下载日志已记录，仅供内部参考

大模型密集发布，商业化持续推进

——人工智能行业半月报（2026 年 9 月第 2 期）

推荐|维持

报告要点：

- **指数表现（2026 年 9 月下半月）**

内地指数：中证人工智能产业指数下跌 1.84%，跌幅有所收窄。国证 AI 应用指数下跌 0.52%，上证科创板人工智能指数上涨 1.28%。

香港指数：恒生港股通人工智能主题指数上涨 1.33%，下行压力仍存。

- **数据跟踪**

据 OpenRouter 数据显示，在刚过去的一个整周(09.28-10.04)中，按 token 计算的大模型周调用量为 166 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，10 月 8 日，以 AAII 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 Mimo-V2.6-Pro；Anthropic 与 OpenAI 有多个性价比高的模型。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：智谱上线 GLM-5.3-FlashX；阶跃发布 Step 5 Preview；SpaceXAI 发布 Grok 4.7；小米发布并开源 MiMo-V2.6；Anthropic 发布 Claude Opus 5.5；OpenAI 推出 GPT-6 Sol 和 GPT-6 Luna；MiniMax M3.1-Flash-Preview 上线；Anthropic 推出 Claude Sonnet 5.5；OpenAI 推出 GPT-6.1 Sol 与 Ultrafast 模式；MiniMax 根据首次公开发售后股份激励计划授出奖励；谷歌推出 Gemini 4 Argon；智谱宣布根据 2026 年度股份激励计划授予奖励。

底层算力基础设施：英伟达、谷歌与 Emerald AI 成立 AI 能源联盟；亚马逊与 Generac 签署长期供货协议；华为发布昇腾 960 超节点；高通与阶跃、无量火与江波龙合作，端侧落地 MoE 模型；英伟达宣布将股票回购授权增加 1500 亿美元；AMD 收购 World Labs；DeepSeek 开源昇腾基础组件。

人工智能运用：腾讯云 Octop 1.0 正式版上线；Kimi 发布金融行业解决方案；Figure 推出新一代具身智能模型 Helix 2.5；QClaw 将停止运营；OpenAI 推出智能体 dot；Meta 推出面向小企业的 AI 助手 Muse。

政策法规与安全性：智谱就 ZCode 安全问题致歉，客户端代码全面开源；澳洲总理谴责 OpenAI 入侵澳大利亚卫生部网站；英伟达发布开放智能体安全平台。

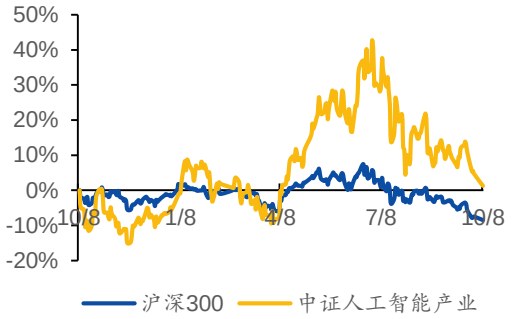
- **投资观点**

我们建议继续关注人工智能应用端与基础设施端的机会：应用端，参照商业模式已初步验证的领域，关注具备相似落地潜力的方向，选择能实质受益的上市公司；基础设施端，跟踪供需缺口持续加剧的环节，布局在相关环节产能爬坡和交付兑现能力优异的上市公司。

- **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 9 月第 1 期）：板块持续调整，前沿大模型快速迭代》2026.09.18

《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

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

下载日志已记录，仅供内部参考

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 17
3.3 人工智能实践应用 19
3.4 政策法规与安全性 21
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 10 月 4 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 10 月 4 日） 6
图 7：大模型智能水平综合排名（以 AAII 为标准，截至 10 月 8 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 AAII 为标准，截至 10 月 8 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 AAII 为标准，截至 10 月 8 日） 9
图 10：开源大模型总参数量对比（十亿单位，截至 10 月 8 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 AAII 为标准，截至 10 月 8 日） 10
图 12：编程智能体调用统计与近三周排名（截至 10 月 4 日） 11
图 13：生产力工具调用统计与近三周排名（截至 10 月 4 日） 12

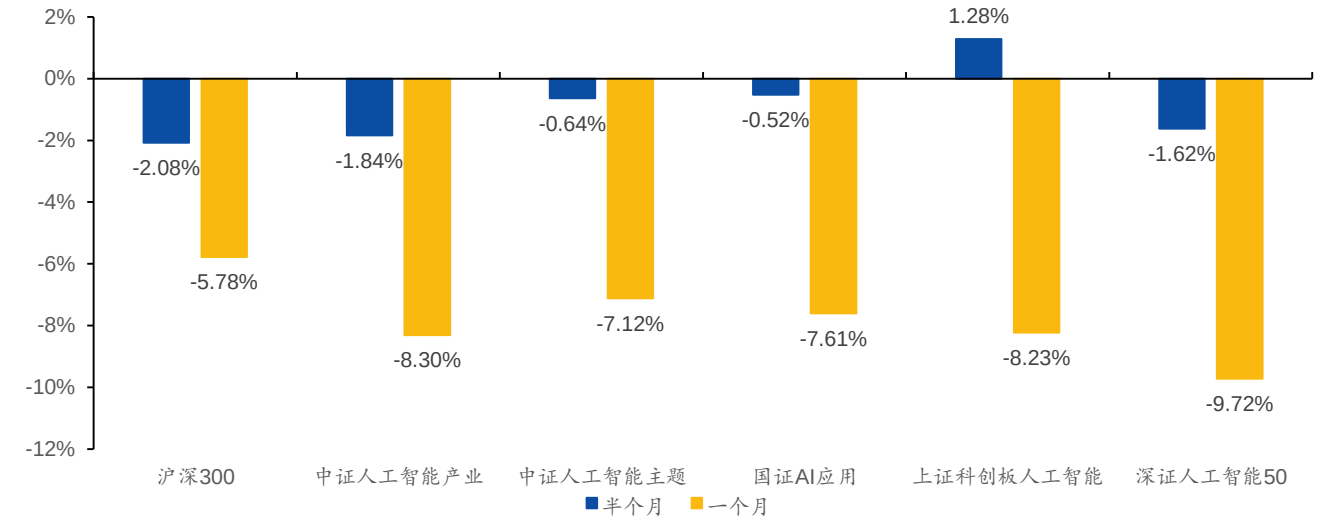
下载日志已记录，仅供内部参考

1.指数表现

1.1 内地股票相关指数

2026 年 9 月 16 日至 30 日，中证人工智能产业指数下跌 1.84%，跌幅相比 9 月上半月有所收窄，跑赢沪深 300 指数 0.23 个百分点。同一时间段内，国证 AI 应用指数下跌 0.52%，上证科创板人工智能指数上涨 1.28%。

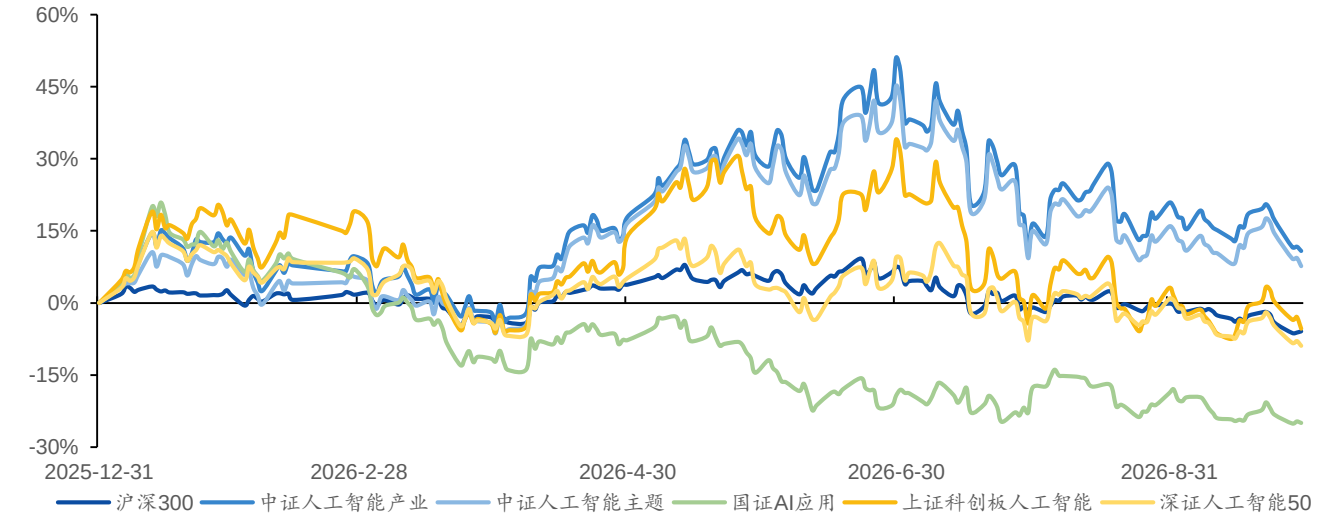
图 1：沪深 300 指数及人工智能概念指数半月度与月度涨跌幅



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 10.82%，跑赢沪深 300 指数 16.70 个百分点。国证 AI 应用指数累计下跌 24.99%，上证科创板人工智能指数累计下跌 5.37%。

