仅供内部参考，请勿外传

大模型竞争转向效率与分层，Agent应用加速落地

——海内外大模型及Consumer / Work Agent产业跟踪

行业研究 · 专题报告

传媒· 计算机

投资评级：优于大市（维持）

证券分析师：张衡
021-60875160
zhangheng2@guosen.com.cn
S0980517060002

证券分析师：熊莉
021-61761067
xiongli@guosen.com.cn
S0980519030002

仅供内部参考，请勿外传

核心要点：模型效率提升推动调用扩张，Agent加速场景落地

- **模型能力继续提升，产业竞争重心加速转向效率与分层。** 海外GPT-6 Astra、Claude、Gemini等模型持续提升复杂推理、Coding、Computer Use及长程Agent能力，头部厂商同时加快将前沿能力向中端及轻量模型下沉，模型体系逐步形成“旗舰模型定义能力上限、中端模型承接复杂任务、轻量模型覆盖高频调用”的分层结构。Jev等专用决策模型进一步将Agent中的高频决策环节从通用LLM中拆分。国内DeepSeek、Qwen、GLM、MiMo等同样加速Flash化及工程效率优化，并向“模型—推理框架—国产算力”协同演进。模型产业正在从单纯提升能力上限，进入能力、成本、速度和任务效率并重的新阶段。。

- **单位智能成本下降正在释放真实调用需求，高性价比模型成为Token增长的主要推动力。**OpenRouter 2026年9月日均Token已达到18.19T，较年初提升近18倍；调用量增长并未集中于最强Frontier模型，而是明显向高性价比模型迁移。OpenAI平台份额从7月的6.0%回升至9月的13.9%，但9月Luna已经占OpenAI调用量约76.3%，Astra仅占约5.4%，Astra更多承担能力定标和复杂任务，而真正承接规模化Token的仍是低成本模型。国产模型的趋势更加显著，9月平台份额达到61.9%，其中Flash模型已占国产Token约52.4%，DeepSeek V4系列、GLM-5.3-Flash等成为主要调用来源。由此来看，模型价格下降并不一定导致大模型商业价值降低，反而可能通过降低单位成本持续释放需求；随着Agent、Coding及复杂工作流增加单任务调用链长度，Token需求、调用频次和任务复杂度有望共同推动行业增长。

- **应用端开始从“生成内容”进入“执行任务”，Consumer / Commerce Agent率先打通消费场景。**Consumer侧以Muse为代表，产品已经能够连接通信、内容、购物、旅行及支付服务，用户只需要提出目标，Agent即可完成任务规划并在后台持续执行；其上线后的下载量和DAU快速增长，初步验证了Consumer Agent的用户需求。更重要的是，Muse正在依托Meta自身的WhatsApp、Instagram、App及硬件入口，并连接Shopify、Walmart、Instacart、Expedia、Stripe、PayPal等外部生态，形成从流量入口—需求理解—商品与服务发现—任务执行—支付交易的闭环。而Sponsored Agents、Mastercard Agent Connect等基础设施的出现，进一步说明Agentic Commerce正在从单一产品探索演进为产业链：AI未来不仅影响用户购买决策，还可能直接参与订单和支付完成，商业模式有望从订阅扩展至广告、交易导流、支付和商业服务。

- **Work Agent开始进入企业真实工作流，企业入口、数据和执行能力成为核心竞争要素。**以WorkBuddy为例，AI已经从传统Copilot式的“提问—生成内容”升级为“提出目标—拆解任务—调用Skills/Connectors/MCP—多Agent执行—直接交付PPT、Excel、文档和数据结果”的完整工作流。从腾讯既有企业生态来看：微信提供入口和身份体系，腾讯文档、会议、邮箱、乐享及知识库提供企业上下文和数据资产，WorkBuddy与Managed Agents负责执行，腾讯云进一步提供模型、Runtime、算力和安全治理，从而形成“企业入口—数据权限—Agent执行—云基础设施”的完整体系。我们认为，具备用户/企业入口、私有数据、任务执行能力以及商业化闭环能力的平台更有可能将模型能力转化为订阅收入、Agent调用、交易服务和云资源收入；与此同时，Agent渗透带来的推理负载扩张，也有望继续推动模型服务、推理算力及相关基础设施需求增长。

- **风险提示：**模型能力提升及Agent渗透不及预期、推理成本下降速度不及预期、应用用户留存及商业化低于预期、行业竞争加剧及数据安全与监管政策变化；

仅供内部参考，请勿外传

国信证券
GUOSEN SECURITIES

1.海内外大模型：能力继续提升，效率与分层成为竞争主线

仅供内部参考，请勿外传

1.1 海外前沿模型密集迭代，Agent能力与运行效率同步提升

- **9月海外头部模型厂商进入新一轮密集发布周期，前沿能力继续提升。** OpenAI于9月3日发布GPT-6 Astra，重点提升复杂推理、软件工程、Computer Use及网络安全能力；Anthropic先后推出Fable 5.1/Mythos 5.1及Opus 5.5；Google六周内连续完成三次Flash模型迭代；SpaceXAI亦推出Grok 4.7。模型更新频率继续提升，复杂任务能力仍在快速演进。

- **竞争的核心变化是Agent能力与运行效率同步提升。** Gemini 3.8 Flash重点强化长程Coding与Agent任务，Grok 4.7强化长任务执行及自我校验，Opus 5.5在多数任务接近Fable 5.1能力水平的同时，典型工作负载成本较Opus 5下降约40%。模型竞争已由单一Benchmark扩展至长任务能力、Token效率、运行成本及自主执行能力。

图：能力上限继续抬升，竞争焦点正从“更聪明”扩展至“更长任务、更低成本、更高Agent效率”



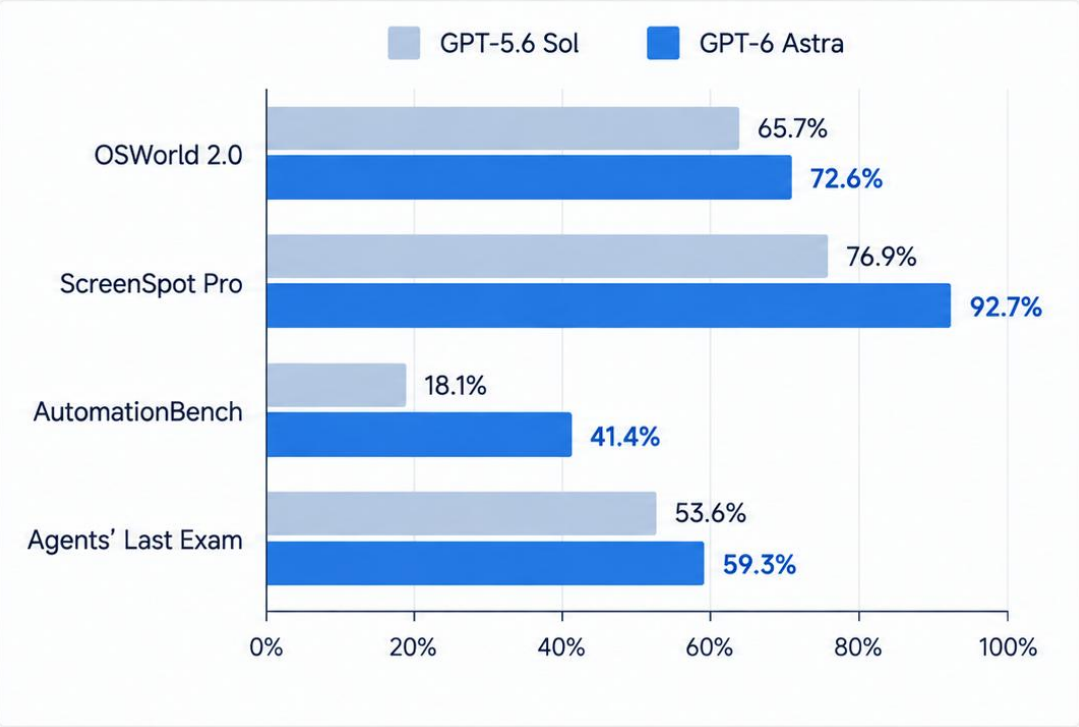
资料来源：Anthropic Institute、OpenAI，国信证券经济研究所整理

