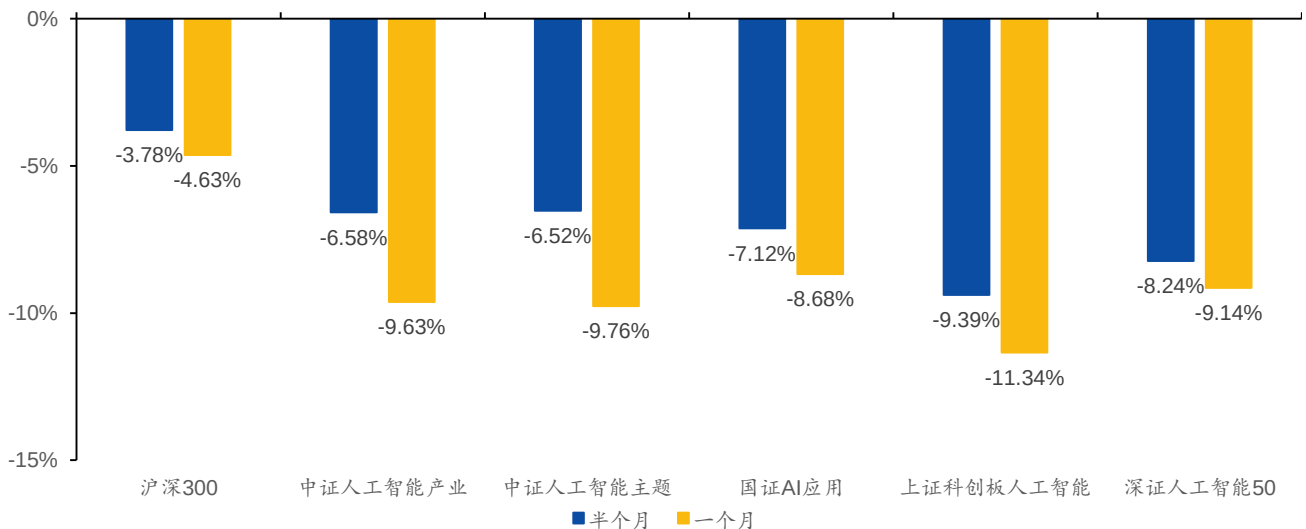
图 1: 沪深 300 指数及人工智能概念指数半月度与月度涨跌幅.....	3
图 2: 沪深 300 指数及人工智能概念指数 2026 年初至今累计涨跌幅.....	3
图 3: 恒生系列核心指数及恒生港股通人工智能主题指数半月度与月度涨跌幅.....	4
图 4: 恒生系列核心指数及恒生港股通人工智能主题指数 2026 年初至今累计涨跌幅.....	4
图 5: 大模型调用统计与近三周排名（截至 9 月 13 日）.....	5
图 6: 大模型厂商市场份额统计与近三周排名（以文本请求为衡量标准，截至 9 月 13 日）.....	6
图 7: 大模型智能水平综合排名（以 AAIL 为标准，截至 9 月 17 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 9 月 17 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 9 月 17 日）.....	9
图 10: 开源大模型总参数量对比（十亿单位，截至 9 月 17 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 9 月 17 日）.....	10
图 12: 编程智能体调用统计与近三周排名（截至 9 月 13 日）.....	11
图 13: 生产力工具调用统计与近三周排名（截至 9 月 13 日）.....	12

1. 指数表现

1.1 内地股票相关指数

2026 年 9 月 1 日至 15 日，中证人工智能产业指数下跌 6.58%，跌幅相比 8 月下半月有所扩大，跑输沪深 300 指数 2.80 个百分点。同一时间段内，国证 AI 应用指数下跌 7.12%，上证科创板人工智能指数下跌 9.39%，深证人工智能 50 指数下跌 9.14%。

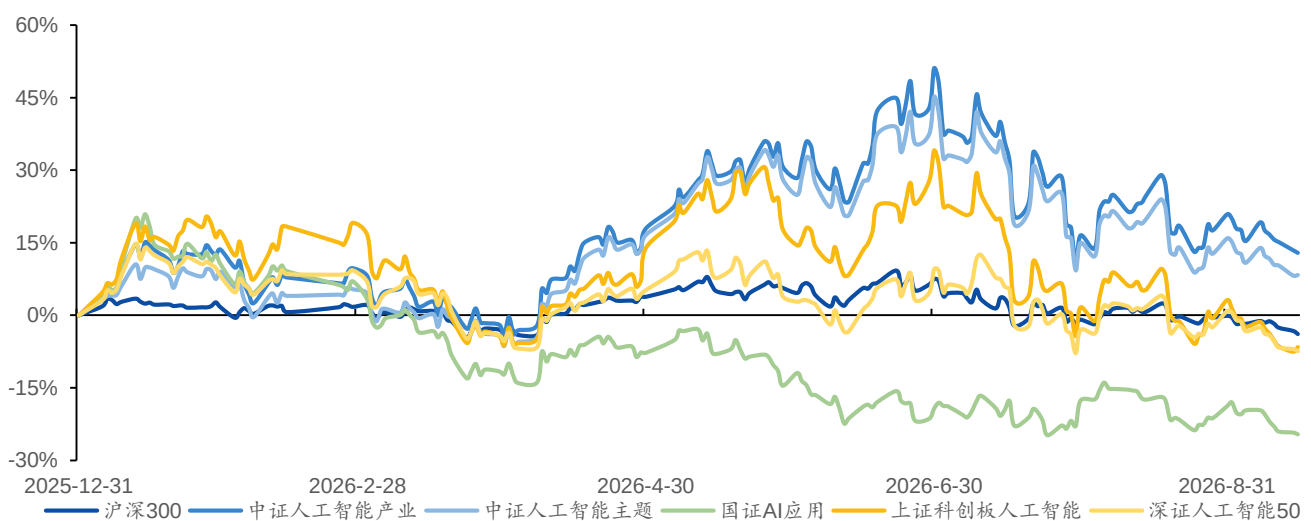
图 1：沪深 300 指数及人工智能概念指数半月度与月度涨跌幅



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 12.90%，跑赢沪深 300 指数 16.79 个百分点。国证 AI 应用指数累计下跌 24.59%，上证科创板人工智能指数累计下跌 6.56%。

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

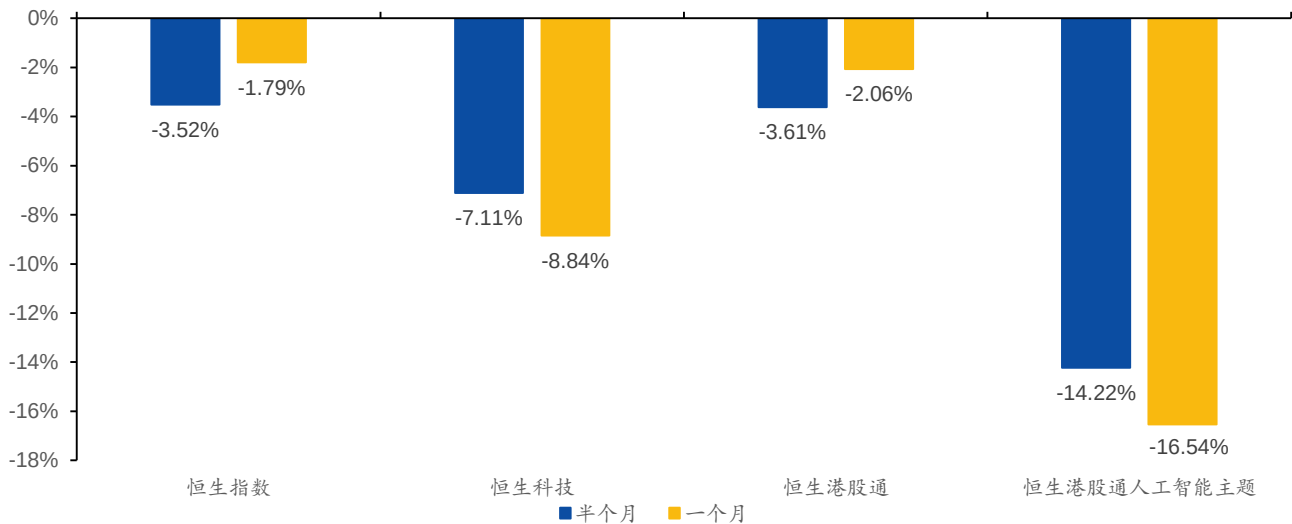


资料来源：iFinD，国元证券研究所

1.2 香港股票相关指数

2026 年 9 月 1 日至 15 日，恒生港股通人工智能主题指数下跌 14.22%，跌幅相比 8 月下半月有所增加，跑输恒生指数 10.71 个百分点。同一时间段内，恒生科技指数下跌 7.11%，恒生港股通指数下跌 3.61%。