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



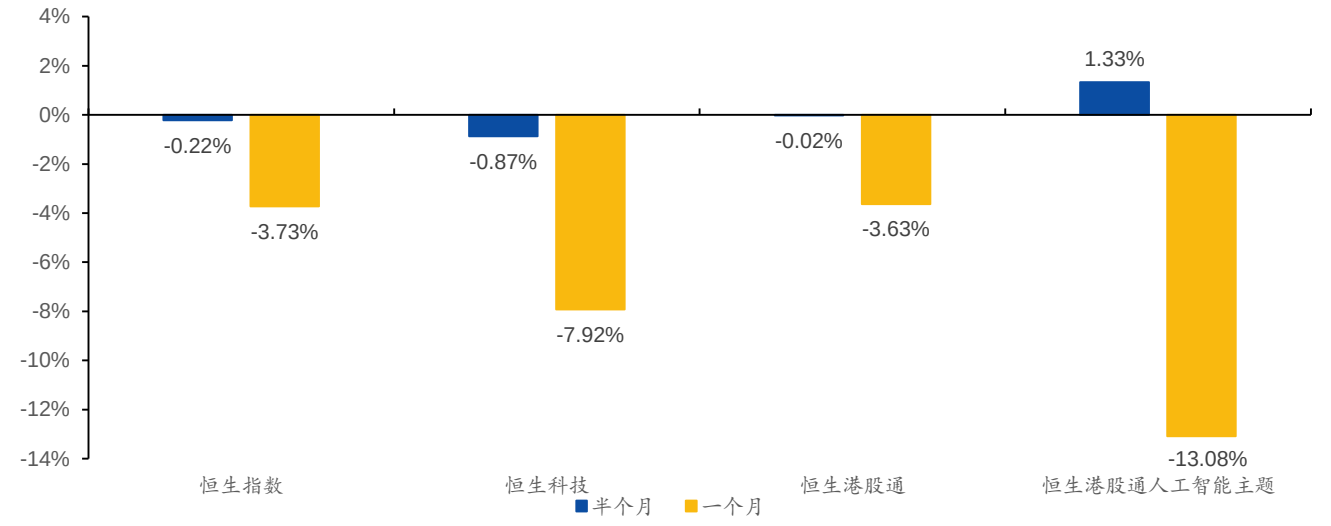
资料来源：iFinD，国元证券研究所

下载日志已记录，仅供内部参考

1.2 香港股票相关指数

2026 年 9 月 16 日至 30 日，恒生港股通人工智能主题指数上涨 1.33%，下行压力边际缓解但仍持续存在，跑赢恒生指数 1.55 个百分点。同一时间段内，恒生科技指数下跌 0.87%，恒生港股通指数下跌 0.02%。

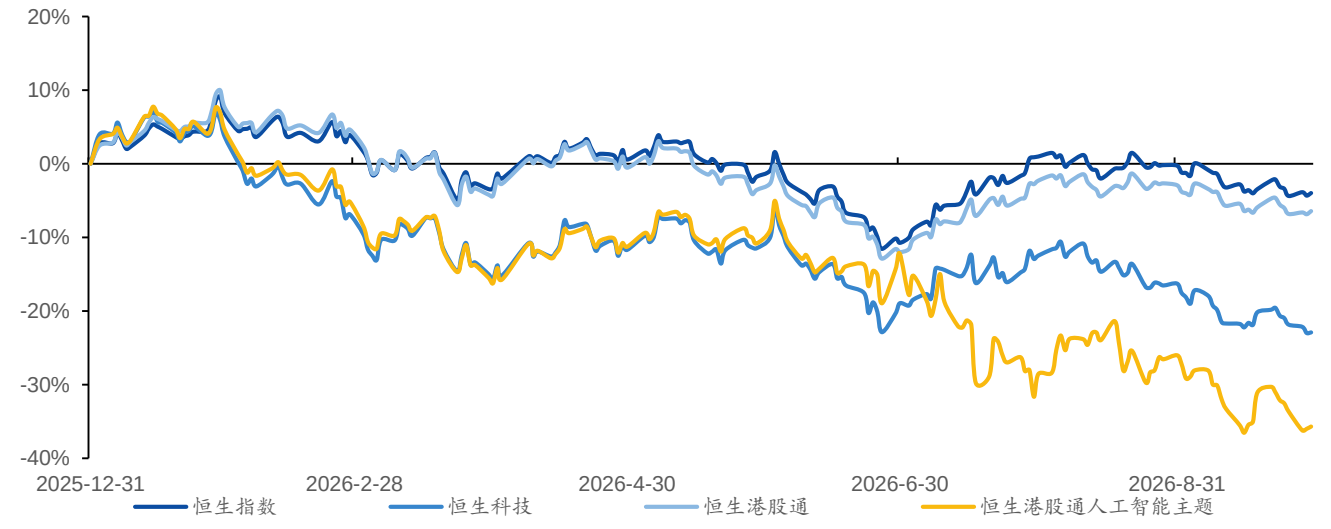
图 3：恒生系列核心指数及恒生港股通人工智能主题指数半月度与月度涨跌幅



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 35.67%，跑输恒生指数 31.70 个百分点。恒生科技指数累计下跌 22.88%，恒生港股通指数累计下跌 6.42%。

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

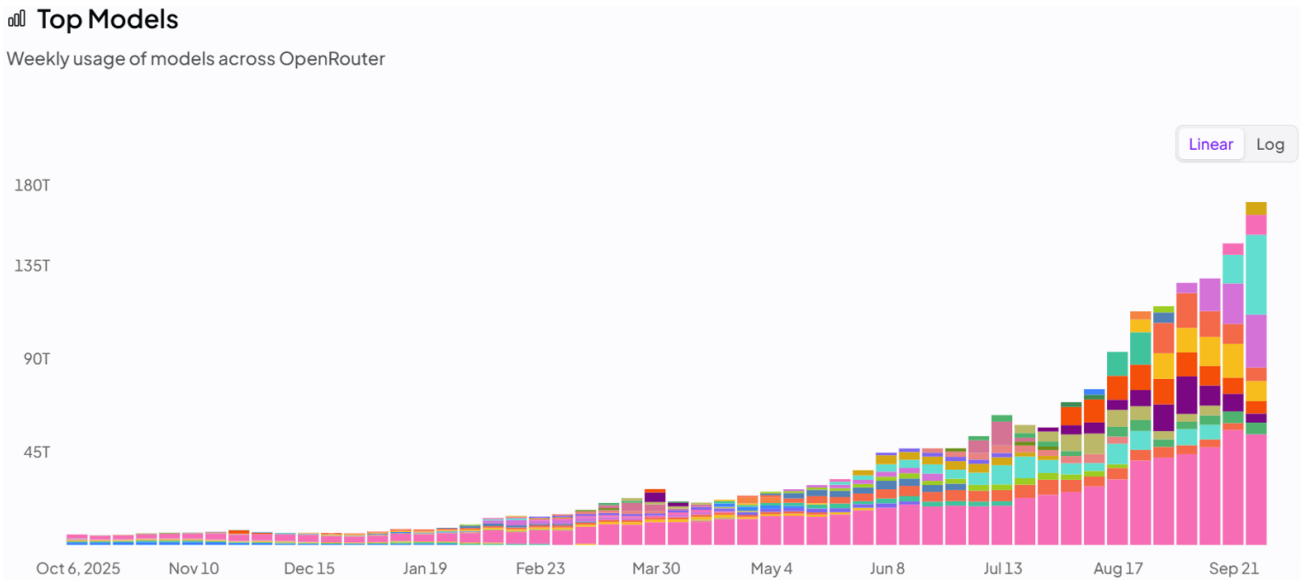
下载日志已记录，仅供内部参考

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(2026.09.28-2026.10.04)中继续创新高，达到 166 万亿 tokens，较上一个整周(2026.09.21-2026.09.27)的 146 万亿 tokens 有所上升。其中，该周调用排名前五的大模型分别为：Space Bunny Alpha、DeepSeek 的 DeepSeek V4.1 Flash、智谱的 GLM 5.3 Flash、小米的 Mimo-V2.6-Flash 以及腾讯的 Hy4 preview。

图 5：大模型调用统计与近三周排名（截至 10 月 4 日）



September 14, 2026		September 21, 2026		September 28, 2026	
Others	47.3T	Others	55.6T	Others	53.5T
DeepSeek V4.1 Flash	15.8T	DeepSeek V4.1 Flash	19.6T	Space Bunny Alpha	38.7T
GLM 5.3 Flash	14.1T	GLM 5.3 Flash	16.3T	DeepSeek V4.1 Flash	25.6T
Hy4 preview	12.5T	Space Bunny Alpha	13.9T	GLM 5.3 Flash	9.73T
GPT-5.6 Luna	9.72T	Hy4 preview	9.64T	MiMo-V2.6-Flash	9.63T
DeepSeek V4 Flash 0731	9.44T	GPT-5.6 Luna	8.53T	Hy4 preview	6.51T
MiMo-V2.5	7.07T	DeepSeek V4 Flash 0731	7.82T	GPT-6 Luna	6.18T
Hy3	4.78T	Nemotron 3 Ultra (free)	5.67T	DeepSeek V4 Flash 0731	5.99T
Nemotron 3 Ultra (free)	4.49T	MiMo-V2.6-Flash	5.51T	Nemotron 3 Ultra (free)	5.65T
DeepSeek V4 Flash 0423	3.77T	DeepSeek V4 Flash 0423	3.35T	GPT-5.6 Luna	4.37T
Total	129T	Total	146T	Total	166T