仅供内部参考，请勿外传

1.2 GPT-6 Astra：复杂任务能力与Token效率同步提升

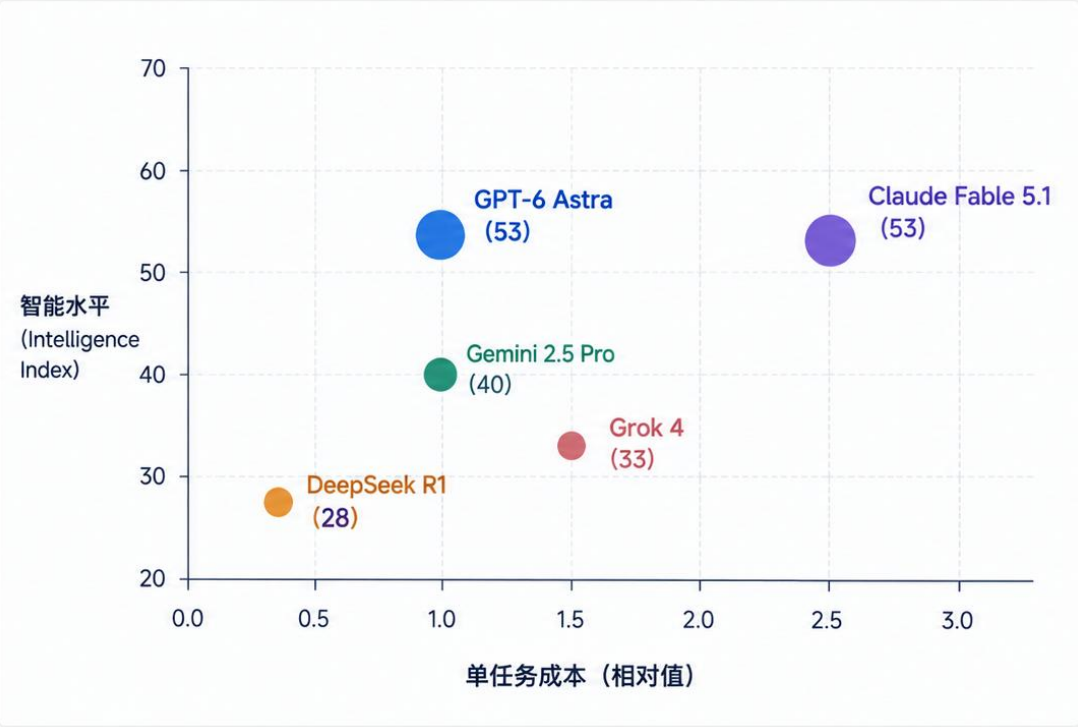
- **GPT-6 Astra进一步抬升复杂任务能力上限。** Astra的提升覆盖复杂推理、软件工程、Computer Use及专业任务等多个维度，OSWorld 2.0得分由GPT-5.6 Sol的65.7%升至72.6%，完成同类Computer Use任务所需时间缩短约47%；AutomationBench由18.1%提升至41.4%。模型能力正在由“回答复杂问题”进一步延伸至端到端执行复杂任务。
- **Token效率成为Astra本轮升级的重要特征。** Artificial Analysis测试显示，Astra Intelligence Index与Claude Fable 5.1并列达到53，但达到同等智能水平的单任务成本约为Fable的40%；Coding Agent能力同样接近Fable，成本约为其60%，主要来自更低的任务Token消耗。Astra体现出前沿模型竞争从单纯提升能力，转向复杂任务能力与单位任务效率同步提升。

图：Astra vs GPT-5.6 Sol



资料来源：OpenAI，国信证券经济研究所整理

图：相对单任务成本



资料来源：artificial analysis，国信证券经济研究所整理

仅供内部参考，请勿外传

1.3 模型体系加速分层，中端及轻量模型成为规模化调用的重要载体

- **头部厂商正在加快将前沿能力向更低成本模型下沉。** Anthropic推出Opus 5.5，在多数任务接近Fable 5.1能力水平的同时，典型工作负载成本较Opus 5下降约40%、输出速度提升30%以上；OpenAI推出GPT-6 Sol和Luna，API价格较上一代同档产品进一步下降。模型厂商已经由单纯依靠旗舰模型刷新能力上限，转向同时提升不同价格带产品的能力与效率。

- **模型产品体系正在形成清晰分层。** Frontier模型负责复杂推理、长程任务及能力定标，中端模型承接Agent、Coding等主力复杂任务，Luna、Flash等轻量模型覆盖高频、低成本调用。模型竞争的核心指标由Benchmark进一步扩展至单位任务成本、Token效率、速度和规模化调用能力

图：能力分层+成本效率



资料来源：Anthropic Institute、OpenAI、Google DeepMind，国信证券经济研究所整理

仅供内部参考，请勿外传

1.4 Jev：专用决策模型补充LLM，Agent计算架构开始功能分工

- **Jev将高频决策任务从通用LLM中进一步拆分。** Jev定位System One决策模型，不负责长文本生成，而是根据当前状态及预定义问题直接输出Choice、Score或Yes/No等结构化判断，用于路由、分类、评分及工具选择等高频任务。其端到端延迟约70－500ms，输入价格仅0.042美元/百万Token，目标是在低成本和低延迟条件下完成可直接被软件调用的决策。

- **Jev体现出Agent计算架构由“单一大模型”向功能分层演进。** 复杂规划、推理和生成仍由LLM承担，高频、低延迟的动作选择和路由判断则可由专用决策模型完成，工具层负责最终执行。Cua已经将Jev作为受约束动作选择器接入Computer Use流程，形成“规划—决策—执行”的分工结构。未来Agent计算有望逐步演化为LLM Planner + Decision Model + Tools的异构体系。

图：Jev：非LLM决策模型



资料来源：Jev，国信证券经济研究所整理

仅供内部参考，请勿外传

1.5 国产模型加速Flash化，低激活参数与低推理成本成为竞争重点

- **国产模型近期最明显的变化是Flash化及推理效率持续优化。** DeepSeek V4.1-Flash通过非对称Causal Encoder-Decoder架构显著降低激活参数及KV Cache占用；Qwen3.8-Flash原生支持1M上下文，重点面向Coding、Agent及高并发场景；GLM-5.3-Flash同样通过低激活参数及稀疏/线性注意力降低推理成本。国产模型竞争正在由参数规模和Benchmark进一步转向低延迟、低成本和规模化推理能力。

- **旗舰能力与Flash产品同步推进。** 小米MiMo-V2.6形成Pro、Flash和UltraSpeed三档产品，分别覆盖复杂项目、规模化调用及实时推理。结合DeepSeek、Qwen、GLM近期产品变化，国产模型已经形成“旗舰模型提升能力上限、Flash模型承接规模化调用、工程优化持续降本”的产品体系。

图：四家近期模型变化对比



资料来源：DeepSeek，阿里云，智谱，小米mimo，国信证券经济研究所整理

仅供内部参考，请勿外传

1.6 国产模型竞争延伸至“模型—框架—算力”协同，推理经济性成为核心指标

- **国产模型竞争正在由模型本身向整个推理体系延伸。** DeepSeek V4.1-Flash发布后，寒武纪通过vLLM+NeuWare实现Day-0适配；GLM-5.3驱动的Infra Agent在10万卡级国产集群上完成推理系统优化，两周内将端到端吞吐提升至初始基线约3倍；商汤大装置亦开始提供GLM-5.3-Flash国产算力及Token服务。模型、推理框架与国产芯片之间的协同速度明显提升。

- **产业竞争由“能否运行”进一步进入“能否低成本规模化运行”的阶段。** 模型适配速度、单位Token成本、集群利用率和生产部署效率正在成为国产推理体系的重要竞争指标，国产算力生态开始由可用性验证向经济性验证推进。

图：从“模型能力”进一步延伸至“模型—框架—算力”



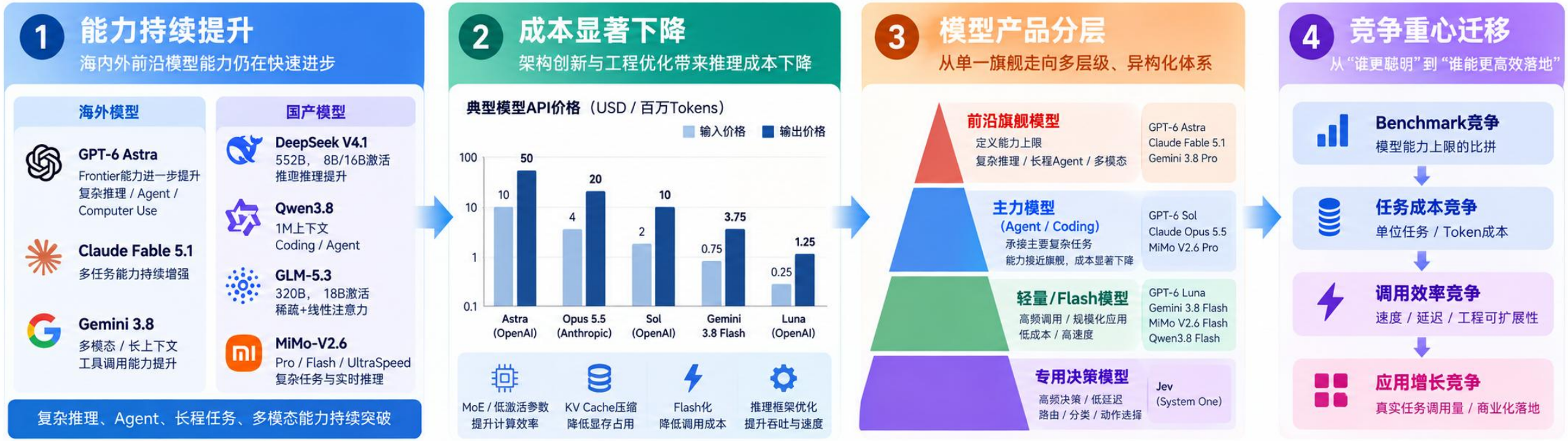
资料来源：DeepSeek，阿里云，智谱，小米mimo，国信证券经济研究所整理

仅供内部参考，请勿外传

1.7 模型竞争进入“能力、成本与调用效率”综合比拼阶段

- **海内外模型能力仍在持续提升，但竞争维度已经明显扩展。** 海外Astra、Claude、Gemini等持续提升复杂推理、Agent、Computer Use及长程任务能力；国内DeepSeek、Qwen、GLM、MiMo则同步推进Flash化、低激活参数、KV Cache压缩及推理工程优化。成本、速度、Token效率和工程可部署性的重要性持续上升
- **模型产品和计算架构同时走向分层。** Frontier模型负责复杂推理及能力定标，中端和Flash模型承接大规模真实调用，Jev等专用模型进一步处理高频决策任务，国产体系则叠加模型、框架和算力协同优化。模型产业正在由“模型能力更强”，转向“能够以更低成本、更高效率完成更多真实任务”

图：大模型产业正在从“能力竞赛”进入“能力 + 成本 + 调用效率”的综合竞争阶段



资料来源：Anthropic Institute、OpenAI、Google DeepMind 、DeepSeek、阿里云、智谱、小米mimo ，国信证券经济研究所整理