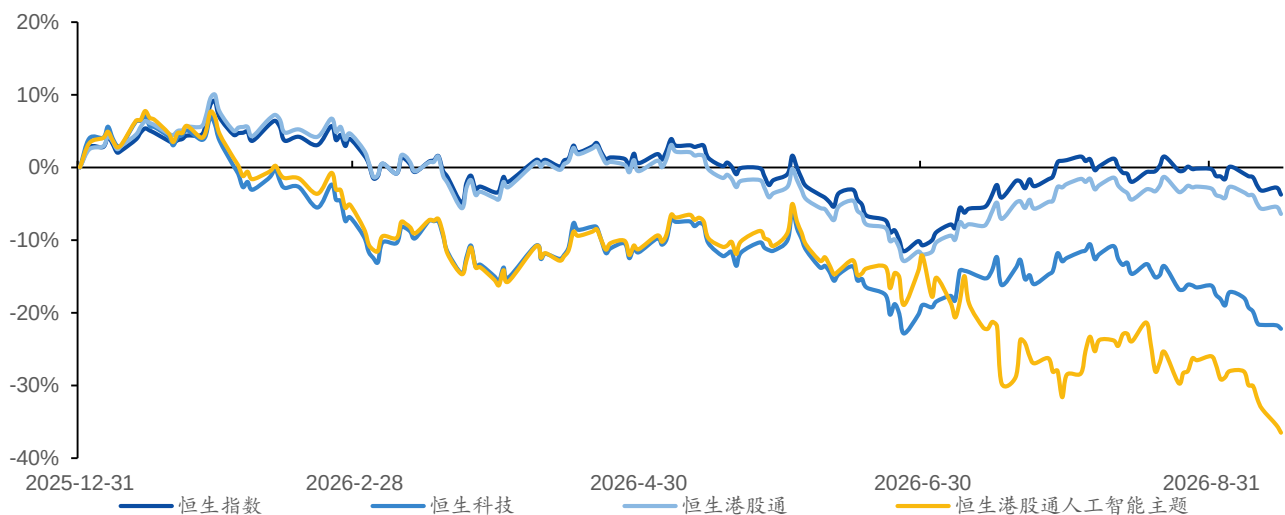
图 3：恒生系列核心指数及恒生港股通人工智能主题指数半月度与月度涨跌幅



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 36.52%，跑输恒生指数 32.76 个百分点。恒生科技指数累计下跌 22.20%，恒生港股通指数累计下跌 6.40%。

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



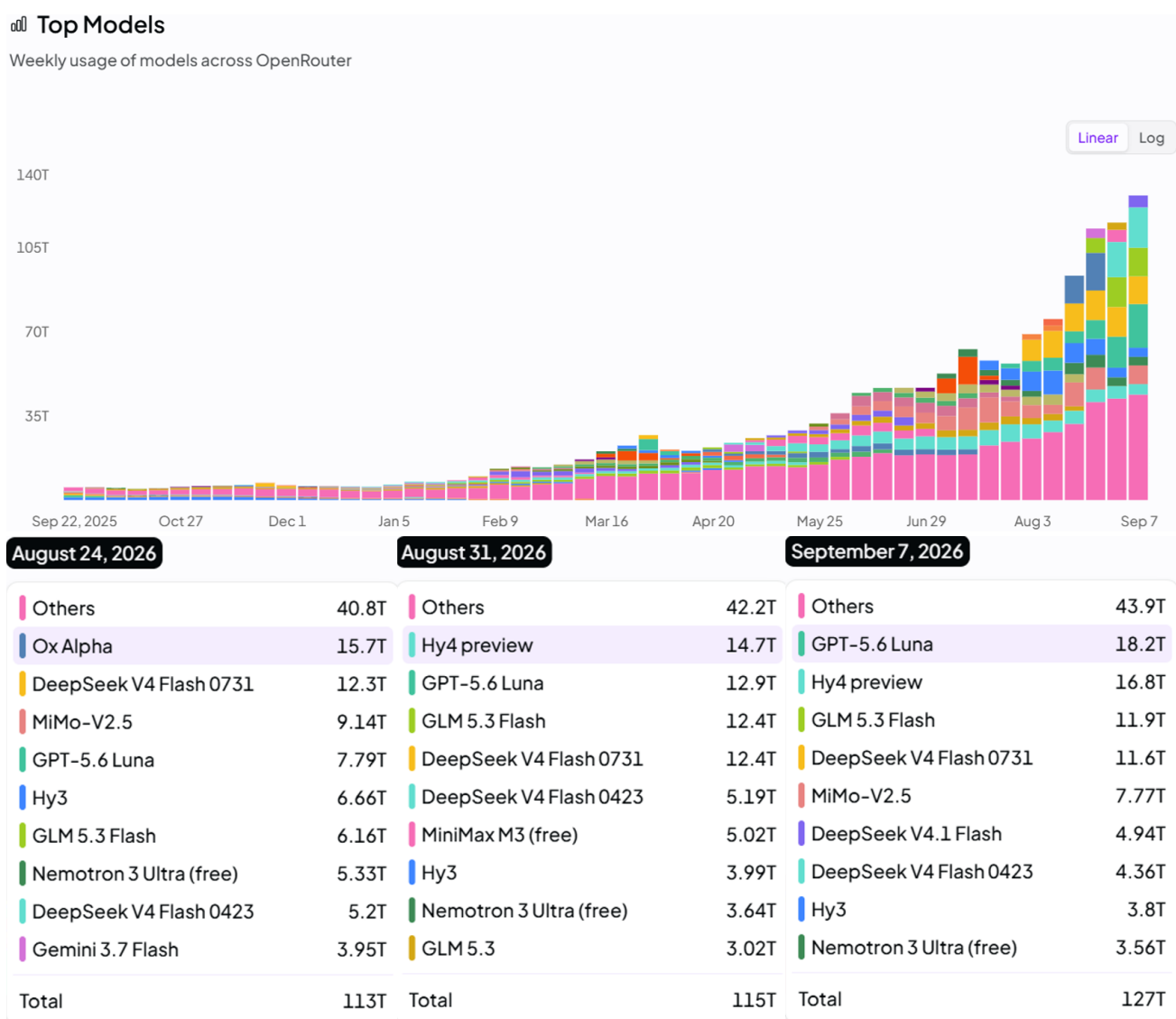
资料来源：iFinD，国元证券研究所

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(2026.09.07-2026.09.13)中继续创新高，达到 127 万亿 tokens，较上一个整周(2026.08.31-2026.09.06)的 115 万亿 tokens 有所上升。其中，该周调用排名前五的大模型分别为：OpenAI 的 GPT-5.6 Luna、腾讯的 Hy4 preview、智谱的 GLM 5.3 Flash、DeepSeek 的 DeepSeek V4 Flash 0731 以及小米的 MiMo-V2.5。