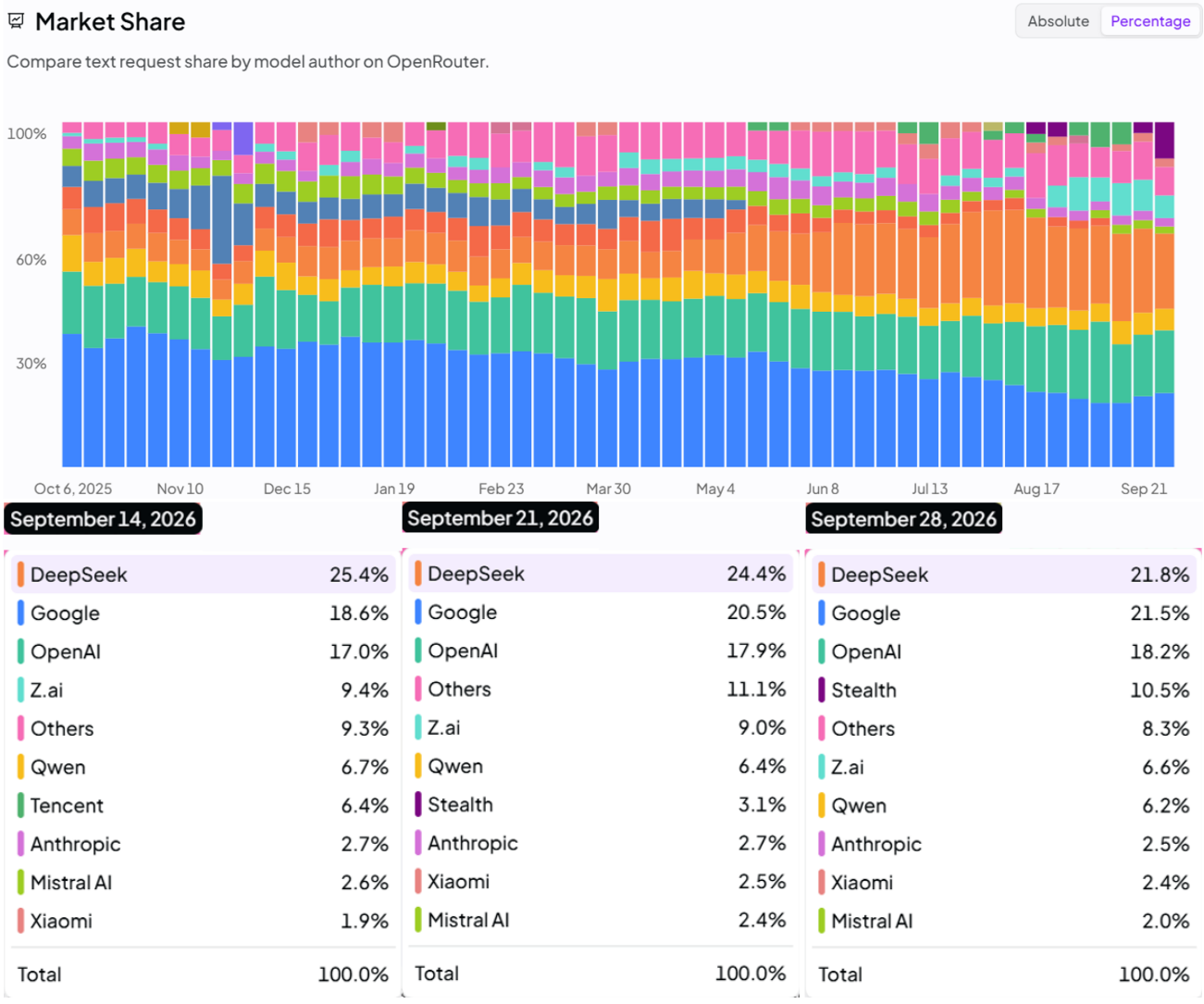
资料来源：OpenRouter，国元证券研究所

下载日志已记录，仅供内部参考

以文本请求为衡量标准，在最近过去的一个整周(2026.09.28-2026.10.04)，市场份额排名首位的是 DeepSeek(21.8%)，排名第二的是谷歌(21.5%)。DeepSeek 再次回归市场份额的榜首。

OpenAI、匿名模型（即 Stealth）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 10 月 4 日）



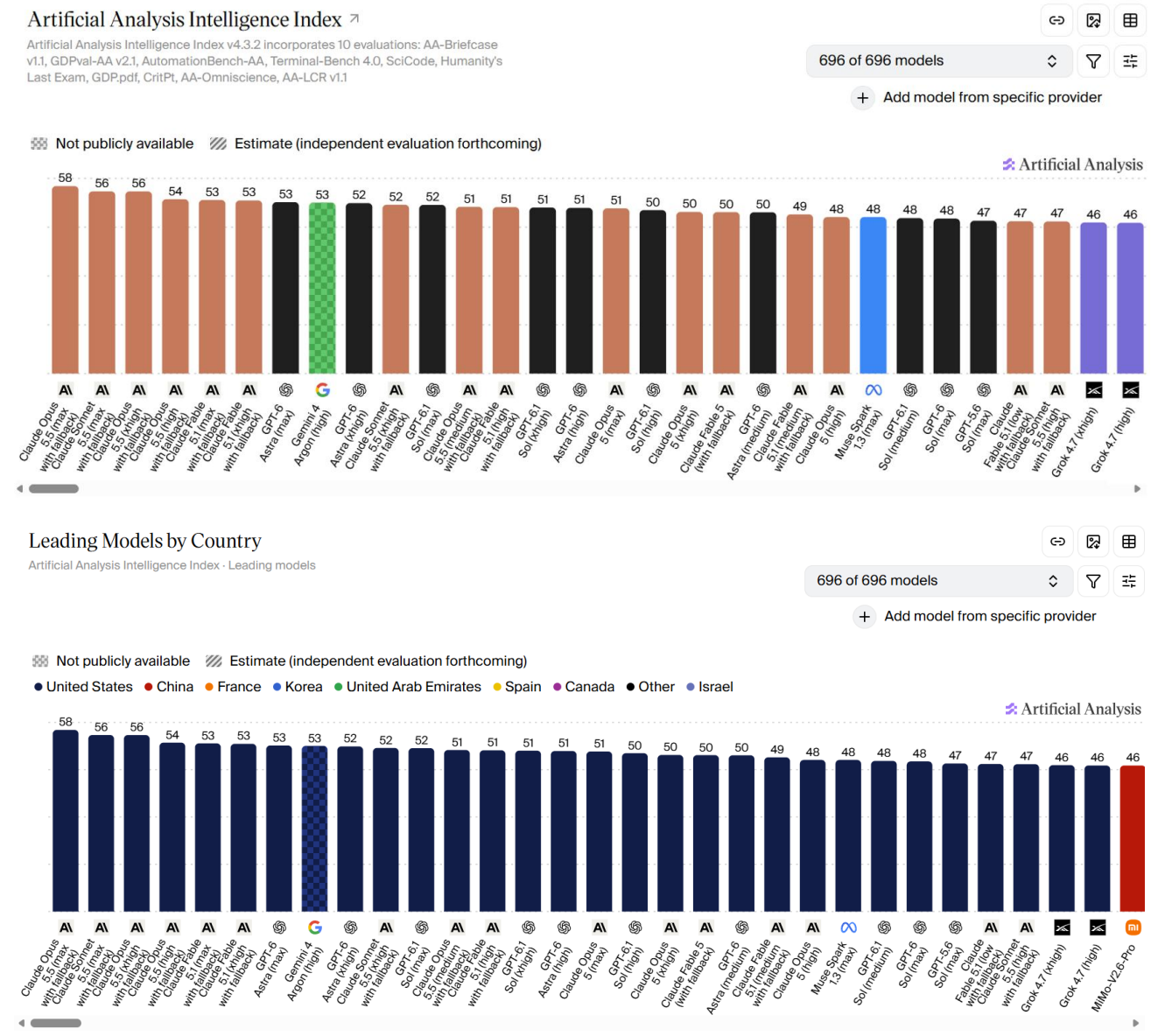
资料来源：OpenRouter，国元证券研究所

下载日志已记录，仅供内部参考

当地时间 9 月 19 日，Artificial Analysis 发布了 AAII v4.3.2 版。相比于 v4.3 版，在新版 AAII 纳入的评测中，AA-Briefcase 升级至 v1.1 版，GDPval-AA 由 v2 版升级至 v2.1 版。各评测领域在新版 AAII 计算中的权重保持不变：智能体占 30%、编程占 20%、通用占 30%、科学推理占 20%。当前 Artificial Analysis 平台显示的 AAII 数据基于 v4.3.2 版计算得出。

综合 Artificial Analysis 的测评结果看，10 月 8 日，Anthropic Claude 与 OpenAI GPT 系列模型的智能水平总体上保持领先；谷歌的 Gemini 4 Argon (high)成为当前唯一得分超过 50 分、且并非来自 Anthropic 或 OpenAI 的大模型；当前智能水平排名前三的国产大模型分别为：Mimo-V2.6-Pro、Qwen3.8 Max (0902)以及 GLM-5.3 (max)。