仅供内部参考，请勿外传

国信证券
GUOSEN SECURITIES

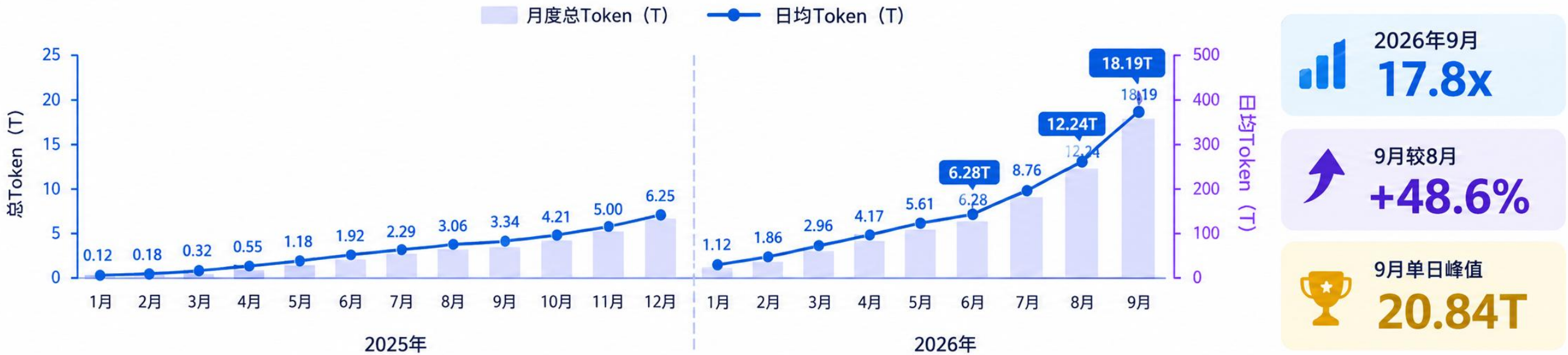
2. OpenRouter观察：调用规模快速扩张，高性价比模型是主要增量

仅供内部参考，请勿外传

2.1 OpenRouter调用规模持续加速，9月日均Token较年初提升近18倍

- **OpenRouter模型调用量在2026年进一步加速。** 日均Token由1月的1.02T提升至6月6.28T、8月12.24T，截至9月22日进一步达到18.19T，较8月提升48.6%，较年初增长约17.8倍；9月单日峰值达到20.84T。
- **Token增长同时来自调用频次提升和单任务负载增加。** 9月前20个有效日累计Token已达到363.7T，相当于8月全月的95.9%；Top应用请求量代理指标9月日均升至约3361万次，较8月增长约45%。真实推理需求仍处于快速扩张阶段。

图：OpenRouter调用规模持续加速



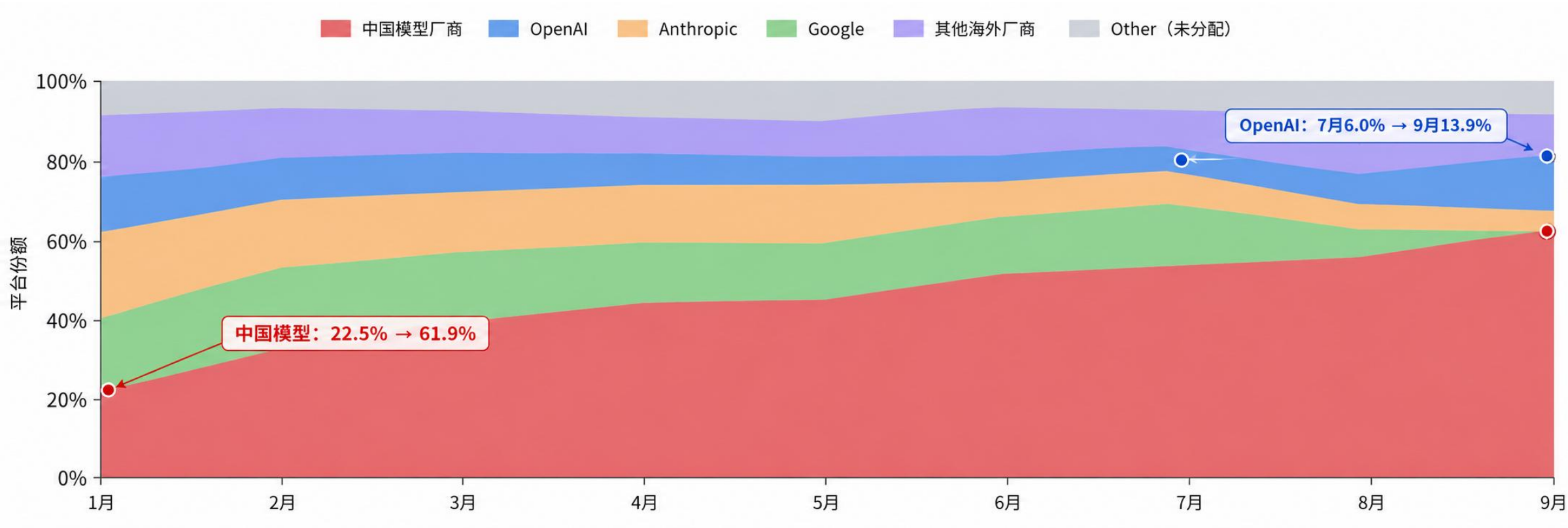
资料来源：OpenRouter，国信证券经济研究所整理

仅供内部参考，请勿外传

2.2 国产模型份额升至61.9%，OpenAI份额连续两月回升

- OpenRouter调用结构继续向国产模型迁移。 国产模型厂商份额由2026年1月的22.5%提升至7月61.5%，8月短暂回落后，9月再次升至61.9%；同期Google和Anthropic份额分别降至5.1%和3.4%。国产模型已经成为平台主要Token来源。
- OpenAI份额自8月以来明显修复。 OpenAI平台份额由7月6.0%升至8月10.7%、9月进一步升至13.9%，9月Token量达到50.6T。进一步拆分内部模型结构看，本轮份额回升主要由Luna等高性价比模型放量推动