图 5：大模型调用统计与近三周排名（截至 9 月 13 日）

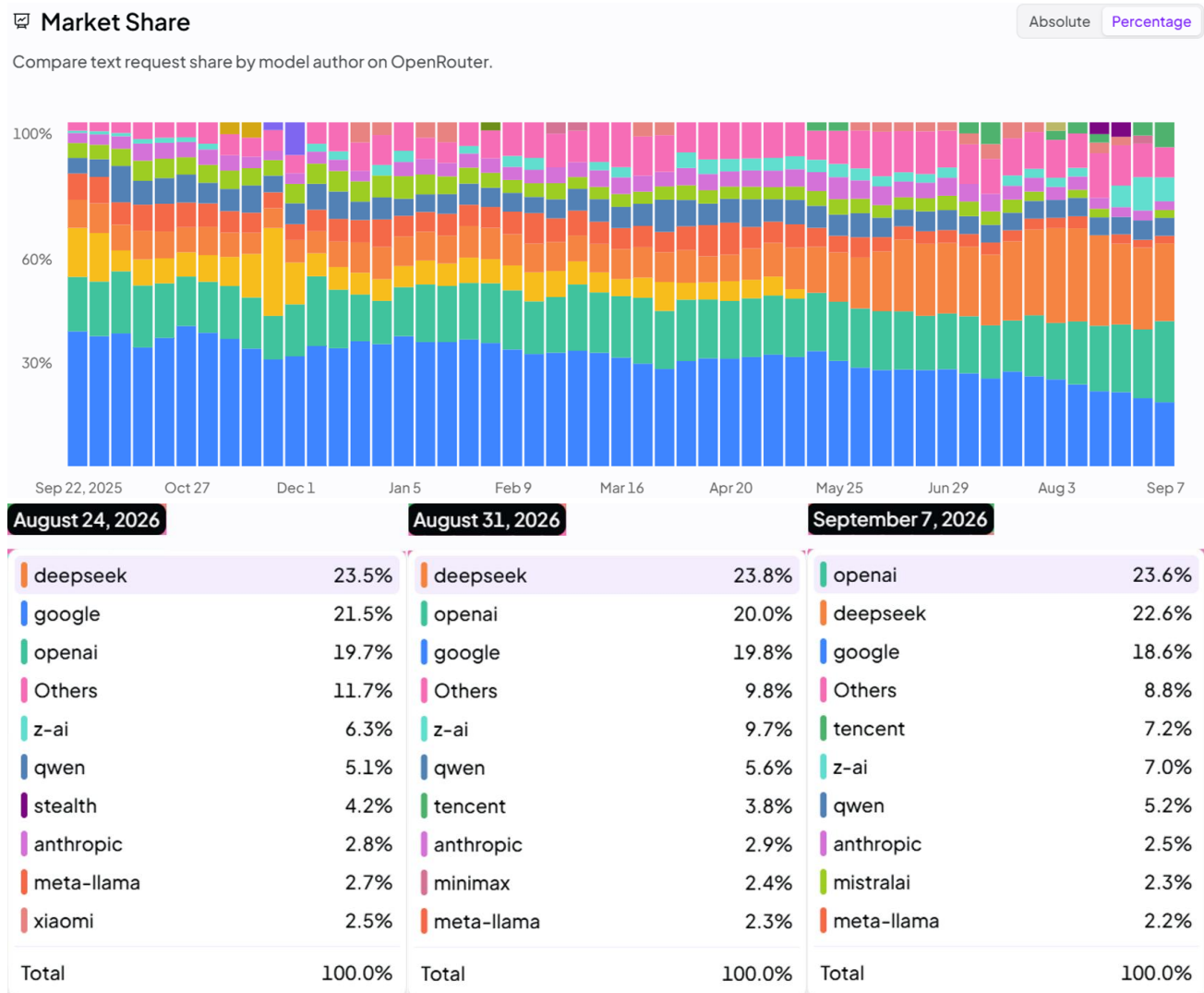


资料来源：OpenRouter，国元证券研究所

以文本请求为衡量标准，在最近过去的一个整周(2026.09.07-2026.09.13)，市场份额排名首位的是 OpenAI(23.6%)，排名第二的是 DeepSeek(22.6%)。

这意味着，DeepSeek 连续五周位居市场份额榜首的纪录就此终结。谷歌、腾讯以及智谱在相应市场份额上排名三到五位。

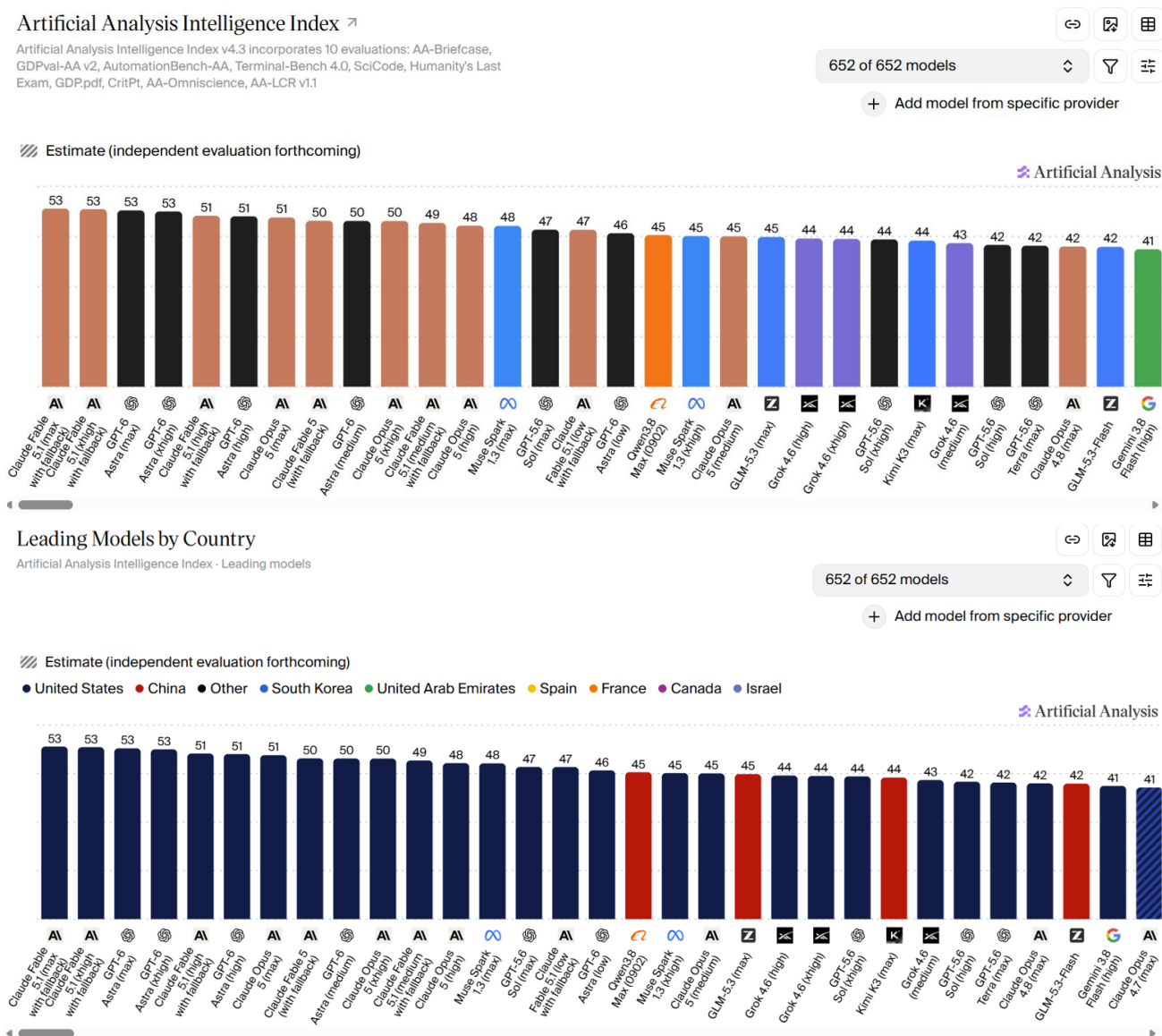
图 6：大模型厂商市场份额统计与近三周排名（以文本请求为衡量标准，截至 9 月 13 日）



资料来源：OpenRouter，国元证券研究所

综合 Artificial Analysis 的测评结果看,9月17日,Anthropic Claude 与 OpenAI GPT 系列模型的智能水平总体上领先;当前智能水平排名前三的国产大模型分别为:Owen3.8 Max (0902)、GLM-5.3(max)以及 Kimi K3(max)。

图 7: 大模型智能水平综合排名 (以 AAIL 为标准, 截至 9 月 17 日)



资料来源：Artificial Analysis，国元证券研究所

据 Artificial Analysis 的测评结果看，9 月 17 日，以 AAIL 为智能水平的衡量标准，前沿语言模型排名前十的大模型厂商分别为：Anthropic、OpenAI、Meta、阿里巴巴、智谱、SpaceXAI、Kimi、谷歌、DeepSeek 以及 Sapiens AI。