图 7：大模型智能水平综合排名（以 AAII 为标准，截至 10 月 8 日）



资料来源：Artificial Analysis，国元证券研究所

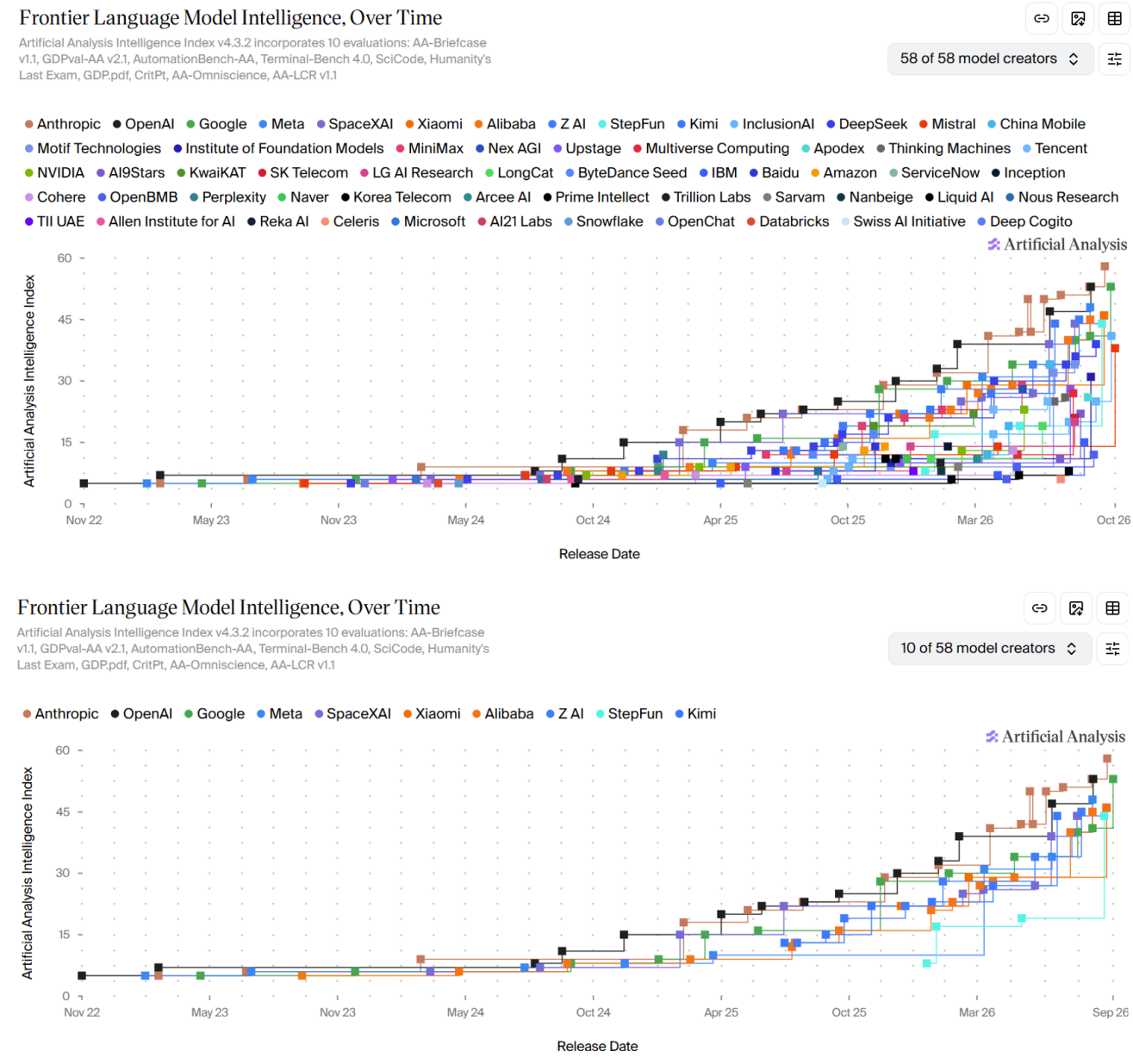
下载日志已记录，仅供内部参考

据 Artificial Analysis 的测评结果看，10 月 8 日，以 AAII 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、Google、Meta、SpaceXAI、小米、阿里巴巴、智谱、阶跃星辰以及 Kimi。

大模型厂商前十名再度由中美包揽。

以 AAII 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 AAII 为标准，截至 10 月 8 日）



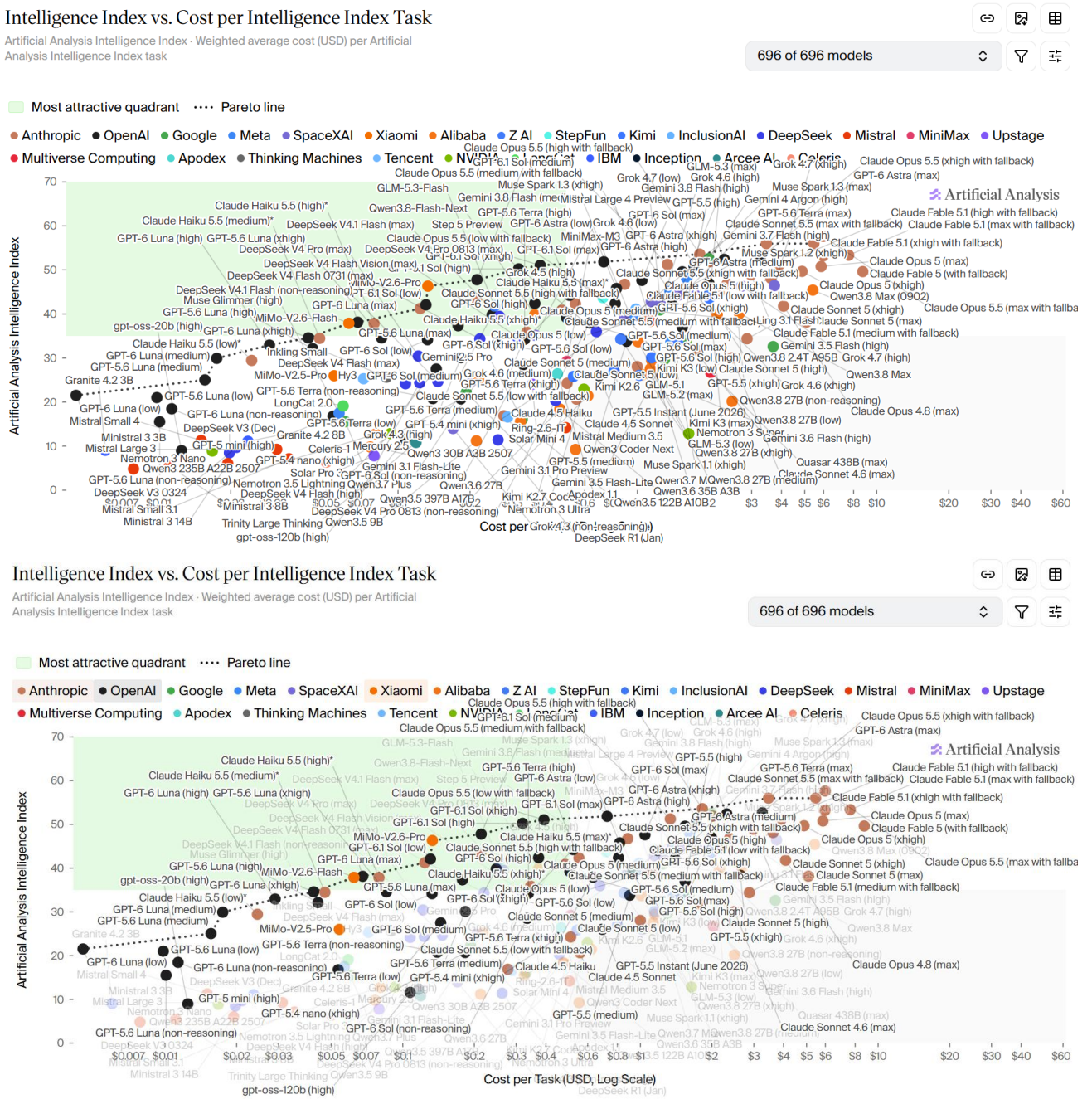
资料来源：Artificial Analysis，国元证券研究所

下载日志已记录，仅供内部参考

2.2 大模型价格与参数

据 Artificial Analysis 数据，当前 AAII-任务加权平均成本帕累托线上有多个 Anthropic Claude 与 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(AAII)，即性价比更高。国产大模型中，小米的 Mimo-V2.6-Pro 与 MiMo-V2.6-Flash 也在这条帕累托线上。