图：2026年1-9月核心模型厂商调用份额



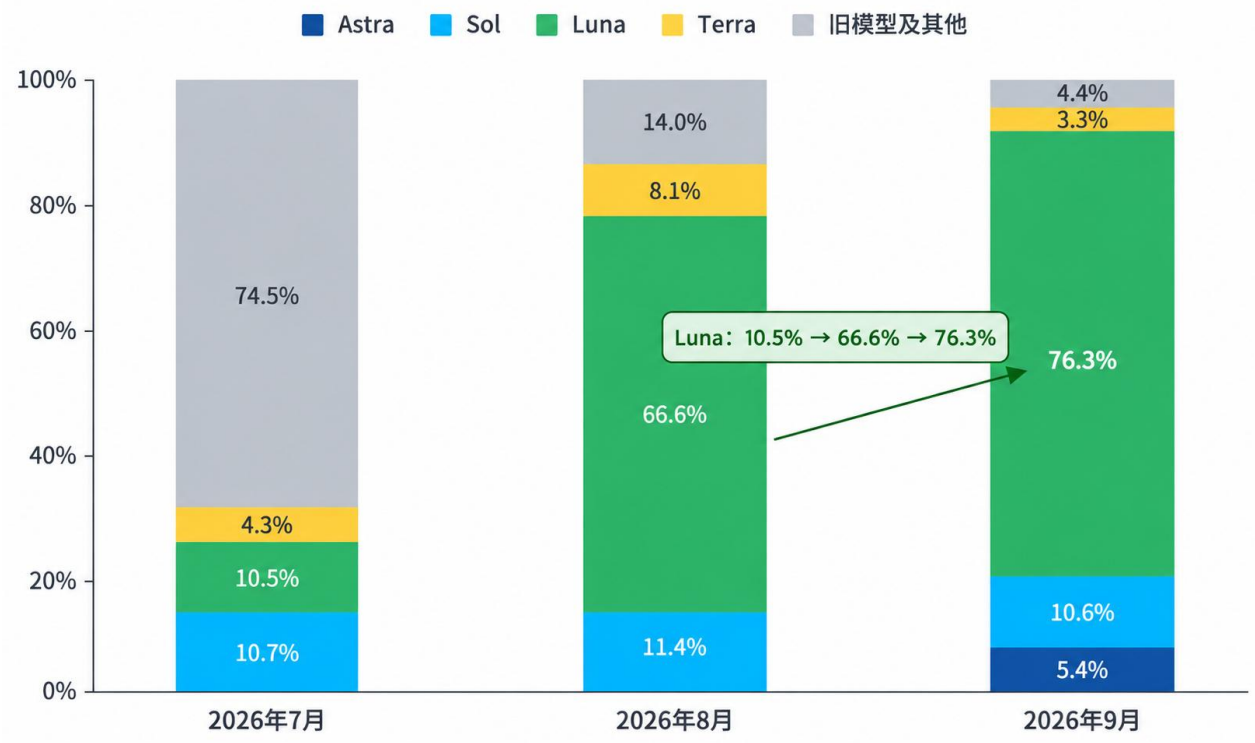
资料来源：OpenRouter，国信证券经济研究所整理

仅供内部参考，请勿外传

2.3 Luna驱动OpenAI份额回升，Astra主要承担前沿能力定标

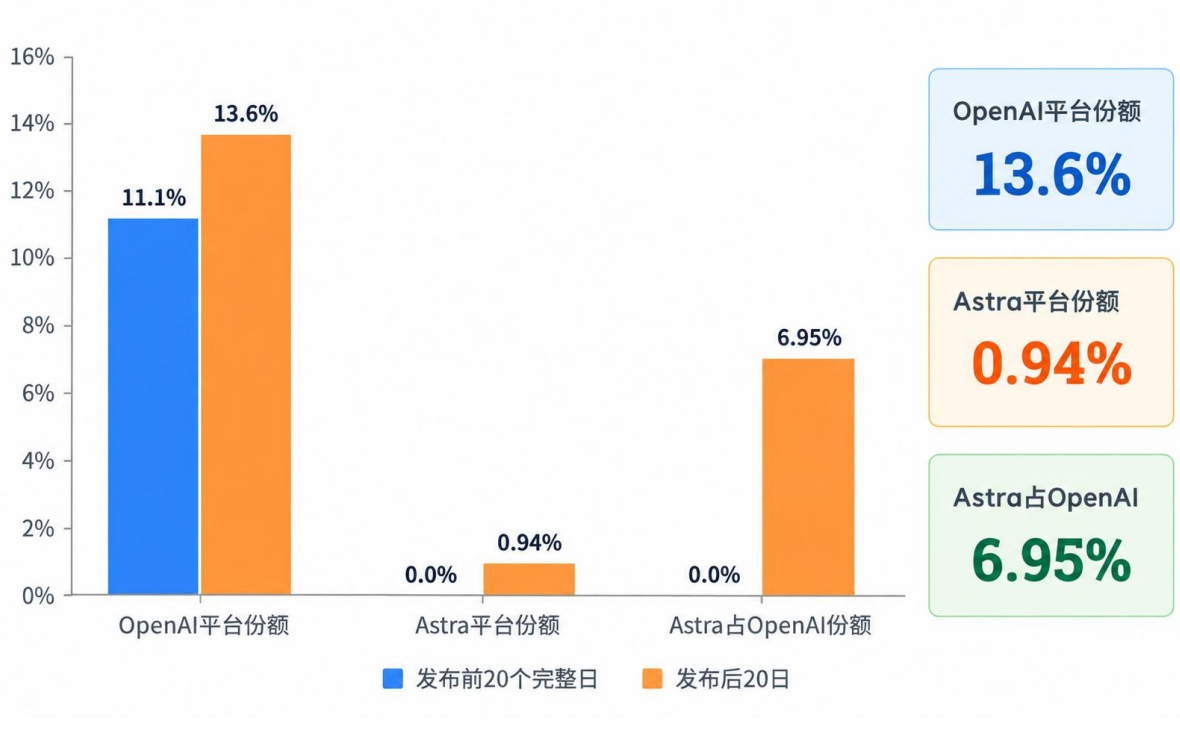
- OpenAI近期Token增长主要由Luna系列承接。 Luna占OpenAI内部调用比例由7月10.5%快速升至8月66.6%，9月进一步达到76.3%；同期OpenAI平台份额由6.0%升至13.9%。截至9月22日，Sol占OpenAI调用10.6%，Astra仅占5.4%，高性价比模型是本轮规模化调用增长的主体。
- Astra主要承担前沿能力和复杂任务定位。 Astra发布后OpenAI整体平台份额有所提升，但Astra自身平台份额仍不足1%。OpenAI当前已经形成较清晰的产品分工：Astra负责能力上限与复杂任务，Luna等模型负责大规模Token调用。

图： OpenAI内部模型结构



资料来源：OpenRouter，国信证券经济研究所整理

图： Astra发布前后20日对比



资料来源：OpenRouter，国信证券经济研究所整理

仅供内部参考，请勿外传

2.4 国产模型份额维持60%+，调用增长由单点驱动转向多厂商共同放量

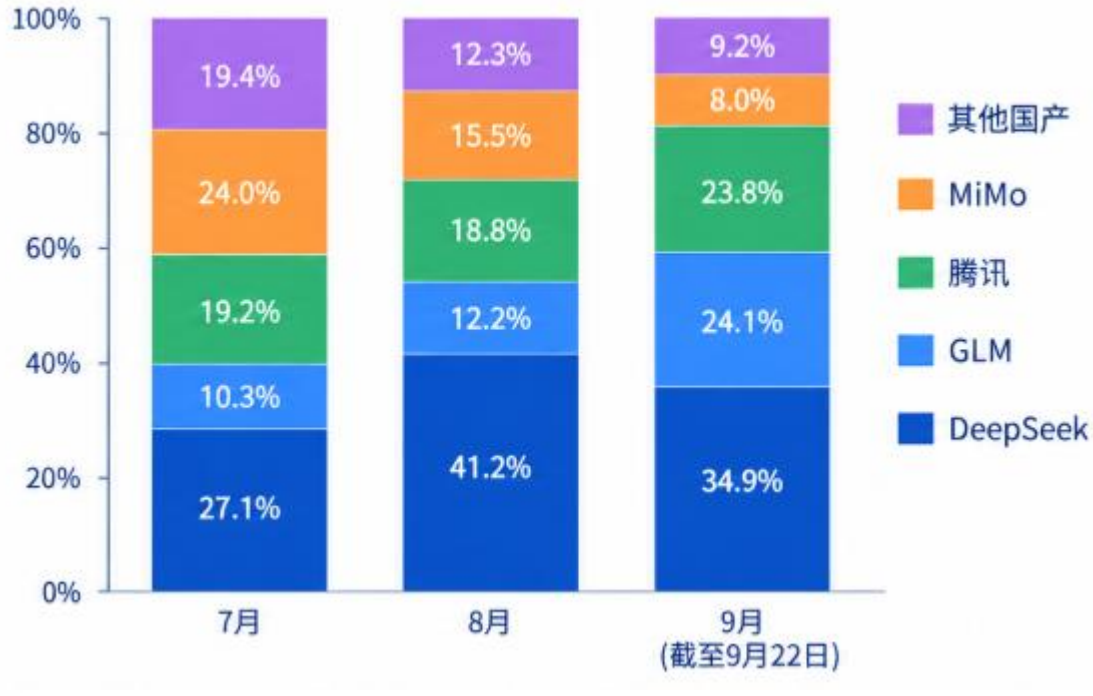
- 国产模型已成为OpenRouter最主要的调用来源。 9月前20个覆盖日国产模型Token达到225T，平台份额升至61.9%，较年初22.5%大幅提升，调用增长已经由阶段性爆款进入较高基数下的持续放量阶段。
- 内部结构由单一模型驱动向多厂商共同增长演变。 9月DeepSeek仍位居首位，占国产模型Token约34.9%；GLM占比由8月12.2%升至24.1%，腾讯由18.8%升至23.8%，三者合计达到国产调用约83%。国产模型调用增长正在由早期DeepSeek等单点驱动，向DeepSeek、GLM、腾讯等多厂商共同放量转变。