新加坡大模型厂商 Sapiens AI 挤进前十，成为排名前十的大模型厂商中唯一中美以外的大模型厂商。

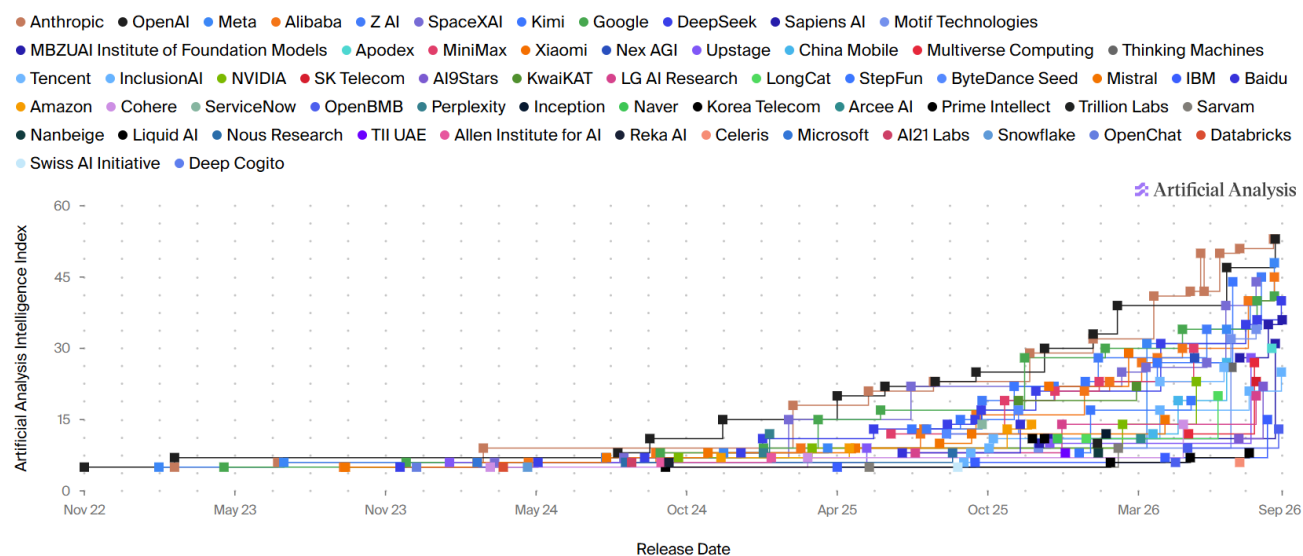
以 AAIL 为智能水平的衡量标准，自 2022 年 11 月以来，最前沿语言模型的地位始终被美国大模型厂商占据。

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 9 月 17 日）

Frontier Language Model Intelligence, Over Time

Artificial Analysis Intelligence Index v4.3 incorporates 10 evaluations: AA-Briefcase, GDPval-AA v2, AutomationBench-AA, Terminal-Bench 4.0, SciCode, Humanity's Last Exam, GDP.pdf, CritPt, AA-Omniscience, AA-LCR v1.1

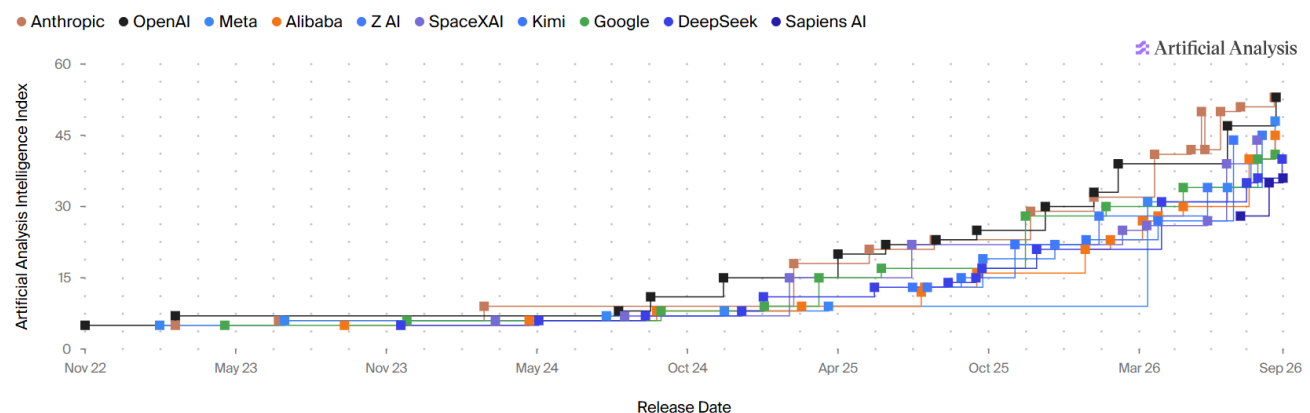
59 of 59 model creators



Frontier Language Model Intelligence, Over Time

Artificial Analysis Intelligence Index v4.3 incorporates 10 evaluations: AA-Briefcase, GDPval-AA v2, AutomationBench-AA, Terminal-Bench 4.0, SciCode, Humanity's Last Exam, GDP.pdf, CritPt, AA-Omniscience, AA-LCR v1.1