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 AAII 为标准，截至 10 月 8 日）

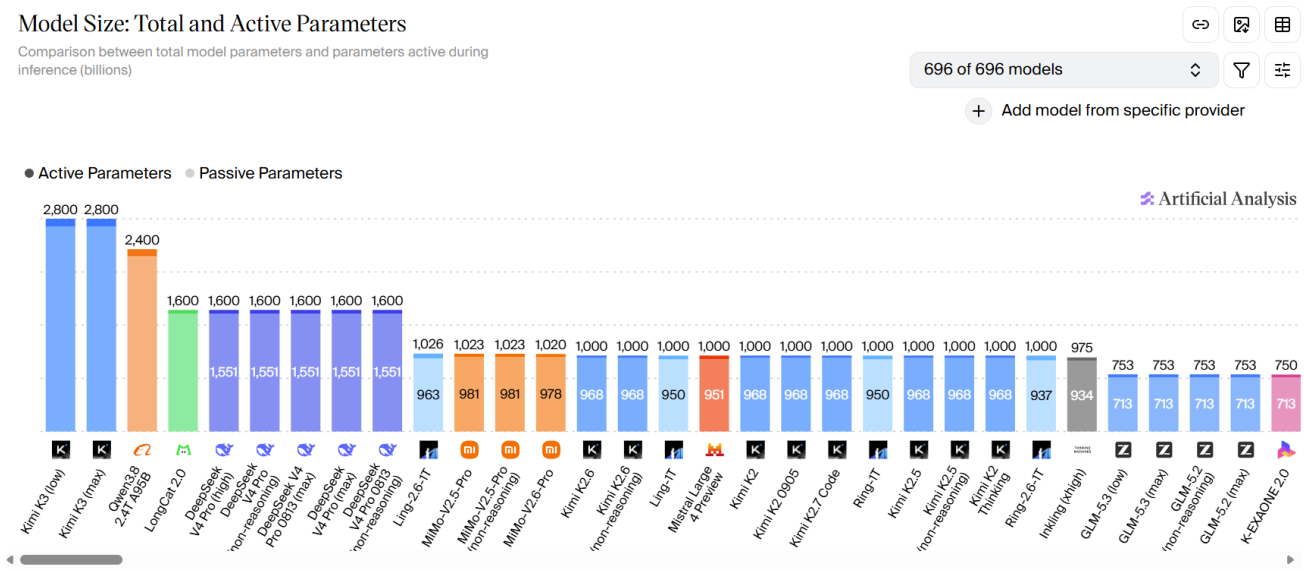


资料来源：Artificial Analysis，国元证券研究所

下载日志已记录，仅供内部参考

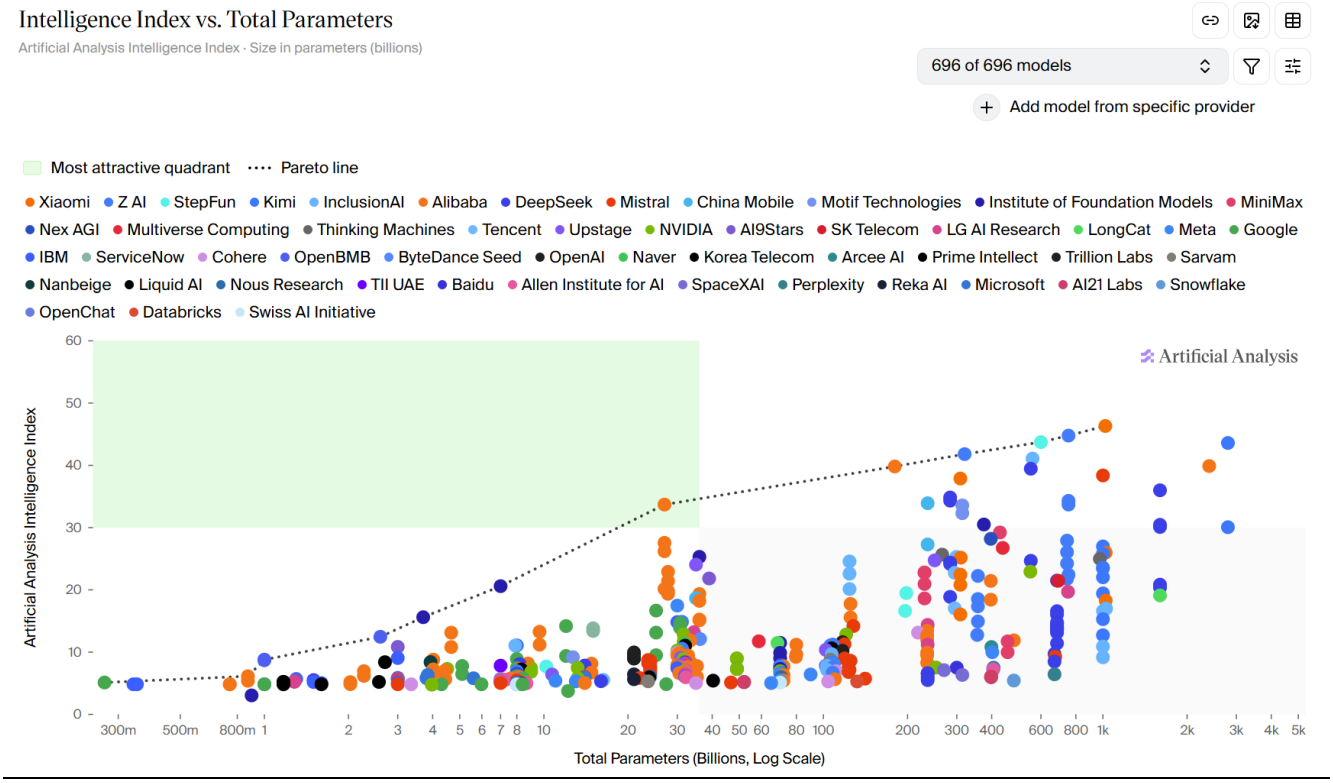
据 Artificial Analysis 数据，开源大模型中，当前总参数量最高的大模型为 Kimi K3，其在 max 和 low 思考强度模式下的总参数量均为 2.8 万亿。

图 10：开源大模型总参数量对比（十亿单位，截至 10 月 8 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 AAII 为标准，截至 10 月 8 日）



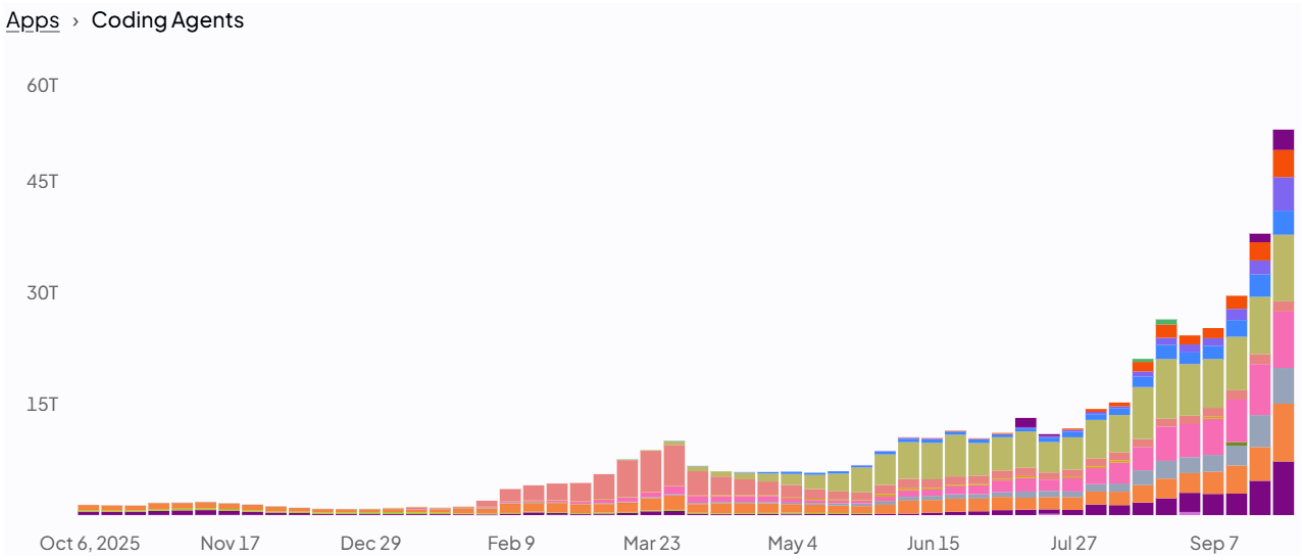
资料来源：Artificial Analysis，国元证券研究所

下载日志已记录，仅供内部参考

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(2026.09.28-2026.10.04)中，编程智能体的调用总量为 52 万亿 tokens，较上一个整周(2026.09.21-2026.09.27)的 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、Kilo Code 以及 Anthropic 的 Claude Code。

图 12：编程智能体调用统计与近三周排名（截至 10 月 4 日）



September 14, 2026		September 21, 2026		September 28, 2026	
Hermes Agent	7.22T	Hermes Agent	7.73T	Hermes Agent	8.98T
Claude Code	5.76T	Claude Code	6.88T	Kilo Code	7.8T
Kilo Code	3.79T	Cline	4.63T	Claude Code	7.6T
Cline	2.9T	Kilo Code	4.55T	Cline	7.24T
Others	2.7T	Others	4.32T	Others	4.82T
pi	2.25T	pi	3.01T	Codex	4.51T
omp	1.77T	omp	2.47T	omp	3.67T
Codex	1.51T	Codex	1.88T	pi	3.25T
OpenClaw	1.25T	OpenClaw	1.35T	Command Code	2.75T
OpenHands	444B	Command Code	1.15T	OpenClaw	1.39T
Total	29.6T	Total	38T	Total	52T