图：国产模型平台份额



资料来源：OpenRouter，国信证券经济研究所整理

图：国产模型内部份额结构



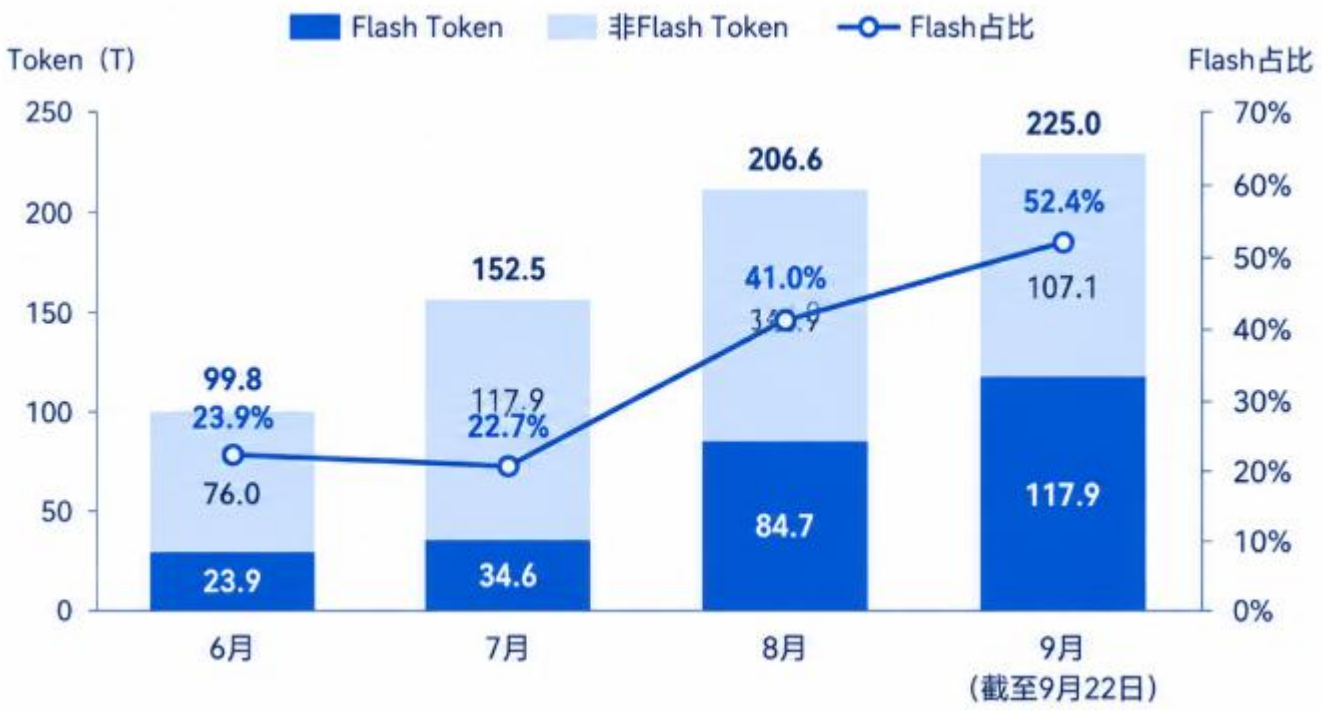
资料来源：OpenRouter，国信证券经济研究所整理

仅供内部参考，请勿外传

2.5 国产调用加速向Flash集中，9月Flash占比升至52.4%

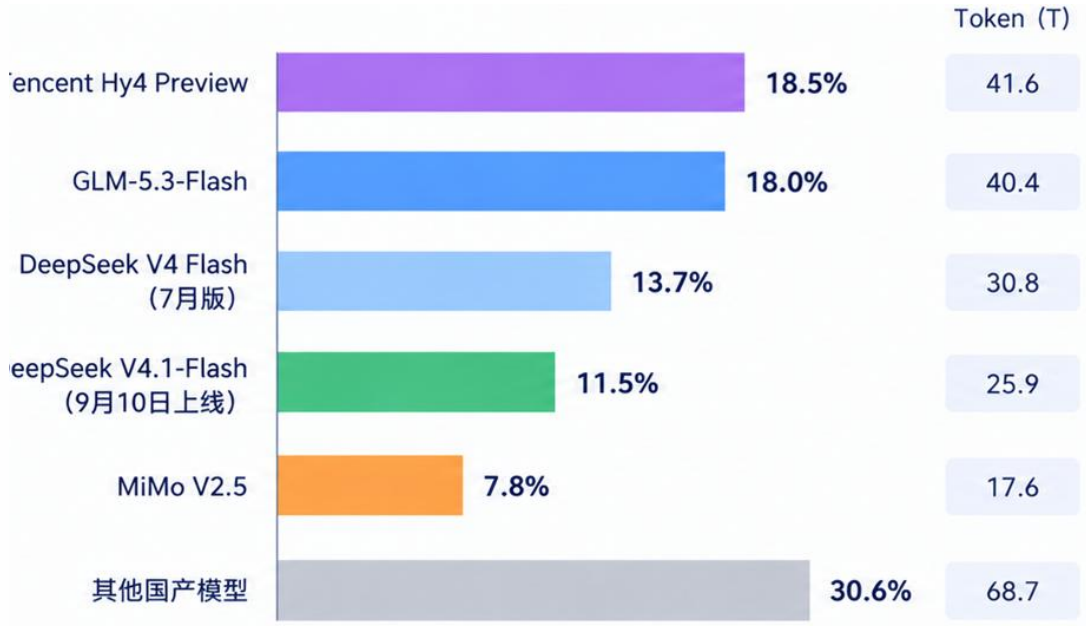
- 国产模型调用增量正在加速向Flash模型集中。 国产Flash模型占Token比例由6月23.9%、7月22.7%快速提升至8月41.0%，9月进一步达到52.4%。9月国产模型总Token较8月增加约18.4T，而Flash模型Token增加33.2T，非Flash模型调用有所下降，近期国产Token增量主要由Flash模型贡献。
- DeepSeek及GLM Flash系列成为主要负载来源。 9月GLM-5.3-Flash占国产模型调用约18.0%，DeepSeek V4 Flash约13.7%，9月10日上线的V4.1-Flash已贡献11.5%。Flash正在由产品补充形态转变为真实推理负载的主要承载形态。

图：国产Flash模型占国产Token比例



资料来源：OpenRouter，国信证券经济研究所整理

图：9月国产模型Top 5调用结构



资料来源：OpenRouter，国信证券经济研究所整理

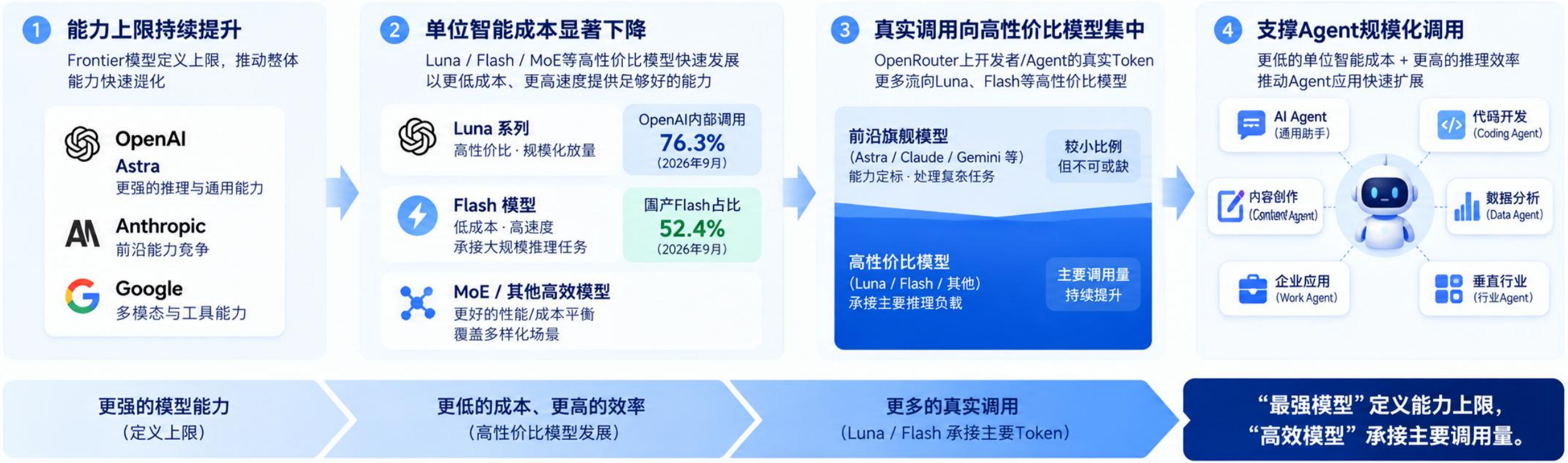
仅供内部参考，请勿外传

2.6 调用趋势：从“最强模型”转向“能力×成本×效率”的综合选择

- **OpenRouter数据验证了模型分层和效率竞争已经进入真实调用环节。** 9月OpenAI内部Luna占比达到76.3%，Astra仅为5.4%；国产模型平台份额达到61.9%，其中Flash模型占国产Token升至52.4%。大规模Token正在明显向能够以更低成本、更高速度完成任务的模型集中，Frontier模型则主要负责能力定标和高复杂度任务。

- **模型产业的核心指标正在由Benchmark扩展至真实使用效率。** 不同任务根据所需智能水平进行模型路由，Frontier模型负责高难度任务，高性价比模型负责规模化调用。未来模型竞争将更多体现为能力、单位任务成本、速度及真实调用量的综合效率。 OpenRouter以开发者、API及Agent场景为主，其高频数据能够较好反映模型调用结构的边际变化。

图：从“最强模型”转向“能力×成本×效率”



资料来源：OpenRouter，国信证券经济研究所整理

仅供内部参考，请勿外传

国信证券
GUOSEN SECURITIES

3 Agent应用：从问答走向任务执行，Consumer / Work双线加速落地

仅供内部参考，请勿外传

3.1 应用：Agent进入任务执行阶段，Consumer / Commerce与Work两条路径加速落地

- **AI应用正在由“生成答案”向“执行任务”升级。** Consumer / Commerce Agent从信息问答进一步延伸至需求理解、任务执行及交易完成；Work Agent则由办公助手升级为能够连接企业数据、调用工具并直接交付工作成果的生产力平台。应用竞争的重点由单一功能展示转向入口、场景、数据连接及任务执行能力。
- **Consumer和Work构成当前Agent落地的两条主要路径。** Muse开始打通用户入口、任务执行及交易支付链路；WorkBuddy则进入企业数据、权限和真实工作流。两者共同体现出Agent产品形态由“回答问题”向“完成任务”演进，AI应用价值开始由使用时长进一步延伸至交易规模、工作效率及企业IT支出。

图：Agent应用正沿“Consumer / Commerce”与“Work”两条路径加速落地



资料来源：meta、腾讯，国信证券经济研究所整理

仅供内部参考，请勿外传

3.2 Muse：Consumer Agent开始验证大众用户需求

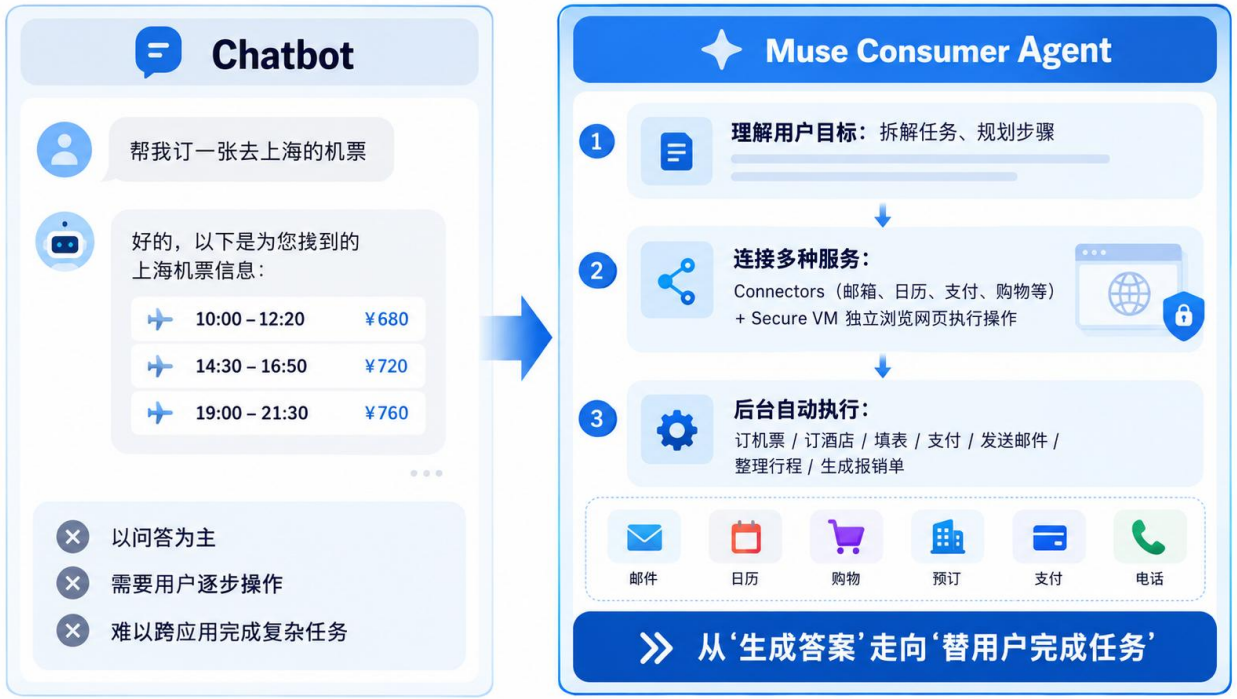
- **Muse将消费级AI交互由“提问”升级为“交付任务”。** Muse能够连接邮件、日历、支付及购物等服务，通过Secure VM浏览网页并在后台持续执行订票、购物、填表、发送邮件等多步骤任务。Connectors进一步扩展第三方服务覆盖范围，使Consumer Agent具备持续扩展真实消费场景的能力。
- **用户增长初步验证Consumer Agent的大众需求。** Muse上线12天累计下载约280万次，美国和加拿大iOS下载约180万次；9月19日美国单日下载达到26.4万次，9月18日DAU约44.8万，并一度登顶美国App Store。Consumer Agent开始由技术产品进入大众用户市场。