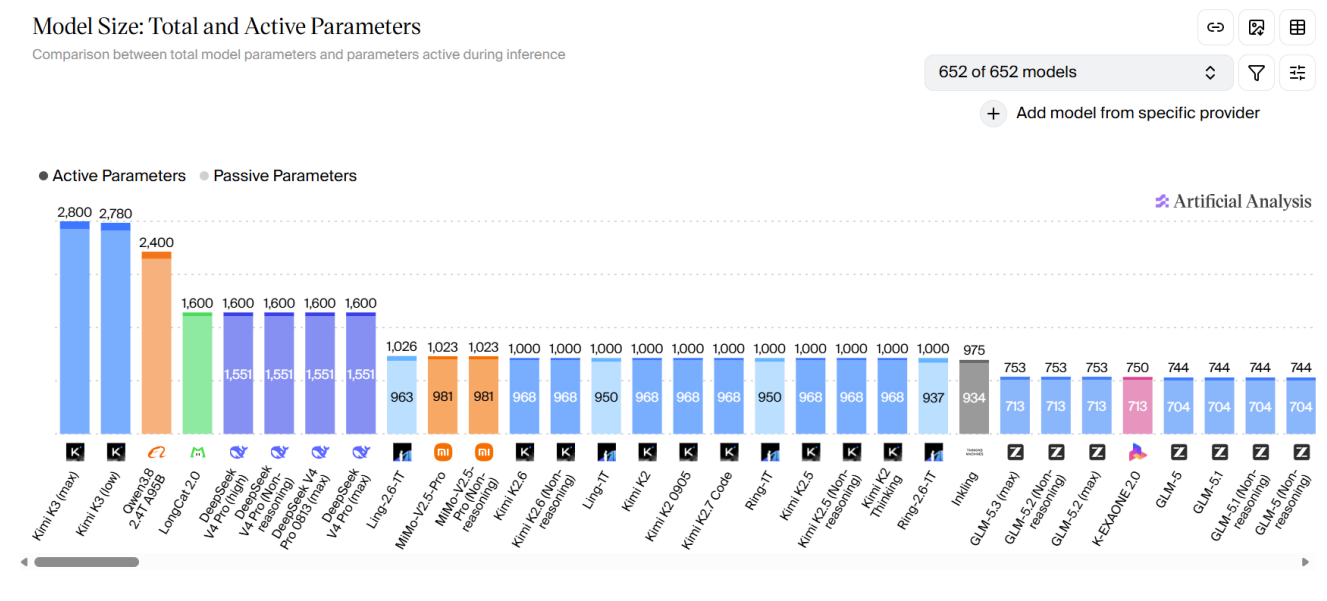
10 of 59 model creators



资料来源：Artificial Analysis，国元证券研究所

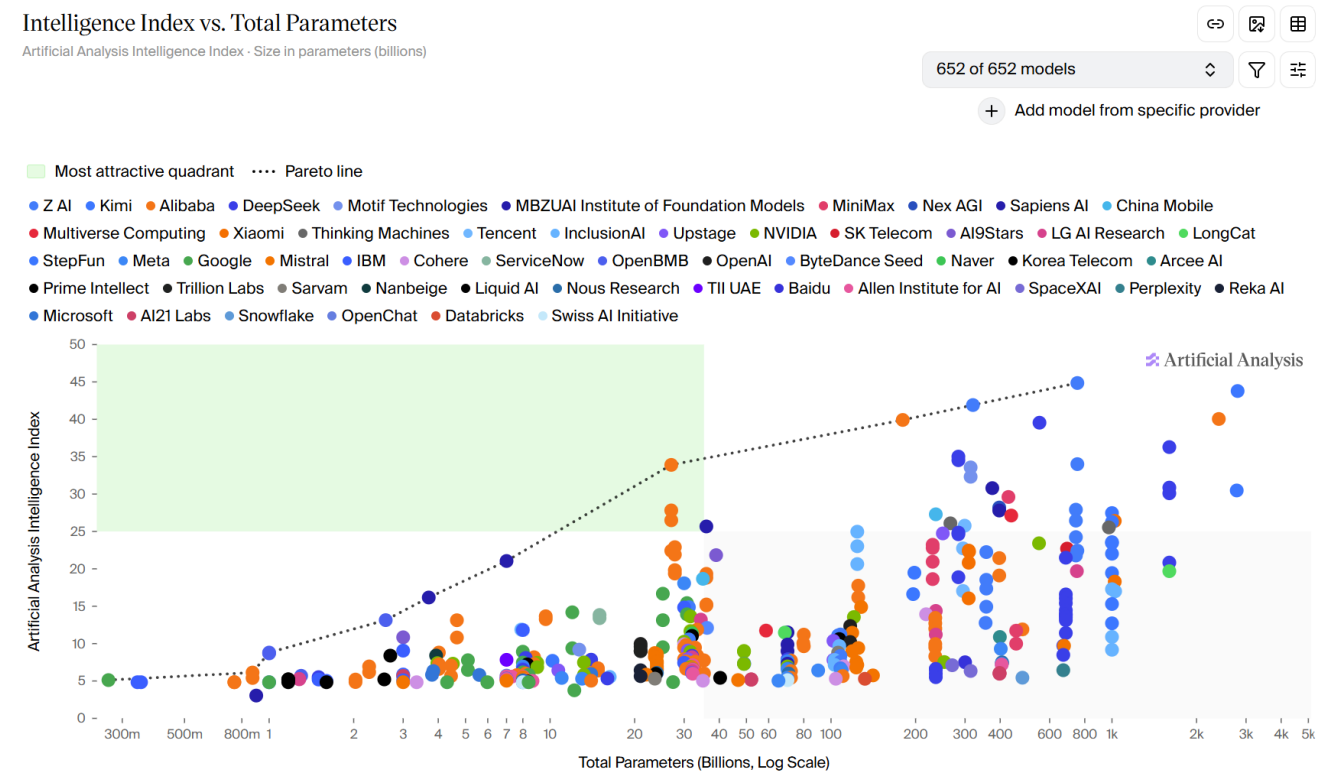
据 Artificial Analysis 数据，开源大模型中，当前总参数量最高的大模型为 Kimi K3，其在 max 和 low 思考强度模式下的总参数量分别为 2.8 万亿和 2.78 万亿。

图 10：开源大模型总参数量对比（十亿单位，截至 9 月 17 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 9 月 17 日）

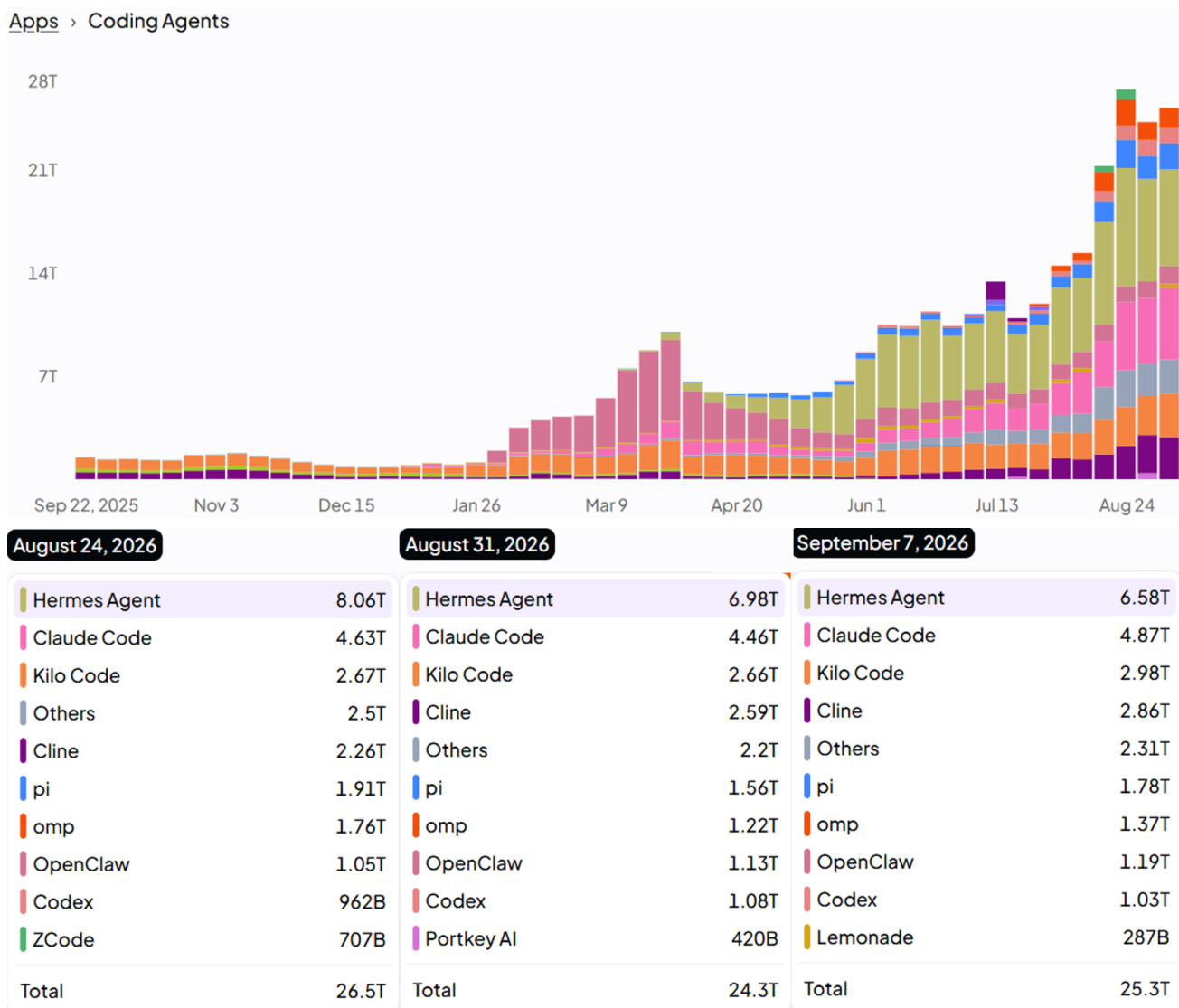


资料来源：Artificial Analysis，国元证券研究所

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(2026.09.07-2026.09.13)中,编程智能体的调用总量为 25.3 万亿 tokens,较上一个整周(2026.08.31-2026.09.06)的 24.3 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 9 月 13 日)