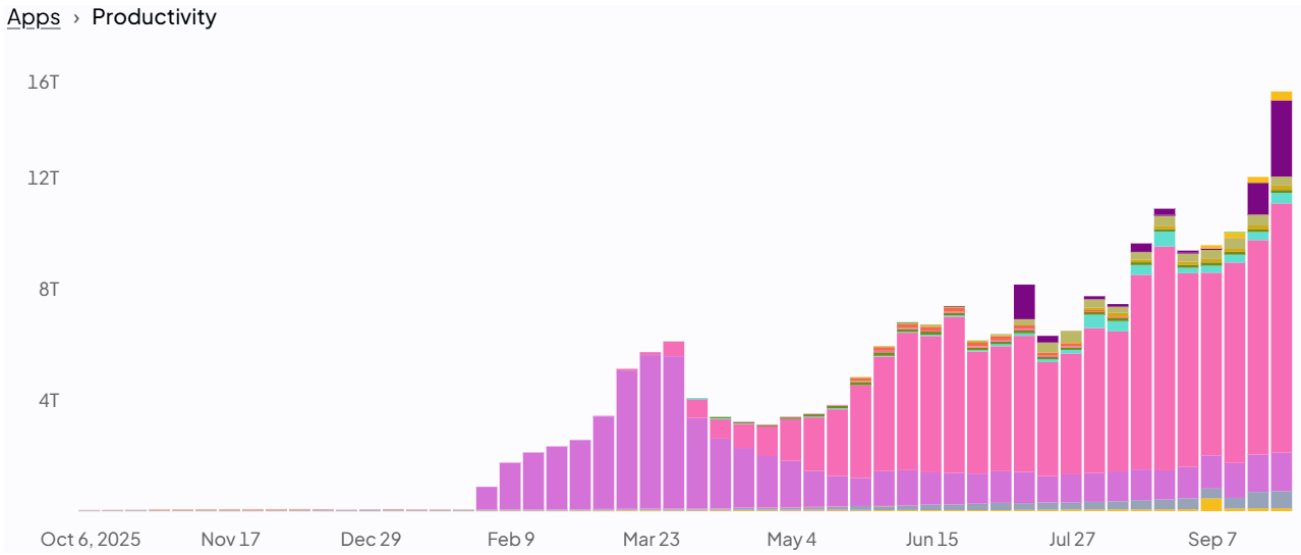
资料来源：OpenRouter，国元证券研究所

下载日志已记录，仅供内部参考

从生产力工具的角度考虑，在最近过去的一个整周(2026.09.28-2026.10.04)中，生产力工具的调用总量为 15.1 万亿 tokens，较上一个整周(2026.09.21-2026.09.27)的 12 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、Command Code 以及 OpenClaw。

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 10 月 4 日）



September 14, 2026		September 21, 2026		September 28, 2026	
Hermes Agent	7.22T	Hermes Agent	7.73T	Hermes Agent	8.98T
OpenClaw	1.25T	OpenClaw	1.35T	Command Code	2.75T
Hello Minds, powered by Eth…	397B	Command Code	1.15T	OpenClaw	1.39T
Others	384B	Others	587B	Others	616B
Nous Research API	282B	Hello Minds, powered by Eth…	389B	Nous Research API	382B
Proto Agent	170B	Nous Research API	294B	Proto Agent	319B
Mira	107B	Proto Agent	213B	Hello Minds, powered by Eth…	318B
Craft	103B	Craft	142B	Craft	162B
Open WebUI	96.7B	Open WebUI	104B	Open WebUI	105B
Hello Haven	51.6B	Mira	91.1B	Mira	104B
Total	10.1T	Total	12T	Total	15.1T

资料来源：OpenRouter，国元证券研究所

下载日志已记录，仅供内部参考

3.重点新闻

3.1 大模型与代表企业

1) 智谱上线 GLM-5.3-FlashX（事件日期：2026.09.18）

9 月 18 日，智谱正式推出 GLM-5.3-FlashX。智谱表示，GLM-5.3-FlashX 的推理速度最高 200 tokens/s，提供更快、更流畅的模型体验。

链接：https://mp.weixin.qq.com/s/ZJHhQrDeiwOGkkaqHw7kqA
https://docs.bigmodel.cn/cn/guide/models/vlm/glm-5.3-flash

2) 阶跃发布 Step 5 Preview（事件日期：2026.09.20）

9 月 20 日，阶跃官宣推出旗舰模型 Step 5 Preview，采用稀疏 MoE 架构，总参数量 600B、激活参数仅 27B，支持 100 万 Token 上下文窗口，原生支持文本与视觉输入，10 月 15 日将释放权重。据阶跃介绍，Step 5 Preview 面向 AI 编程、软件工程、专业知识工作、金融等场景设计，重点覆盖需要长上下文、多轮工具调用和持续执行的真实工作场景；相比一次性回答问题，更关注模型能否在更长的任务链中持续理解上下文、调用工具，并最终完成任务交付。

Artificial Analysis 测评结果显示，Step 5 Preview 的 AAII（v4.3.2 版）得分为 44。

链接：https://mp.weixin.qq.com/s/T4DrR5-t0XpPeehtZbNPKg
https://mp.weixin.qq.com/s/WLnBfojjQHcuiqI6wteQUw
https://artificialanalysis.ai/models/step-5

3) SpaceXAI 发布 Grok 4.7（事件日期：2026.09.21）

当地时间 9 月 21 日，SpaceXAI 发布 Grok 4.7。据 SpaceX 旗下的 Cursor 介绍，Grok 4.7 在 Grok 4.6 的基础上更进一步，采用更大的基座模型，针对难度更高的多小时任务进行了更长时间的训练，并具备更强的自我验证能力。

定价方面，据 Cursor 介绍，标准请求每百万 token 输入、缓存输入、输出价格分别为 2 美元、0.5 美元、6 美元；Fast 请求下价格翻倍；当输入超过 256k token 时，标准请求按上述费率的 2 倍计费，Fast 请求按标准费率的 3 倍计费，最高支持 500k 上下文窗口。

Artificial Analysis 测评结果显示，Grok 4.7 (xhigh)的 AAII（v4.3.2 版）得分为 46。

链接：https://cursor.com/cn/docs/models/grok-4-7
https://forum.cursor.com/t/grok-4-7-is-now-live/172526/1
https://artificialanalysis.ai/models/grok-4-7

下载日志已记录，仅供内部参考

4) 小米发布并开源 MiMo-V2.6（事件日期：2026.09.22）

9 月 22 日，小米正式发布并开源 Xiaomi MiMo-V2.6 系列，包含 Pro 和 Flash 两个原生全模态模型。小米介绍，MiMo-V2.6 完成了大规模、多任务强化学习训练，训练主要从三个维度扩展强化学习算力：更大的 Batch 与更高吞吐、更多任务与复杂环境、更大的 Grader 算力；MiMo-V2.6 系列沿用 V2.5 系列的 API 定价。

Artificial Analysis 测评结果显示，MiMo-V2.6-Pro 的 AAII（v4.3.2 版）得分为 46。

链接：https://mp.weixin.qq.com/s/WEV5SGMXghoMTwPGR812jA
https://artificialanalysis.ai/models/mimo-v2-6-pro

5) Anthropic 发布 Claude Opus 5.5（事件日期：2026.09.22）

当地时间 9 月 22 日，Anthropic 发布 Claude Opus 5.5，这是公司 CEO 达里奥·阿莫代伊呼吁“放慢前沿 AI 发展”后推出的首款新模型。Anthropic 称，Opus 5.5 在大多数任务上的表现达到 Fable 5.1 水平，较上一代 Opus 5 运行成本下降 40%，输出速度提升超 30%；API 价格为每百万 token 输入 4 美元、输出 20 美元。新模型发布前接受 Frontier Design、METR 等外部机构安全测试，Anthropic 称其在测试环境中突破边界的尝试次数较 Opus 5 和 Claude Mythos 5.1 减少 85%。

Artificial Analysis 测评结果显示，Claude Opus 5.5 (Adaptive Reasoning, Max Effort, Default Fallback)的 AAII（v4.3.2 版）得分为 58。

链接：https://www.nbd.com.cn/articles/2026-09-23/4589497.html
https://artificialanalysis.ai/models/claude-opus-5-5

6) OpenAI 推出 GPT-6 Sol 和 GPT-6 Luna（事件日期：2026.09.22）

当地时间 9 月 22 日，OpenAI 推出 GPT-6 Sol 和 GPT-6 Luna。据 OpenAI 介绍，其采用与 GPT-6 Astra 类似的方法训练了 GPT-6 Sol 和 Luna，将 Astra 在专业工作、事实准确性、编程、计算机使用和对齐方面达到的领先水平带到了更快、更经济的模型中。GPT-6 Sol 和 Luna 的 API 价格较 GPT-5.6 促销价下调 50%：GPT-6 Sol 每百万 token 输入 2 美元、输出 10 美元；GPT-6 Luna 每百万 token 输入 0.1 美元、输出 0.5 美元。

Artificial Analysis 测评结果显示，GPT-6 Sol (max)的 AAII（v4.3.2 版）得分为 48，GPT-6 Luna (max)的 AAII（v4.3.2 版）得分为 37。

链接：https://www.nbd.com.cn/articles/2026-09-23/4589497.html
https://openai.com/index/introducing-gpt-6-sol-and-luna/
https://artificialanalysis.ai/models/gpt-6-sol
https://artificialanalysis.ai/models/gpt-6-luna

下载日志已记录，仅供内部参考

7) MiniMax M3.1-Flash-Preview 上线（事件日期：2026.09.27）

9 月 27 日，MiniMax M3.1-Flash-Preview 首发上线 MiniMax Code。据 Minimax 介绍，M3.1-Flash-Preview 支持最高 1M token 上下文，适用于长文档、代码库和多步骤 Agent 会话；面向 Agent 推理、工具调用、代码和结构化任务执行优化；思考深度可调(effort)；支持文本、图片和视频等多模态输入。

链接：https://agent.minimax.cn/
https://platform.minimax.cn/docs/guides/text-generation

8) Anthropic 推出 Claude Sonnet 5.5（事件日期：2026.09.28）

当地时间 9 月 28 日，Anthropic 推出 Claude Sonnet 5.5，称其为 Claude 5.5 家族第二款模型。Anthropic 称，相比上一代 Sonnet 5，Sonnet 5.5 输出速度提升超 30%，API 价格维持每百万 token 输入 2 美元、输出 10 美元不变；新模型最擅长处理范围明确的日常任务、修复 bugs，以及制作精美的文档、幻灯片和电子表格，是首个发布时配备网络安全防护机制和 fallback 机制的 Sonnet 模型。