图：Muse上线后的用户增长



资料来源：Sensor Tower、Meta、Reuters，国信证券经济研究所整理

图：从Chatbot到Consumer Agent



资料来源：meta，国信证券经济研究所整理

仅供内部参考，请勿外传

3.3 Muse打通Meta入口、外部服务与支付链路，商业化路径逐步清晰

- **Meta既有消费入口显著降低Muse的用户触达成本。** Muse除独立App外，已经能够通过WhatsApp交互，并连接Instagram内容、AI眼镜和Mac等入口，使社交通信、内容消费、移动端及硬件设备逐步汇聚至同一个Personal Agent。Meta庞大的既有用户和内容生态为Agent提供持续的需求入口和任务触发场景。

- **外部服务与支付体系正在补齐从需求产生到交易完成的链路。** Muse已接入Shopify、Walmart、Best Buy、Instacart等购物服务，并逐步连接Expedia、Notion、GitHub等旅行及工作场景，支付端接入Stripe Link、Shop Pay和PayPal。Muse已经能够沿着需求产生—服务发现—Agent执行—支付交易完成端到端任务，其商业价值有望由订阅进一步延伸至交易导流、支付及商业服务。

图：Muse商业闭环逻辑



资料来源：meta，国信证券经济研究所整理

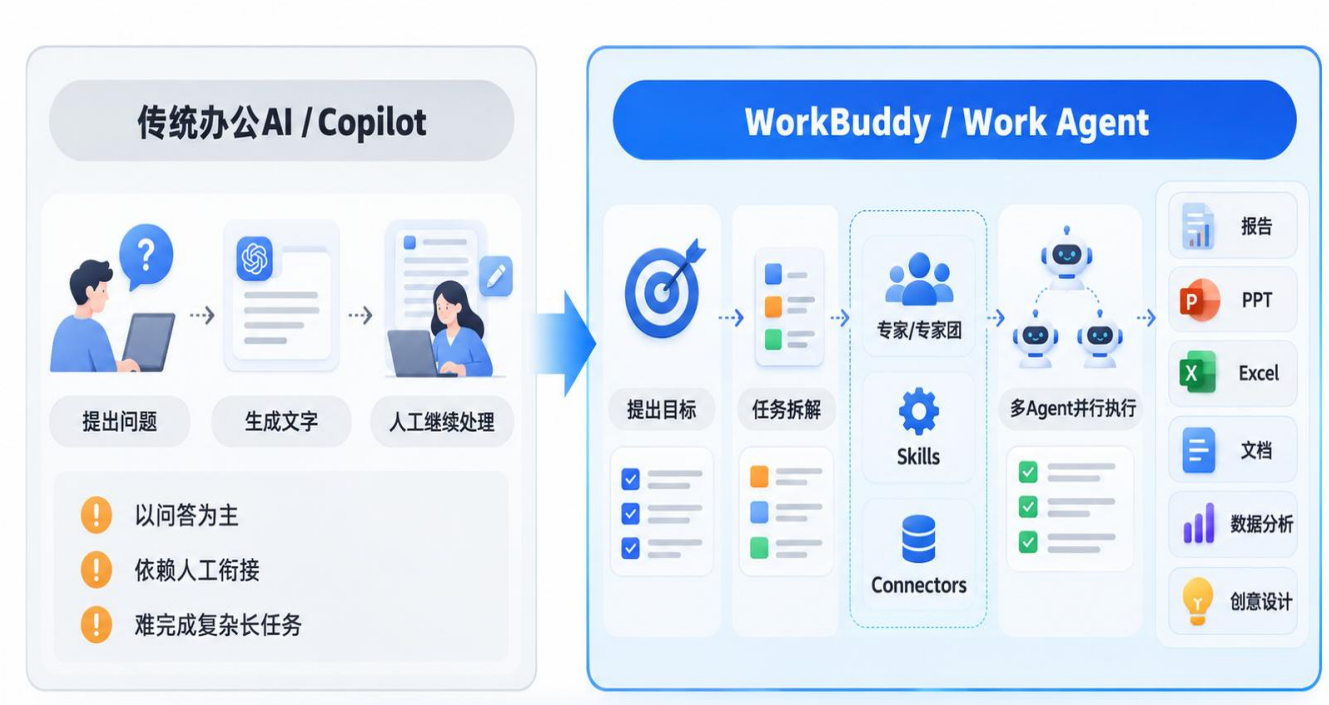
仅供内部参考，请勿外传

3.4 WorkBuddy：从办公助手升级为任务执行型Work Agent

- **WorkBuddy已经由内容生成工具升级为任务执行平台。** 用户提出目标后，系统能够通过专家/专家团、Skills、Connectors及多Agent并行执行完成报告、PPT、表格、数据分析及创意设计等任务，并直接交付工作成果。最新版本进一步增加行业Buddy、资料库引用、后台任务等能力，产品核心由“回答工作问题”转向“理解目标—调用工具—执行任务—交付成果”。

- **与企业数据和工作流的连接是Work Agent的关键能力。** WorkBuddy已支持腾讯文档、会议、乐享、TAPD、邮箱等连接器，并可读取企业微信中的文档、微盘文件、日程、待办、会议及智能纪要；企业版本进一步通过Managed Agents提供独立云端Runtime。WorkBuddy正在形成办公入口—企业数据—任务执行—成果交付的完整工作链路。

图：从Copilot到Work Agent



图：WorkBuddy企业闭环



资料来源：腾讯，国信证券经济研究所整理

仅供内部参考，请勿外传

3.5 腾讯企业生态强化WorkBuddy闭环能力，商业模式向Agent调用与云资源延伸

- **腾讯既有企业服务体系为WorkBuddy提供入口、数据及权限基础。** 企业微信及WorkBuddy构成Agent入口，腾讯文档、会议、乐享知识库、网盘等沉淀企业内部数据和业务上下文，并由Agent直接读写和操作。WorkBuddy由此能够突破通用问答场景，进入企业内部的真实业务流程。
- **Agent执行层与腾讯云进一步形成商业化协同。** WorkBuddy、Skills、Connectors、MCP及Managed Agents负责执行，Managed Agents提供可7×24小时运行的独立云端Runtime，并支持Agent部署、分发及权限管理。商业模式也由单一办公软件订阅进一步扩展至企业席位、Agent调用、云端Runtime、腾讯云资源及企业软件生态增量。

图：生态闭环



资料来源：腾讯，国信证券经济研究所整理

仅供内部参考，请勿外传

国信证券
GUOSEN SECURITIES

4 投资观点：模型效率提升推动调用扩张，Agent打开应用与推理空间

仅供内部参考，请勿外传

4 投资观点：模型效率提升推动调用扩张，Agent打开应用与推理空间

- **模型能力持续提升，效率下降正在成为产业扩张的重要驱动力。** Astra等Frontier模型继续提升复杂推理、Coding及Agent任务能力，海内外厂商同时加快推出Sol、Luna、Flash等高性价比模型，产品体系逐步形成“旗舰模型定义能力上限、中端及Flash模型承接规模化调用、专用模型处理高频决策”的分层结构。随着Token成本、单位任务成本及运行延迟持续下降，模型能力向更多应用和工作流渗透的门槛进一步降低。

- **单位智能成本下降正在转化为更大的真实调用需求。** OpenRouter数据显示，2026年9月日均Token较年初提升近18倍；OpenAI近期调用增长主要由Luna承接，国产模型平台份额超过60%，其中Flash模型占国产Token超过一半。真实调用已经由单纯追求最强模型转向按任务匹配能力和成本。随着Agent、Coding及复杂工作流持续增加单任务调用链长度，Token需求、调用频次及任务复杂度有望共同增长，推理成为AI产业下一阶段的重要增量环节。

- **Agent正在推动AI应用由内容生成进入任务执行和商业闭环阶段。** Consumer / Commerce Agent以Muse为代表，逐步打通用户入口、需求理解、任务执行及支付交易链路；Work Agent以WorkBuddy为代表，开始进入企业数据、权限和真实工作流。两类产品的共同变化是AI由“给出答案”升级为“完成任务”，应用价值也由使用时长进一步延伸至交易规模、生产效率及企业IT支出。

- **投资建议：重点关注两条主线：**1）具备入口、数据、执行能力及商业化闭环的AI应用平台，以及受益于Token需求增长的推理基础设施。 2）应用侧，拥有既有用户流量、企业数据、支付或云生态的平台更容易将Agent能力转化为订阅、交易导流、Agent调用及云资源收入；推理侧，随着调用向高性价比模型、Flash架构及Agent工作负载迁移，模型服务、推理算力及相关基础设施有望持续受益于Token需求扩张。