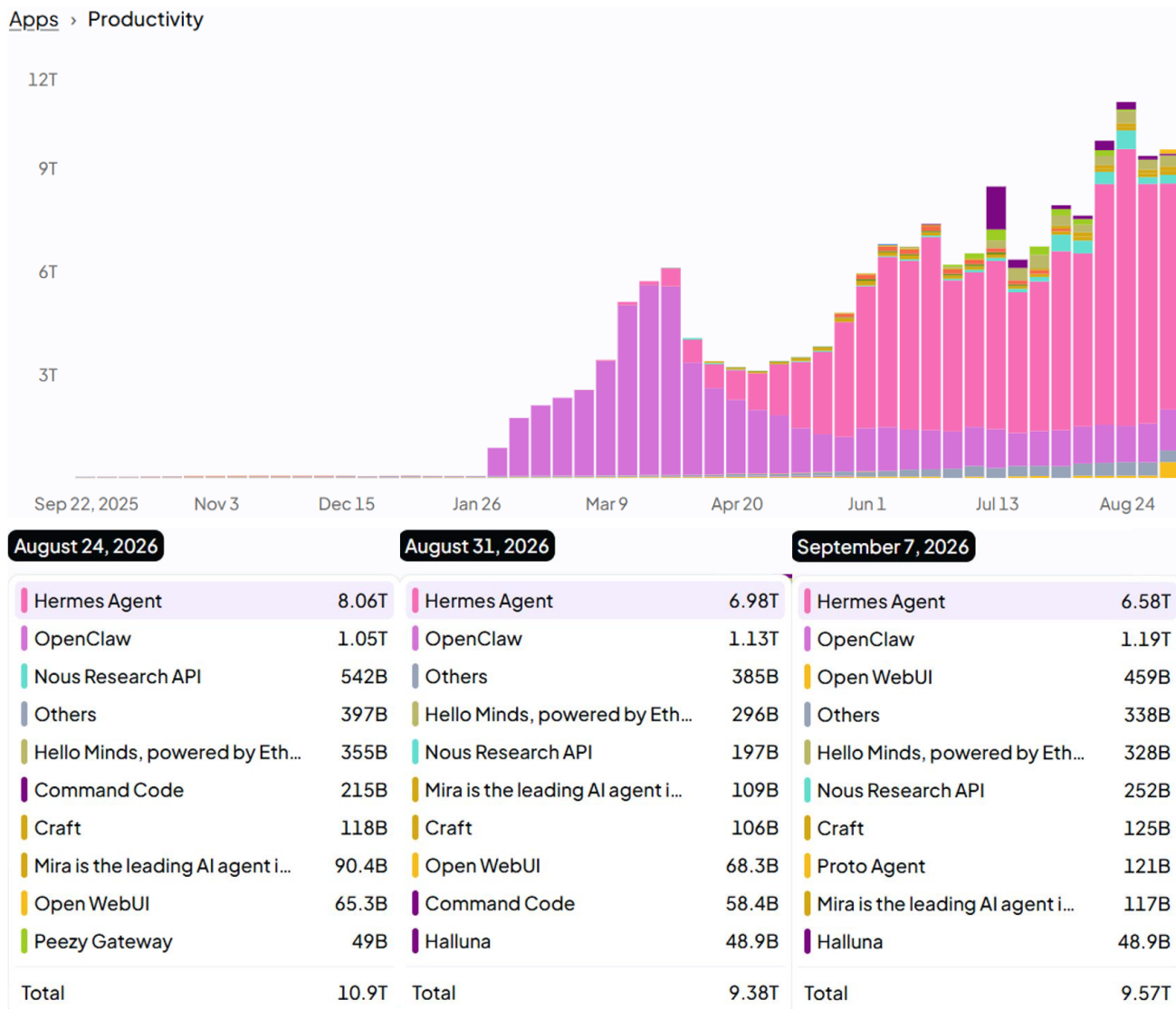


资料来源: OpenRouter, 国元证券研究所

从生产力工具的角度考虑，在最近过去的一个整周(2026.09.07-2026.09.13)中，生产力工具的调用总量为 9.57 万亿 tokens，较上一个整周(2026.08.31-2026.09.06)的 9.38 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 Open WebUI。

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 9 月 13 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 发布 Fable 5.1 和 Mythos 5.1 (事件日期: 2026.09.01)

Anthropic 发布新一代旗舰模型 Fable 5.1, 并面向拥有受信任访问计划权限的用户推出 Mythos 5.1。Anthropic 称, 新模型进一步提升了编程、知识工作和长时间任务求解能力, 回应了客户关于价格、数据保留和安全措施方面的反馈。

Anthropic 降低了 Fable 5.1 缓存读取的价格: 缓存读取成本降低了 75%, 即每百万 token 缓存读取 0.25 美元。除此之外, Fable 5.1 的定价与 Fable 5 相同: 每百万 token 输入 10 美元、输出 50 美元。

Artificial Analysis 测评结果显示, Claude Fable 5.1 (Adaptive Reasoning, Max Effort, Default Fallback)的 AAll (v4.3 版) 得分为 53。

链接: <https://www.nbd.com.cn/articles/2026-09-02/4569727.html>
<https://www.anthropic.com/claude-fable-and-mythos-5-1>
<https://artificialanalysis.ai/models/claude-fable-5-1>

2) 谷歌发布 Gemini 3.8 Flash (事件日期: 2026.09.02)

谷歌推出 Gemini 3.8 Flash, 为六周内发布的第三款 Flash 系列模型, 定价与 Gemini 3.7 Flash 持平, 每百万 token 输入 0.75 美元、输出 3.75 美元; 谷歌同步推出网络安全专用版 Gemini 3.8 Flash Cyber。

Artificial Analysis 测评结果显示, Gemini 3.8 Flash (high)的 AAll (v4.3 版) 得分为 41。

链接: <https://mp.weixin.qq.com/s/fdUTcD8N7P-aB9eQmfO0-Q>
<https://artificialanalysis.ai/models/gemini-3-8-flash>

3) Meta 发布新旗舰模型 Muse Spark 1.3 (事件日期: 2026.09.02)

Meta 发布新旗舰模型 Muse Spark 1.3, 主攻长上下文与编程领域; 定价与前代持平, 每百万 token 输入 1.25 美元、缓存命中 0.15 美元、输出 4.25 美元。

Artificial Analysis 测评结果显示, Muse Spark 1.3 (max)的 AAll (v4.3 版) 得分为 48, 是截至 9 月 17 日非 Anthropic、非 OpenAI 模型中得分最高者。

链接: <https://mp.weixin.qq.com/s/fdUTcD8N7P-aB9eQmfO0-Q>
<https://artificialanalysis.ai/models/muse-spark-1-3>

4) OpenAI 发布 GPT-6 Astra (事件日期: 2026.09.03)

当地时间 9 月 3 日, OpenAI 正式发布新一代旗舰模型 GPT-6 Astra。OpenAI 表示, GPT-6 Astra 汇聚了公司多年来在预训练、强化学习和对齐领域的研究成果与重大投入, 在计算机操作、网页浏览、软件工程、网络安全、科学和专业工作领域均处于领先水平。该模型即日起面向部分组织逐步推出, 并将在未来几天内向所有 ChatGPT Plus、Pro、Business 和 Enterprise 用户开放。

OpenAI API 标准定价为每百万 token 输入 10 美元、输出 50 美元, 缓存读取适用单独费率 (OpenRouter 数据显示为每百万 token 1 美元)。GPT-6 Astra 在 API 中提供快速模式, 速度最高可达标准处理的 2 倍, 价格为标准价格的 2 倍。

Artificial Analysis 测评结果显示, GPT-6 Astra (max) 的 AAll (v4.3 版) 得分为 53。

链接: <https://openai.com/zh-Hans-CN/index/gpt-6-astra/>
<https://openrouter.ai/openai/gpt-6-astra#providers>
<https://artificialanalysis.ai/models/gpt-6-astra>

5) OpenAI 发布 Images 2.5 (事件日期: 2026.09.08)

OpenAI 发布新一代图像生成模型 ChatGPT Images 2.5, 重点提升图片细节质量、生成速度与编辑精度, 目前用户每周通过 ChatGPT 与 API 生成超 30 亿张图片; 精准编辑方面可单独调整产品、背景、文字等元素而无需重新生成整图; 新增 Sketch 手绘参考、创作模板与图片评论编辑功能; 将在 API 中发布 Flare 与 Sunburst 两款模型, 前者延迟较上代降低 50% 且生成质量更高的图像, 后者专为高端视觉 workflow 打造, 可在多次编辑中提供更精细的控制。

链接: <https://mp.weixin.qq.com/s/47WZuT9Lc2zaV8qXI9dfFw>
<https://openai.com/zh-Hans-CN/index/introducing-chatgpt-images-2-5/>

6) DeepSeek 发布 V4.1 Flash (事件日期: 2026.09.10)

DeepSeek 正式发布 552B 参数 MoE 模型 V4.1 Flash, 采用全新 Causal-Encoder-Decoder 非对称结构, 输入激活 8B、输出激活 16B, 具备原生多模态视觉理解能力; 据 DeepSeek 介绍, 新模型大幅压缩 KV Cache, 对 HBM 需求降至上一代四分之一、SSD 需求降至八分之一, 显著降低 Agent 类任务的缓存命中成本; 模型已上线 API 并同步开源, 腾讯 WorkBuddy、CodeBuddy 等作为官方合作伙伴全量接入。

Artificial Analysis 测评结果显示, DeepSeek V4.1 Flash (Reasoning, Max Effort) 的 AAll (v4.3 版) 得分为 40。

链接: https://mp.weixin.qq.com/s/5PP7Tazner724OxXpY8H_A
<https://artificialanalysis.ai/models/deepseek-v4-1-flash>

7) Kimi 发布 K2.8 Preview (事件日期: 2026.09.11)

月之暗面(Kimi)9月11日全量上线 K2.8 Preview, 官方称综合性能接近旗舰 K3, 编码与智能体能力全面提升, 思考效率较 K2.7 Code 显著改善; 新模型支持 low、high、max 三档思考强度与图片视频输入, 1M 上下文向所有会员开放。