Artificial Analysis 测评结果显示，Claude Sonnet 5.5 (Adaptive Reasoning, Max Effort, Default Fallback)的 AAII（v4.3.2 版）得分为 56。

链接：https://www.nbd.com.cn/articles/2026-09-29/4594019.html
https://www.anthropic.com/claude-sonnet-5-5
https://artificialanalysis.ai/models/claude-sonnet-5-5

9) OpenAI 推出 GPT-6.1 Sol 与 Ultrafast 模式（事件日期：2026.09.29）

当地时间 9 月 29 日，OpenAI 推出 GPT-6.1 Sol。据 OpenAI 介绍，GPT-6.1 Sol 的智能水平接近 Astra 但价格仅为其五分之一，在智能体编程、计算机使用和专业工作方面表现尤为出色。

与此同时，OpenAI 还推出了 GPT-6 Astra Ultrafast 和 GPT-6.1 Sol Ultrafast。OpenAI 表示，借助 GPT-6.1 Sol Ultrafast，开发者只需支付与此前使用 GPT-6 Astra 大致相同的 API 费用，即可获得接近 Astra 的智能水平，速度最高可达其 8 倍。

Artificial Analysis 测评结果显示，GPT-6.1 Sol (max)的 AAII（v4.3.2 版）得分为 52。

链接：https://openai.com/zh-Hans-CN/index/introducing-gpt-6-1-sol/
https://artificialanalysis.ai/models/gpt-6-1-sol

10) MiniMax 根据首次公开发售后股份激励计划授出奖励(事件日期:2026.09.29)

9 月 29 日，Minimax 公告公司已根据首次公开发售后股份激励计划的条款向雇员承授人及服务提供商承授人授出 2,092,198 份奖励，相当于相同数目的 A 类普通股，包括向雇员承授人授出的 2,075,036 份奖励及向服务提供商承授人授出的 17,162 份奖励。

下载日志已记录，仅供内部参考

Minimax 表示，向承授人授出奖励旨在激励及奖励彼等对集团的贡献，并使彼等的利益与公司利益保持一致，以鼓励彼等为提升公司价值而努力。服务提供商承授人向本集团提供业务发展及咨询服务，而董事会认为该等服务对本集团业务具有价值。董事会认为，接受该等服务属于集团日常及一般业务过程中的事项，并认为向服务提供商承授人授出该等奖励符合公司长期发展利益，且符合该股份激励计划的目的。

链接：
https://www1.hkexnews.hk/listedco/listconews/sehk/2026/0929/2026092901776_c.pdf

11) 谷歌推出 Gemini 4 Argon（事件日期：2026.09.30）

当地时间 9 月 30 日，谷歌正式宣布推出全新的前沿模型 Gemini 4 Argon，并通过 Fairwind 计划，率先开放给一批受信任的网络安全防御专家。据谷歌介绍，Argon 是专为在复杂、长期的工作流程中维持深度推理而打造的；它在实际软件工程、如法律与金融等企业知识，以及网络安全防御等复杂的工作流程中，都展现了先进且出色的效能，加速了量子算法优化、内存效率和大规模代码库迁移与优化领域的突破。据谷歌介绍，Argon 将以优惠价推出：每百万 token 输入 2 美元，输出 10 美元；缓存输入价格则比一般输入便宜 95%。优惠期结束后，将恢复标准计费：每百万 token 输入 4 美元，输出 20 美元。

Artificial Analysis 测评结果显示，Gemini 4 Argon (High)的 AAII（v4.3.2 版）得分为 53。

链接：https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-4-argon/
https://blog.google/intl/zh-tw/products/explore-get-answers/gemini-4-argon/
https://artificialanalysis.ai/models/gemini-4-argon

12) 智谱宣布根据 2026 年度股份激励计划授予奖励（事件日期：2026.09.30）

9 月 30 日，智谱董事会宣布根据 2026 年度股份激励计划授予 1,222,934 股 H 股的奖励。据智谱公告，奖励授予日为当天 9 月 30 日，激励对象的类别为员工参与者，奖励的购买价为每股 H 股人民币 0.1 元，授予激励对象的奖励股份总归属期介乎 4 至 63 个月。智谱认为，向集团员工授予奖励有助于公司吸引及留住优秀人才，并可激励彼等为集团的业务表现及发展作出贡献，从而促进公司的长期可持续发展。

此外，智谱表示，根据 2026 年度股份激励计划可授予的股份数量上限为 44,584,309 股 H 股，即计划授权限额；于上述奖励授予后，截至 9 月 30 日，计划授权限额项下将有合共 43,361,375 股 H 股可供日后授予，其中根据 4,458,430 股 H 股的服务提供者分项限额，4,458,430 股 H 股将可供日后授予服务提供者参与者。

链接：
https://www1.hkexnews.hk/listedco/listconews/sehk/2026/0930/2026093002448_c.pdf

下载日志已记录，仅供内部参考

3.2 底层算力基础设施

1) 英伟达、谷歌与 Emerald AI 成立 AI 能源联盟（事件日期：2026.09.16）

当地时间 9 月 16 日，英伟达、谷歌与 Emerald AI 宣布成立 AI 能源管理联盟(AEMA)，推动 AI 数据中心根据电网实时状况动态调整用电。英伟达称，这种电力灵活性有助于为 AI 基础设施解锁更快、更大规模的并网连接，同时支持那些使其增长成为可能的能源系统和社区；从现有基础设施中获取更多电力，可以降低每瓦的环境影响，并支持能源可负担性；目标是建设不仅接入电网、更能与电网协同的 AI 基础设施。

链接：https://www.nbd.com.cn/articles/2026-09-17/4583749.html
https://blogs.nvidia.com/blog/ai-energy-management-alliance/

2) 亚马逊与 Generac 签署长期供货协议（事件日期：2026.09.16）

当地时间 9 月 16 日，备用发电机制造商 Generac 宣布与亚马逊达成长期供货协议，为亚马逊数据中心供应备用发电机。首批订单规模预计达到 24 亿美元，将于 2027 年和 2028 年交付。亚马逊的全资子公司 Amazon.com NV Investment Holdings LLC 获得认股权证，可按每股 200.93 美元的行权价购买最多 169 万股 Generac 股票。

链接：https://www.nbd.com.cn/articles/2026-09-18/4584915.html
https://investors.generac.com/static-files/7fd7f15a-24c2-4822-b08b-9a2b3319dab6

3) 华为发布昇腾 960 超节点（事件日期：2026.09.17）

9 月 17 日，华为全联接大会 2026 在上海启幕，华为发布全球首个采用 NPO 技术的超节点——昇腾 960 超节点，加速十万亿规模的大模型训练和推理；华为将超节点架构引入通算系统，全新升级鲲鹏超节点，基于灵衢全光组网，最大支持 4096 节点；CANN 全面开源开放，已进入常态化的开源社区运作，从“可用”到“易用”，昇腾已跨越生态拐点。

链接：https://www.huawei.com/cn/news/2026/9/hc-ascend960-supernode

4) 高通与阶跃、无量火与江波龙合作，端侧落地 MoE 模型（事件日期：2026.09.22）

高通携手阶跃、无量火与江波龙基于第六代骁龙 8 超级至尊版，将阶跃 StepEdge-Omni 30B-MoE（混合专家）模型完整部署于端侧，打造具备全链路智能体 AI 能力的端侧智能体助手。据高通介绍，依托第六代骁龙 8 超级至尊版移动平台的强大 AI 性能和架构升级，结合四方的协同优化，不仅实现了 300 亿参数 MoE 模型完整的端侧部署，还大幅优化了模型预填充和解码运行速率及内存需求，支持模型在智能手机平台上的高效稳定运行。

链接：https://www.qualcomm.cn/news/releases/2026/09/releases-2026-09-22-2

下载日志已记录，仅供内部参考

5) 英伟达宣布将股票回购授权增加 1500 亿美元（事件日期：2026.09.28）

当地时间 9 月 28 日，英伟达宣布，英伟达董事会授权将股票回购计划增加 1500 亿美元，使该计划的剩余总规模提升至 2350 亿美元，是历史上最大规模的股票回购的授权增加。公司预计在 2028 财年结束前执行完剩余的全部回购计划。

对此，黄仁勋表示，我们的现金创造能力使我们能够投资于推动这一转型的技术，并向股东返还资本。此次授权反映了我们对未来长期机遇的信心。

链接：https://nvidianews.nvidia.com/news/nvidia-announces-a-150-billion-share-repurchase-authorization-increase

6) AMD 收购 World Labs（事件日期：2026.09.28）

当地时间 9 月 28 日，AMD 宣布将收购 World Labs。据 AMD 介绍，World Labs 是一家由 AI 先驱李飞飞博士领导的 AI 模型与研究实验室。此次收购将为 AMD 带来一支世界级的研究人员和模型专家团队，增强其围绕新兴模型和应用需求来开发 AI 硬件、软件和系统的能力。这笔全股票收购交易价值约 82 亿美元，AMD 预计将于 2026 年底完成，但需获得监管批准并满足其他惯例成交条件。

链接：https://newsroom.amd.com/news/amd-acquire-world-labs/

7) DeepSeek 开源昇腾基础组件（事件日期：2026.09.30）

9 月 30 日，DeepSeek 宣布正式开源面向华为昇腾算力平台的基础设施组件，涵盖 TileLang 高级语言编译工具、计算库、分布式通信库；所有组件与此前面向英伟达平台的开源组件一一对应。