- **风险提示：**模型能力提升及Agent渗透不及预期、推理成本下降速度不及预期、应用用户留存及商业化低于预期、行业竞争加剧及数据安全与监管政策变化等。

仅供内部参考，请勿外传

附录：大模型产业链相关公司

公司名称	股票代码	上市地点	所属国家	核心大模型/产品	商业模式与业务特点
阿里巴巴	9988.HK	港股/美股	中国	通义千问（Qwen）	商业友好开源，深度融入阿里云生态，高频迭代，为多语言开发主流选择之一。
腾讯	0700.HK	港股	中国	混元大模型	整合微信、游戏、广告、办公等生态，推出AI编程工具等产品。
科大讯飞	002230.SZ	A股	中国	星火大模型	深耕教育、医疗等垂直领域
三六零	601360.SH	A股	中国	360智脑4.0	自研千亿参数通用模型，覆盖全应用场景，在税务、企业服务等垂直行业落地
昆仑万维	300418.SZ	A股	中国	天工大模型、Mureka V8	开放API按Token调用计费，切入“词元经济”变现；同时布局AI音乐、漫剧等领域
智谱华章	02513.HK	港股	中国	GLM系列大模型	“模型即产品”平台路线，以MaaS（API调用）为核心，B端/G端服务为主
MiniMax	00100.HK	港股	中国	M2系列大模型、海螺AI、星野、Talkie	C端订阅+B端开放平台双轮驱动，海外收入占比过半。
字节跳动	未上市	—	中国	豆包大模型（2.1系列）、Seed 2.0/2.1系列、Seedance视频生成（2.5）、Seedream图像模型、TRAE企业级AI Coding、扣子Agent平台	B端+C端双轮驱动：①C端以豆包App为核心（日活超2亿），2026年6月上线付费订阅（68/200/500元/月三档）②B端通过火山引擎MaaS平台提供API调用，2026年MaaS营收目标150亿元（2025年约15亿元）；③视频生成模型Seedance贡献主要企业收入。国内公有云大模型Token调用份额49.5%位列第一，日均Token调用量达180万亿。
月之暗面（Kimi）	未上市（筹备赴港IPO）	—	中国	Kimi K3	2026年7月发布全球参数最大开源模型Kimi K3，在Code Arena超越Claude Fable 5和GPT-5.6 So
深度求索（DeepSeek）	未上市	—	中国	DeepSeek系列大模型	国产开源模型龙头，Token调用量稳居前列。
微软（Microsoft）	MSFT	纳斯达克	美国	Copilot、Azure AI（整合OpenAI技术）	向OpenAI投资超130亿美元，将GPT系列深度整合至全线产品；Azure云提供大模型API服务。
Alphabet（谷歌）	GOOGL / GOOG	纳斯达克	美国	Gemini系列大模型	自研全栈AI模型，云平台（GCP）提供AI基础设施；持有Anthropic约14%股权。
亚马逊（Amazon）	AMZN	纳斯达克	美国	AWS Bedrock、Titan大模型	AWS云平台提供模型托管与分发服务，持有Anthropic股权。
Meta Platforms	META	纳斯达克	美国	Llama系列大模型	自研开源大模型，深度整合至广告推荐与内容生成，为GPU综合应用大户。
Anthropic（Claude）	未上市（已保密递交IPO申请）	—	美国	Claude系列大模型	2026年6月递交IPO申请，估值达9650亿美元，最早2026年10月挂牌；Alphabet（14%）、Amazon、Microsoft参股。
OpenAI（ChatGPT）	未上市	—	美国	GPT系列、ChatGPT、Sora	全球估值最高AI初创，无公开交易股票；微软为最大战略投资者。

资料来源：Wind，公司公告，官网，国信证券经济研究所整理

仅供内部参考，请勿外传

附录：中美主要CSP情况

公司名称	股票代码	云业务品牌	2026年预估资本支出	业务结构
亚马逊 Amazon	AMZN.O	AWS	约2,200亿美元	综合型公有云；重点投入AWS、AI基础设施、数据中心及自研芯片。
Alphabet	GOOGL.O / GOOG.O	Google Cloud	1,950亿–2,050亿美元	TPU、Gemini、Vertex AI与公有云构成全栈体系。
微软 Microsoft	MSFT.O	Microsoft Azure	约1,750亿美元	Azure结合Microsoft 365、GitHub和Copilot；部分租赁分类影响报告口径。
Meta Platforms	META.O	自建AI基础设施	1,300亿–1,450亿美元	主要支持Llama、推荐广告系统、Meta AI及新型AI产品。
甲骨文 Oracle	ORCL.N	OCI	约600亿–700亿美元	聚焦大规模GPU集群、AI训练推理、数据库云和多云部署。
阿里巴巴 Alibaba	9988.HK / BABA.N	阿里云	约1,300亿–1,600亿元人民币	云基础设施、通义千问与AI开发平台一体化。
腾讯控股 Tencent	0700.HK	腾讯云	约1,100亿–1,300亿元人民币	结合微信、游戏、广告和企业服务，发展混元模型与GPU云。
字节跳动 ByteDance	未上市	火山引擎	约2,000亿元人民币；潜在上修至4,000亿–5,000亿元	服务抖音、TikTok、豆包及推荐系统，并向外部提供云和模型服务。
百度 Baidu	9888.HK / BIDU.O	百度智能云	约220亿–280亿元人民币	重点布局GPU云、文心模型、搜索与自动驾驶相关AI应用。
华为 Huawei	未上市	华为云	未单独披露	昇腾、盘古、华为云、存储及网络设备构成软硬件全栈体系。

资料来源：Wind，公司公告，官网，国信证券经济研究所整理

仅供内部参考，请勿外传

附录：美股B端场景垂类AI应用

公司	股票代码	市值（亿美元）	定位/产品	AI逻辑/商业化方式
Snowflake	SNOW	1098	AI Data Cloud、Cortex、Snowflake Intelligence、Cortex Code	企业数据治理、RAG和Agent运行底座；按数据存储、计算和AI用量收费，同时向业务用户延伸AI应用
Datadog	DDOG	1008	云监控、LLM/Agent可观测性、安全监控	监测模型调用、延迟、Token成本、Agent链路与生产故障；按主机、日志、事件和用量计费
Cloudflare	NET	1036	Workers AI、边缘推理、AI Gateway、网络与安全	将推理、缓存、安全和Agent运行放到边缘节点；按请求、算力、存储和安全服务收费
MongoDB	MDB	303	Atlas、向量检索、AI应用数据库	为RAG、AI应用和实时业务系统提供数据库与检索层；按云数据库用量收费
Elastic	ESTC	72	Elasticsearch、向量搜索、企业搜索、可观测性	企业知识检索、RAG、安全分析和日志智能化；订阅＋云用量模式
Palantir	PLTR	3807	Foundry、AIP、Ontology、企业AI操作平台	将模型连接企业数据、权限与实际业务动作；平台许可、项目部署和持续扩容
Microsoft	MSFT	36197	Azure AI、Microsoft Foundry、Microsoft 365 Copilot、GitHub Copilot	同时覆盖云、模型、开发工具和办公入口；云用量＋Copilot席位＋Agent/平台调用收费
Salesforce	CRM	1581	Agentforce、Data 360、Slack Agent、AI CRM	以CRM数据和销售/客服工作流驱动Agent执行；订阅、数据平台及按任务/用量收费
ServiceNow	NOW	1212	Now Assist、AI Control Tower、Autonomous Workforce	在IT、HR、客服、采购等工作流中部署和治理Agent；平台订阅＋AI功能包及用量收费
Adobe	ADBE	1031	Firefly、Acrobat AI Assistant、GenStudio、创意Agent	将生成式AI嵌入创意、文档与营销工作流；订阅＋生成积分/Token＋企业营销产品收费
Intuit	INTU	897	财税、会计、支付和个人金融AI Agent	基于财税与经营数据辅助记账、报税、咨询和风控；订阅＋交易及增值服务收费
HubSpot	HUBS	125	Breeze AI、营销/销售/客服Agent	利用CRM和营销数据自动化获客、内容、销售跟进和客服；订阅套餐＋AI增值模块
Zoom	ZM	295	AI Companion、会议与协作Agent	会议摘要、知识提取、任务分配及协作自动化；基础订阅带动留存，高端功能增值
GitLab	GTLB	61	Duo、代码生成、测试、安全与DevSecOps Agent	结合代码仓库、CI/CD和安全流程提升研发效率；按开发者席位及AI用量收费
CrowdStrike	CRWD	2137	Charlotte AI、终端安全与安全Agent	利用威胁情报和终端数据自动研判、响应和处置；安全平台订阅＋模块化扩容
Veeva Systems	VEEV	347	生命科学云、内容/合规与商业运营AI	基于药企合规数据和研发、商业流程提供AI能力；行业软件订阅＋服务收费
Tempus AI	TEM	84	医疗数据、精准诊疗、临床AI	将临床、病理和基因组数据用于诊断与治疗决策；检测、数据服务和软件收入
C3.ai	AI	15	企业AI应用平台、工业与能源AI应用	面向能源、制造、国防等行业提供预制AI应用；订阅及项目型部署收费
SoundHound AI	SOUN	28	语音AI、车载和客服Agent	语音交互、点餐、车载助手和客服自动化；按部署、调用量或客户合同收费