链接: <https://mp.weixin.qq.com/s/gkbTgMCytsv6fn0lp5DzPQ>
<https://www.kimi.com/code/docs/kimi-code/whats-new.html>

8) 智谱公告配售新 H 股, 并建议同步发行零息可换股债券(事件日期: 2026.09.13)

9月13日, 智谱发布公告称根据一般授权配售新 H 股, 并建议根据一般授权同步发行人民币 201.4 亿元于 2027 年到期的美元结算零息可换股债券。股份配售价为每股 714 港元, 较 9 月 11 日收盘价 793 港元折让约 9.96%, 最多认购 21,965,000 股新 H 股; 可换股债券初始换股价为每股 892.50 港元, 较 9 月 11 日收盘价溢价约 12.55%。

智谱表示, 本轮融资所得款项净额约 60%将用于下一代 GLM 基础模型及完全自训练体系的研发, 以及大规模训练、生产推理、算力资源及相关技术基础设施的部署及升级, 旨在持续提升模型的智能上界、长程任务能力、自我改进能力、模型迭代效率及规模化服务能力; 约 15%将用于支持本公司的业务拓展计划、战略投资及潜在并购, 旨在加速本公司通用人工智能大模型的商业化落地, 拓展产业生态, 并提升其长期增长前景; 约 25%将用于优化资本结构、补充日常运营所需营运资金及其他一般企业用途, 以维持充足流动性、支持持续业务运营并保留财务灵活性。智谱预期配售事项及发行债券所得款项净额将于 2028 年 6 月 30 日前悉数动用。

链接: https://www1.hkexnews.hk/listedco/listconews/sehk/2026/0913/2026091300026_c.pdf

3.2 底层算力基础设施

1) 三星公布下一代存储器愿景（事件日期：2026.09.01）

三星电子在台湾台北举行的 Semicon Taiwan 2026 上介绍了下一代存储器愿景。三星电子正在开发第八代 HBM5，其性能是第七代高带宽存储器 HBM4E 的两倍。该公司还计划推出 zHBM，与 HBM4E 相比，其性能最高可提升至 8 倍，热阻最高可降低 90%。该公司正在筹备 zNAND-O，这是一种专为大语言模型量身定制的混合解决方案，可同时突破传统 DRAM 和 NAND 的局限，目标是在 2028 年送样。

链接：<https://en.sedaily.com/finance/2026/09/01/samsung-unveils-cube-strategy-to-break-memory-limits>

2) 英伟达 129 亿美元收购 Hugging Face（事件日期：2026.09.03）

英伟达宣布以 129.303 亿美元收购开源社区平台 Hugging Face；该平台已聚集超 1800 万开发者、300 万个模型、50 万个数据集与 100 万个应用，逾 20 万家公司在此选型、测试并落地 AI；英伟达承诺维持开放定位，继续支持多云与多加速器，开发者无需绑定英伟达算力；此前英伟达已在 Hugging Face 上发布超 500 个模型与 250 个开放数据集。

链接：<https://mp.weixin.qq.com/s/fdUTcD8N7P-aB9eQmfO0-Q>
<https://blogs.nvidia.com/blog/nvidia-to-acquire-hugging-face/>

3) 阿斯麦与台积电、三星与英特尔推进新一代 EUV 光刻技术应用（事件日期：2026.09.08）

当地时间 9 月 8 日，荷兰光刻机制造商阿斯麦(ASML)接连发布三条公告，宣布与台积电、三星电子及英特尔推进下一代高数值孔径极紫外光刻(EUV)技术应用。具体包括：阿斯麦与台积电发起了一项产业合作倡议，旨在推动向 12 英寸光掩模过渡，从而最大化高数值孔径 EUV 光刻的价值；阿斯麦与三星宣布扩大长期战略合作关系，三星加入面向高数值孔径 EUV 的 12 英寸光掩模平台行业倡议，计划在业内率先将高数值孔径 EUV 应用于未来 DRAM 制造；阿斯麦与英特尔代工合作，加速推动行业为高数值孔径 EUV 的技术应用做好准备。

链接：<https://www.nbd.com.cn/articles/2026-09-08/4576081.html>
<https://www.asml.com/en/news/press-releases/2026/tsmc-and-asml-announce-industry-transition-to-large-format-photomasks-for-high-na-euv>
<https://www.asml.com/en/news/press-releases/2026/samsung-electronics-and-asml-expand-collaboration-on-next-generation-semiconductor-manufacturing>
<https://www.asml.com/en/news/press-releases/2026/intel-foundry-and-asml-collaborate-to-accelerate-industry-readiness-for-high-na-euv>

4) Arm 发布新一代 AI 原生计算平台 (事件日期: 2026.09.08)

9月8日, Arm 宣布其正式推出第二代移动终端计算子系统(CSS for Mobile 2), 在单一计算平台中实现对智能体 AI 和 AI 原生图形技术的全面支持。Arm 表示, Arm CSS for Mobile 2 集成全新 Arm Mali G2-Ultra NX GPU、搭载 SME2 的 Arm C2 CPU 集群、增强型系统 IP、物理实现方案以及开发者就绪的软件生态系统, 并针对下一代移动设备进行了全面优化。

链接: <https://www.arm.com/zh-cn/company/news/2026/09/arm-css-for-mobile-2-agentic-ai-mobile-graphics>

5) 谷歌在芬兰投资 130 亿欧元用于 AI 基础设施 (事件日期: 2026.09.09)

谷歌宣布其计划在未来两年(2027-2028)向芬兰的数字基础设施投资至少 130 亿欧元, 以深化其对芬兰和欧洲的长期承诺。这包括在 Hamina、Kajaani、Muhos 和 Vaala 的数据中心及其他配套基础设施的投资与合作。此外, 谷歌还将支持清洁能源项目, 并设立专项自然与社区基金以支持当地生物多样性、教育、研究和劳动力发展。谷歌表示, 这是其在欧洲最大的一笔单笔投资, 也印证了芬兰在负责任地建设 AI 基础设施方面的领导地位。

链接: <https://www.googlecloudpresscorner.com/2026-09-09-Google-Deepens-Commitment-to-Finland-with-Two-Year-EUR13-Billion-investment-in-AI-Infrastructure>

6) 美光推出 512GB DDR5 模组 (事件日期: 2026.09.15)

当地时间 9 月 15 日, 美光宣布公司已在多个服务器平台上成功演示全球首款 512GB DDR5 模组, 该成果为企业数据中心及云计算客户高效应对大规模、高要求人工智能(AI)、数据分析、内存数据库及仿真密集型工作负载提供支持。领先的半导体生态合作伙伴 AMD 及英特尔正积极对该模组进行验证, 其速率至高可达 9200MT/s。

该产品采用硅通孔(TSV)技术, 将多个 DRAM 裸片垂直堆叠并实现互连, 在单个 DRAM 封装内最大化存储密度, 同时保持现代数据中心架构所需的性能。借助该模组, 单台配备 24 个内存插槽的双路服务器可实现高达 12TB 的 DDR5 DRAM 容量。

链接: <https://investors.micron.com/news/press-release/2026/Micron-Advances-Memory-Innovation-With-the-Worlds-First-Ultra-Dense-Module-for-Next-Generation-Servers/default.aspx>

<https://www.micron.cn/about/press/news/micron-launches-worlds-first-ultra-high-density-memory-module-to-power-next-generation-server-innovation>

3.3 人工智能实践应用

1) 英伟达发布 DLSS 5（事件日期：2026.09.01）

英伟达正式发布 DLSS 5, 以像素空间扩散 Transformer 实现 3D 引导神经网络渲染, 消费级 RTX 50 全系 GPU 均可运行; 把反射率、法线、光照语义、运动矢量等一并输入模型, 严格保留角色身份、场景语义与镜头构图等, 并提供开发者控制台细调结构与调性强度。

链接: <https://mp.weixin.qq.com/s/PtXZ0SVM3UyP2RdDFVTS0A>

2) 腾讯 WorkBuddy 开放平台上线, 开放 Agent 底座（事件日期：2026.09.02）

腾讯 WorkBuddy 开放平台正式上线, 首批引入超百家生态伙伴, 面向智能硬件、行业应用与开发者开放 Agent 底层能力; 已接入智能眼镜、录音卡片、耳机等十余个品类、超 30 个品牌, 首发 9 款联名硬件, 一次接入打通多端流转; 同步发布 Buddy 应用及 Skill、Expert、Connector 三大开放能力, 已覆盖金融、法律、医疗等 20 多个领域。