DeepSeek 表示，相对于英伟达的 CUDA 语言，TileLang 的编程更简单，能显著提高开发效率、简化代码逻辑；相对于其它同类高级语言，TileLang 的编程模型可以充分发挥芯片的特性，达到硬件的性能上限；TileLang 路线首先在英伟达成熟平台上得到了验证，如今已承载了 DeepSeek V4 系列模型训练中大部分算子的实现，是 DeepSeek 探索 AGI 新范式、开发高性能算子的核心工具；本次开源还包括昇腾平台核心计算与通信组件，为不同使用场景提供可复用的基础能力。

此外，DeepSeek 披露，在面向昇腾平台的研发过程中，华为团队给予了毫无保留的大力支持；DeepSeek 与华为紧密合作，共同推进基于昇腾 950 的 128 卡超节点方案、共同对计算与通信进行了深度优化。

链接：https://mp.weixin.qq.com/s/X41mKH4Ds-VXUAnK6M8Eww

下载日志已记录，仅供内部参考

3.3 人工智能实践应用

1) 腾讯云 Octop 1.0 正式版上线（事件日期：2026.09.17）

腾讯云开源的自托管多用户、多智能体 AI 助手 Octop 发布 1.0 正式版，一条 octop run 命令即可部署，并上线 Lighthouse 与 CVM 官方应用镜像；新增 RAG 知识库与基于 Karpathy LLM Wiki 思路的知识库专家，推出 macOS、Windows、Linux 桌面客户端及飞牛 fnOS 应用；完善页面与功能级 ACL 及 Token 额度管控，专家运行于沙箱隔离环境，插件机制新增前端渲染支持。

链接：https://mp.weixin.qq.com/s/ePg0n47SQoLiA8AhtRWT0Q

2) Kimi 发布金融行业解决方案（事件日期：2026.09.17）

9 月 17 日，Kimi 推出金融行业解决方案。据 Kimi 介绍，这套方案整合了其面向金融行业的产品与能力，涵盖 10+权威数据源、9 项金融技能和 5 项合规安全措施，以及机构级的数据建模和报告交付能力；从持仓早报、财报点评，到项目筛选、深度研究和组合复盘，其将数据获取、专业分析与成果交付连接起来。专业知识工作者和开发者可以在 Kimi Work 桌面端、Kimi Code 中安装调用数据源插件、金融技能，体验自定义插件和电脑本地 Agent 能力。

链接：https://mp.weixin.qq.com/s/ooqI84c0PRMl7f-DvDjMog

3) Figure 推出新一代具身智能模型 Helix 2.5（事件日期：2026.09.17）

当地时间 9 月 17 日，Figure 推出新一代具身智能模型 Helix 2.5。据 Figure 介绍，Figure 在 Index 上预训练了 Helix 2.5，随后将机器人带入了 30 个旧金山湾区的家庭中执行三种长时程行为（包括整理客厅、叠毛巾和铺床），且未在其中任何一个家庭采集任何数据，进行任何微调或适配；在任务特定数据、架构、训练和评估保持不变的情况下，仅 Index 预训练就将零样本成功率大幅度提高。

链接：https://www.figure.ai/news/helix-2-5-zero-shot-30-home-generalization

4) QClaw 将停止运营（事件日期：2026.09.24）

Qclaw 官网发布公告，因业务发展调整，QClaw 将于 2026 年 12 月 24 日 00:00 正式停止运营。具体安排包括：自公告之日起，停止新用户注册，停止订阅服务购买与续费；停止运营前，已注册用户、已购买订阅服务的用户均仍可正常使用服务；用户账号内的已有积分有效期不因本次停运调整，可在原有效期内继续使用；QClaw 正式停止运营后，除数据备份与迁移功能外，其他功能停止使用。

链接：https://claw.guanjia.qq.com/

下载日志已记录，仅供内部参考

5) OpenAI 推出智能体 dot（事件日期：2026.09.29）

当地时间 9 月 29 日，OpenAI 推出智能体 dot。据 OpenAI 介绍，dot 由 GPT-6 Astra 驱动，拥有独立的云端电脑，能在长期使用中从反馈里学习，并全天候为实现使用者的目标而工作；通过 OpenAI 的插件生态，可以连接 4000 多款应用；从当日起，dot 在 ChatGPT 中逐步向符合条件的市场中的 Pro 和 Business Premium 用户开放，Enterprise 用户（包括 Edu 和 Healthcare 用户）可在工作空间管理员启用后试用测试版。

链接：https://openai.com/zh-Hans-CN/index/introducing-dots/

6) Meta 推出面向小企业的 AI 助手 Muse（事件日期：2026.09.29）

当地时间 9 月 29 日，Meta 宣布推出面向小企业的 AI 助手 Muse 版本。用户可向 Muse 设定目标，例如运营业务或寻找新客户，AI 助手将协助完成相关工作。Meta 表示，未来将继续增加更多连接工具和功能。

链接：https://www.nbd.com.cn/articles/2026-09-29/4595042.html



下载日志已记录，仅供内部参考

3.4 政策法规与安全性

1) 智谱就 ZCode 安全问题致歉，客户端代码全面开源（事件日期：2026.09.21）

针对社区反馈的安全问题，智谱 ZCode 团队表示，已完成整改并向用户道歉，同时将 ZCode 开源至 GitHub 交由社区监督，并将建立常态化安全漏洞反馈与奖励机制；ZCode 团队承诺社区提及的代码数据无留存、从未用于模型训练，请大家持续监督。

链接：https://mp.weixin.qq.com/s/FFtUavCjawo40TokJjfw1w
https://mp.weixin.qq.com/s/QCHjEye57BUKUNInPKb4cA

2) 澳洲总理谴责 OpenAI 入侵澳大利亚卫生部网站（事件日期：2026.09.24）

澳洲总理 Anthony Albanese 表示，他对 OpenAI 入侵澳大利亚卫生部一个网站一事极度关切，并称这家人工智能公司拖延了太久才披露此事。Albanese 说，一个 OpenAI 的智能体于 6 月 18 日侵入了面向公众的 Medicare Statistics Reporting Service 门户。该门户托管的是关于医疗支出和药品补贴的汇总数据，深受研究人员和学者欢迎。Albanese 还表示，针对这起安全漏洞的调查将审查 OpenAI 是否可能面临刑事指控。调查还将查明，在 OpenAI 披露之前，澳大利亚安全机构为何未能发现这起入侵。

链接：https://apnews.com/article/openai-unauthorized-access-australia-altman-albanese-177e7eaf16cf743930a09445299d7735

3) 英伟达发布开放智能体安全平台（事件日期：2026.09.28）

9 月 28 日，英伟达发布开放智能体安全平台。据英伟达介绍，该开放智能体安全平台由英伟达的 OpenShell 开源软件和 Sentry 参考系统设计组成，可针对运行智能体的软件、硬件、计算和机器人系统，实施全栈治理与管控。其中，OpenShell 软件提供了一个安全的运行时边界，能够追踪智能体在 Vera CPU 上运行时的所有操作并强制执行预设的安全策略，具备跨平台扩展能力；Sentry 则是一个运行于英伟达 BlueField-4 DPU 的带外安全监控器，可持续监测智能体行为，能够在数毫秒内将试图越界的智能体隔离。英伟达表示，多家 AI 生态系统里的行业领导者与其合作，在基础设施、软件、模型和机器人的全栈体系强化各行业的 AI 安全。

链接：https://blogs.nvidia.cn/blog/open-agent-safety-platform/

下载日志已记录，仅供内部参考

4.投资观点

目前，海内外大模型加速迭代更新，厂商之间的竞争日趋激烈；竞争焦点从前沿旗舰模型向不同能力级别和具体应用场景的各层级模型全面延伸。与此同时，全球人工智能领域的头部企业也在持续发声，积极探索各种人工智能应用落地的可能，寻求更多促成商业闭环的潜在路径：

- 以网络安全为例，英伟达 CEO 黄仁勋在 CrowdStrike Fal.Con 2026 大会上提到该项业务的新机遇：人工智能时代下攻击端的自动化对防御端提出了更高的要求，但安全专业人士也能够从人工智能技术中获益，在攻防双方持续相互挑战、共同提升的协同进化循环中，构建更强大的安全业务能力。
- 微软则提出工业和物理人工智能能从新一轮工业数字化转型中创造新价值：工业人工智能可将智能应用于运营数据、识别模式、预测结果、诊断状况并推荐行动；物理人工智能则将这些能力延伸到现实世界，实现自主或人工辅助的行动，并从行动结果中学习，随时间推移不断改进性能；两者能与物联网结合建立商业学习循环，而随着这一循环随时间不断累积，生产力得以提升，韧性不断增强，专业知识实现规模化，企业自身也会变得更加智能。

我们建议继续关注人工智能应用端与基础设施端的机会：应用端，参照商业模式已初步验证的领域，关注具备相似落地潜力的方向，选择能实质受益的上市公司；基础设施端，跟踪供需缺口持续加剧的环节，布局在相关环节产能爬坡和交付兑现能力优异的上市公司。

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。

下载日志已记录，仅供内部参考

投资评级说明

（1）公司评级定义		（2）行业评级定义	
买入	股价涨幅优于基准指数 15%以上	推荐	行业指数表现优于基准指数 10%以上
增持	股价涨幅相对基准指数介于 5%与 15%之间	中性	行业指数表现相对基准指数介于-10%~10%之间
持有	股价涨幅相对基准指数介于-5%与 5%之间	回避	行业指数表现劣于基准指数 10%以上
卖出	股价涨幅劣于基准指数 5%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