资料来源：Wind，公司公告，国信证券经济研究所整理

仅供内部参考，请勿外传

附录：A股B端场景垂类AI应用

行业	公司	代码	市值（亿元人民币）	主要AI产品及应用	B端客户
流程工业	中控技术	688777.SH	844	TPT工业时序大模型、生产优化、预测性维护、无人化操作、能耗优化	石化、化工、电力、冶金等流程工业企业
钢铁工业	宝信软件	600845.SH	459	钢铁大模型、MES/ERP/APS智能化、设备运维、质量预测	宝武体系及外部钢铁、制造企业
建筑	广联达	002410.SZ	161	AI算量、AI组价、AI评标、AI清标、方案设计、施工安全	设计院、施工企业、造价机构、建设单位
制造业IT	赛意信息	300687.SZ	111	善谋GPT、制造智能平台、质量/设备/物流分析、企业Agent	制造业大型企业
制造业ERP	鼎捷数智	300378.SZ	99	雅典娜平台、企业AI Agent、生产排程、供应链及经营分析	制造业中大型企业
金融IT	恒生电子	600570.SH	442	LightGPT、智能投研、投顾、客服、运营、合规和知识助手	券商、基金、银行、保险、期货公司
银行IT	宇信科技	300674.SZ	119	信贷、风控、营销、客服、研发智能体	银行及其他金融机构
财税AI	税友股份	603171.SH	203	“犀友”财税大模型、财税Agent、合规税优、涉税风控、经营分析	税务机关、中小企业、财税服务机构
医疗IT	卫宁健康	300253.SZ	172	WiNEX Copilot、病历生成、临床决策、医院运营及医保应用	医院、卫健委、医保机构
法律科技	华宇软件	300271.SZ	47	智慧庭审、审判辅助、法律知识服务、仲裁及破产重组AI	法院、检察院、政法机关、高校及企业法务
城市治理	数字政通	300075.SZ	71	人和大模型、城市智能体、智能执法、12345、城市生命线	地方政府、城管及城市运营部门
网络安全	深信服	300454.SZ	552	安全GPT、威胁研判、自动处置、AI+SASE、私有化AI办公	政企、金融、医疗、教育等
教育/医疗/司法	科大讯飞	002230.SZ	1061	星火大模型、教育、医疗、司法、汽车等行业智能体	学校、医院、政府、车企、央国企
企业数据与AI平台	星环科技	688031.SH	143	数据治理、多模数据库、知识图谱、LLMOps、Agent开发、异构算力管理	金融、电力、交通、政企等行业客户
企业软件平台	用友网络	600588.SH	405	企业AI、Data Agent、财务/供应链/人力等业务智能体	大中型企业、央国企、制造与金融客户
企业数字化服务	汉得信息	300170.SZ	183	ERP智能化、供应链与制造业Agent、企业数据治理	制造、零售、消费等大型企业

资料来源：Wind，公司公告，国信证券经济研究所整理

仅供内部参考，请勿外传

附录：传媒娱乐AI应用产业链

领域	公司	代码	市值（亿元人民币）	主要AI产品及应用	落地场景
AI营销	蓝色光标	300058.SZ	565	AI创意生成、智能投放、数字人及出海营销服务	品牌广告主、电商及出海客户
AI营销/出海	易点天下	301171.SZ	241	AI营销平台、创意生成与投放优化	跨境电商、游戏、应用及AI原生企业
AI营销	因赛集团	300781.SZ	48	因赛AI引擎、营销策划/海报/商品图生成	品牌客户、营销代理及电商商家
创意软件/视频	万兴科技	300624.SZ	110	Filmora等视频创作软件的AI剪辑、生成与Agent功能	中小企业、内容机构、专业创作者；B/C混合
视觉内容/版权	视觉中国	000681.SZ	132	视觉素材版权库、AIGC内容管理、视觉内容服务	媒体、广告公司、品牌方、平台
游戏研发与运营	三七互娱	002555.SZ	448	AI用于美术、文案、客服、投放、运营及玩法探索	以内部游戏研发和发行体系为主
游戏研发工具	恺英网络	002517.SZ	393	SOON全流程AI开发平台：动画渲染、建模、脚本生成	主要内部研发团队；可关注未来工具外溢
游戏AI/内容互动	巨人网络	002558.SZ	571	自研大模型能力、多模态模型、智能体及AI产品探索	游戏研发、运营及玩家互动场景
IP内容/阅读	中文在线	300364.SZ	193	网文IP、AI辅助创作、短剧/动漫等多模态内容开发	内容平台、版权合作方、创作者生态
互动娱乐/游戏	完美世界	002624.SZ	224	AI辅助游戏研发、美术与内容生产	主要内部研发、发行及运营
AI漫剧/IP内容	中文在线	300364.SZ	193	逍遥AI、次元神笔；用于AI漫剧、AI仿真人剧、网文及短剧生产	内容平台、版权合作方、制作团队；兼具C端内容分发
AI短剧/数字阅读IP	掌阅科技	603533.SH	104	AI短剧全流程：剧本生成、视觉制作、成片交付；数字阅读IP转化	短剧平台、内容制作与发行合作方
AI影视/短剧	华策影视	300133.SZ	143	AI漫剧、AI真人剧、AI本地化翻译及国际发行	长短视频平台、海外发行渠道、品牌与内容合作方
AI短剧出海	昆仑万维	300418.SZ	571	DramaWave短剧平台、AI漫剧与AI短剧自动化制作、多语种发行	海外内容消费者、渠道及广告/分发合作方
AI影视制作	百纳千成	300291.SZ	52	AI创作、AI精准营销赋能影视、剧集和互动影视游戏	视频平台、品牌方、内容联合出品方
AI影视/IP	欢瑞世纪	000892.SZ	40	AI漫剧、AI真人剧；剧本孵化、素材生成、镜头设计与后期制作	视频平台、品牌及IP合作方

资料来源：Wind，公司公告，国信证券经济研究所整理

仅供内部参考，请勿外传

风险提示

模型能力提升及Agent渗透不及预期

推理成本下降速度不及预期

应用用户留存及商业化低于预期

行业竞争加剧

数据安全与监管政策变化

经营管理风险等。

仅供内部参考，请勿外传

免责声明

国信证券 GUOSEN SECURITIES

国信证券投资评级

投资评级标准	类别	级别	说明
报告中投资建议所涉及的评级（如有）分为股票评级和行业评级（另有说明的除外）。评级标准为报告发布日后6到12个月内的相对市场表现，也即报告发布日后的6到12个月内公司股价（或行业指数）相对同期相关证券市场代表性指数的涨跌幅作为基准。A股市场以沪深300指数（000300.SH）作为基准；新三板市场以三板成指（899001.CSI）为基准；香港市场以恒生指数(HSI.HI)作为基准；美国市场以标普500指数(SPX.GI)或纳斯达克指数（IXIC.GI）为基准。	股票投资评级	优于大市	股价表现优于市场代表性指数10%以上
		中性	股价表现介于市场代表性指数±10%之间
		弱于大市	股价表现弱于市场代表性指数10%以上
		无评级	股价与市场代表性指数相比无明确观点
	行业投资评级	优于大市	行业指数表现优于市场代表性指数10%以上
		中性	行业指数表现介于市场代表性指数±10%之间
		弱于大市	行业指数表现弱于市场代表性指数10%以上

分析师承诺

作者保证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理解，通过合理判断并得出结论，力求独立、客观、公正，结论不受任何第三方的授意或影响；作者在过去、现在或未来未就其研究报告所提供的具体建议或所表述的意见直接或间接收取任何报酬，特此声明。

重要声明

本报告由国信证券股份有限公司（已具备中国证监会许可的证券投资咨询业务资格）制作；报告版权归国信证券股份有限公司（以下简称“我公司”）所有。本报告仅供我公司客户使用，本公司不会因接收人收到本报告而视其为客户。未经书面许可，任何机构和个人不得以任何形式使用、复制或传播。任何有关本报告的摘要或节选都不代表本报告正式完整的观点，一切须以我公司向客户发布的本报告完整版本为准。

本报告基于已公开的资料或信息撰写，但我公司不保证该资料及信息的完整性、准确性。本报告所载的信息、资料、建议及推测仅反映我公司于本报告公开发布当日的判断，在不同时期，我公司可能撰写并发布与本报告所载资料、建议及推测不一致的报告。我公司不保证本报告所含信息及资料处于最新状态；我公司可能随时补充、更新和修订有关信息及资料，投资者应当自行关注相关更新和修订内容。我公司或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并进行交易，还可能为这些公司提供或争取提供投资银行、财务顾问或金融产品等相关服务。本公司的资产管理部门、自营部门以及其他投资业务部门可能独立做出与本报告中意见或建议不一致的投资决策。

本报告仅供参考之用，不构成出售或购买证券或其他投资标的要约或邀请。在任何情况下，本报告中的信息和意见均不构成对任何个人的投资建议。任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。投资者应结合自己的投资目标和财务状况自行判断是否采用本报告所载内容和信息并自行承担风险，我公司及雇员对投资者使用本报告及其内容而造成的一切后果不承担任何法律责任。

证券投资咨询业务的说明

本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，是指从事证券投资咨询业务的机构及其投资咨询人员以下列形式为证券投资人或者客户提供证券投资分析、预测或者建议等直接或者间接有偿咨询服务的活动：接受投资人或者客户委托，提供证券投资咨询服务；举办有关证券投资咨询的讲座、报告会、分析会等；在报刊上发表证券投资咨询的文章、评论、报告，以及通过电台、电视台等公众传播媒体提供证券投资咨询服务；通过电话、传真、电脑网络等电信设备系统，提供证券投资咨询服务；中国证监会认定的其他形式。

发布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。

仅供内部参考，请勿外传

国信证券
GUOSEN SECURITIES

国信证券经济研究所

深圳

深圳市福田区福华一路125号国信金融大厦36层

邮编：518046　　总机：0755-82130833

上海

上海浦东民生路1199弄证大五道口广场1号楼12楼

邮编：200135

北京

北京西城区金融大街兴盛街6号国信证券9层

邮编：100032