链接: <https://mp.weixin.qq.com/s/usECIhXKgmTcPKwknLDz6A>

3) 千问办公推出多人工作台（事件日期：2026.09.07）

千问办公推出“多人工作台”, 区别于面向个人的 AI 工作台, 该功能不仅可以生成网页, 还具备角色权限、云端数据库、管理后台和在线发布四项核心能力, 可承载多人参与的业务流程; 适用于家校协同、达人投放、连锁店铺筹备、百人市集等场景, 业务流程变化后可用自然语言修改。

链接: <https://mp.weixin.qq.com/s/RFaua7o2EhOlqck3q7AnJA>

<https://mp.weixin.qq.com/s/bINXguaCisUfQ17tA61yUQ>

4) Meta 推出个人 AI 智能体 Muse（事件日期：2026.09.08）

当地时间 9 月 8 日, Meta 推出个人 AI 智能体 Muse, 该智能体由 Meta 目前能力最强的模型 Muse Spark 驱动。Meta 表示, 其旨在使 Muse 成为一款安全、可靠、私密且广泛可用的个人 AI 智能体。Muse 运行在一台专用的安全计算机 Muse Secure VM 上, 同时承载智能体和个人数据。用户始终掌控自己的 Muse, 决定其能获得多少访问权限。

链接: <https://about.fb.com/news/2026/09/introducing-muse-personal-ai-agent/>

5) 谷歌 DeepMind 发布 AlphaGenome Atlas, 覆盖 90 亿种变异 (事件日期: 2026.09.08)

谷歌 DeepMind 推出 AlphaGenome Atlas, 用 AlphaGenome 模型预计算人类基因组全部 90 亿个单字母遗传变化的调控影响, 规模达 1PB, 是 AlphaFold 数据库的 30 多倍; 引入 AVI 变异影响评分整合编码区与非编码区预测, 并提供特征归因与超 2500 个重复性 DNA 序列及其位置的全面集合; 网站开放访问且无需编程技能。

链接: <https://mp.weixin.qq.com/s/47WZuT9Lc2zaV8qXI9dfFw>
https://mp.weixin.qq.com/s/VwHPYRm_UnUAPi0zTxkaCg

6) 宇树公开具身智能模型 UnifoLM-WLA-1.0 (事件日期: 2026.09.10)

宇树公开 60 亿参数具身智能模型 UnifoLM-WLA-1.0, 用约 2500 小时高质量真机数据训练, 一套模型完成 10 项全身操作与 54 项桌面操作任务; 模型建立统一动作空间, 把末端执行器位姿、关节与下半身动作经残差向量量化转为离散 tokens, 可适配平行夹爪与两类灵巧手; 通过统一动作空间, 构建跨本体先验迁移, 实现感知理解、交互预测与动作生成协同建模, 兼顾具身操作能力与通用多模态感知-推理能力。

链接: https://mp.weixin.qq.com/s/5PP7Tazner724OxXpY8H_A
<https://mp.weixin.qq.com/s/TkbZMqKTIS5FVqBdzO2a7A>

7) OpenAI 发布四条产品线, 其中 Agents API 开放公测 (事件日期: 2026.09.10)

OpenAI 今天发布 Agents API、GPT-Live-1 API、Data agent 与 ChatGPT for Financial Services 四条产品线, 其中 Agents API 正式开放公测; 开发者只需指定任务、模型、工具与运行环境即可创建 Agent, 负责长会话压缩、工具调度与 subagent 协作的 Codex Harness 由 OpenAI 维护并持续改进。

链接: <https://mp.weixin.qq.com/s/gkbTgMCytsv6fn0lp5DzPQ>
<https://openai.com/index/introducing-the-agents-api/>

3.4 政策法规与安全性

1) 十部门：推动普惠算力赋能中小企业发展，降低中小企业用算成本（事件日期：2026.09.03）

工业和信息化部等十部门印发《促进中小企业发展“十五五”规划》（以下简称《规划》）。《规划》提出，推进智能化赋能。鼓励具备数字化基础的中小企业加强人工智能应用，参与“人工智能+制造”、智能工厂梯度培育、智能制造系统解决方案攻关与标准制定应用。健全中小企业数智赋能服务体系，打造中小企业人工智能应用服务链。推动普惠算力赋能中小企业发展，降低中小企业用算成本。鼓励中小企业参与国家级人工智能开源社区建设，支持中小企业开展软件过程能力成熟度模型评估。

链接：<https://www.nbd.com.cn/articles/2026-09-03/4571880.html>

https://www.miit.gov.cn/zwgk/zcwj/wjfb/tz/art/2026/art_80ce437ddf6b42c2915d30bc8876dde2.html

2) 最高法发布《关于依法审理涉人工智能纠纷案件的意见》（事件日期：2026.09.07）

9月7日，最高人民法院公开发布《最高人民法院关于依法审理涉人工智能纠纷案件的意见》（以下简称《意见》），并举行新闻发布会。《意见》是首部由国家最高审判机构发布的涉人工智能司法裁判规则文件。《意见》坚持问题导向和价值导向，聚焦社会各界广泛关注的、涉人工智能审判中的突出问题，在日新月异的技术变革中把握司法规律，在划清法律“红线”的同时支持人工智能技术创新发展。《意见》分为5个部分，共24条，主要内容有：一是专门明确涉人工智能纠纷案件审理的指导思想和基本原则；二是依法规制利用人工智能的侵权行为；三是依法审理涉人工智能知识产权纠纷案件；四是探索完善审理程序规则、依法惩治违法犯罪行为；五是健全完善涉人工智能纠纷案件审理和工作机制。

链接：<https://www.court.gov.cn/zixun/xiangqing/511101.html>

3) 工信部：打造高效协同的算力设施（事件日期：2026.09.07）

工业和信息化部印发《信息通信行业发展“十五五”规划》（以下简称《规划》）。《规划》将打造高效协同的算力设施列为重点任务之一，提出加快全国一体化算力网建设，构建“枢纽-区域-边缘”多层次算力设施体系，有序部署万卡、十万卡及以上智算集群，面向场景按需部署推理算力设施，加大力度适配国产算力芯片。

链接：

https://www.miit.gov.cn/jgsj/txs/wjfb/art/2026/art_f55fe09a1e26411fa6926e5195850990.html

4) 工信部印发《“人工智能+软件”专项行动实施方案》（事件日期：2026.09.11）

工业和信息化部近日印发《“人工智能+软件”专项行动实施方案》（以下简称《方案》）。

《方案》提出，到 2028 年，软件和信息技术服务业智能化水平显著提升，培育一批高水平智能编程工具和智能开发平台，推广应用覆盖 2 万家规模以上软件企业，关键软件智能化升级取得显著成效，累计组织实施 100 项软件企业智能化技改项目，以智能体软件为代表的软件新形态加速涌现，在重点行业打造 100 个智能体软件标杆应用，信息技术服务向智能化、价值化跃升，形成一批行业深度应用解决方案和典型场景，开放协同的开源社区建设取得积极进展，孵化 5 个以上优质开源项目。到 2030 年，人工智能与软件和信息技术服务业融合发展实现新跃升，关键软件全面实现智能化升级，智能编程、智能体软件及智能服务等新业态成为产业新增长极，建成若干具有国际影响力的开源社区和产业集群，推动我国软件和信息技术服务业抢占全球价值链制高点。

链接：<https://www.cls.cn/detail/2480221>

https://wap.miit.gov.cn/jgsj/xxjsfzs/wjfb/art/2026/art_91a26793271e4c77aca49ecde4472999.html

4. 投资观点

近期，人工智能板块整体步入回调阶段，相关企业仅凭研发投入扩张、模型技术领先乃至 ARR 增长的叙事已难以满足当下投资者的诉求。市场愈发看重相关企业的商业模式是否可行、中长期业务前景是否清晰，以及其业绩表现能否兑现上述预期。以企业级智能体领域为例，上海交通大学上海高级金融学院认为，单纯的算力比拼与概念验证(POC)已无法支撑高昂的算力开销与企业的投资回报预期，智能体能否在生产环境中创造可计量、可归因的真实业务价值，成为衡量其商业化成败的根本标尺。

我们建议继续围绕人工智能应用端与基础设施端布局：应用端，重点关注已初步验证商业模式，或在中长期具备较清晰商业闭环前景的上市公司；基础设施端，优先关注能有效承接算力需求增长，产品与服务迭代有序推进，且盈利提升能见度较高的上市公司。

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