人工智能行业专题（16）

Agent驱动全球模型厂MaaS收入爆发，国产模型各有优势

行业研究 · 海外市场专题

互联网 · 互联网II

投资评级：优于大市（维持）

证券分析师：张伦可
0755-81982651
zhanglunke@guosen.com.cn
S0980521120004

证券分析师：陈淑媛
021-60375431
chenshuyuan@guosen.com.cn
S0980524030003

报告摘要

- 自2026年以来，全球正在经历模型Tokens调用量大爆发的阶段，核心在于模型发展正式进入Agent智能体交互时代。根据OpenRouter数据，截至4月7日中国AI大模型已连续五周调用量超越美国，显示国产模型的崛起。本文主要探讨大模型行业发展趋势、技术、价值链分布等；以及细致梳理国内初创大模型公司和大厂模型的发展现状。

- 国产模型与全球模型厂商比较，主要在性价比领域优势明显。**在国内Deepeek开源基础上，国产模型着重提升工程化、数据能力，来弥补算力限制**。国产模型性价比高主要系：①架构层面算力利用效率领先。以Deepseek为代表的国内厂商针对MoE、注意力机制做了原创性轻量化优化。除此以外，国内厂商（如 Kimi、MiniMax）创始初期（23年左右），锚定长上下文路线，针对超长文本场景做了成本全链路优化。②极致定价，中国厂商愿意以接近成本价来换市场份额。

- **大模型的商业模式较传统互联网，用户网络效应不明显。**AI时代打破传统互联网的发展趋势，用户增长带来的飞轮效应较弱，原因是人为的数据反馈不如机器自己去辨别的更有价值。模型智力水平在当前发展阶段仍是驱动用户、商业化收入增长的最重要因素。

- **通过比较国内初创模型企业和大厂的模型布局，我们认为不同团队之间呈现差异化竞争，比如出身清华人工智能实验室的智谱，更加聚焦模型智能上届的突破，适配国产算力；Minimax从发展第一天起着眼于布局全球市场，成为最早出海并实现可观收入的国产模型厂商之一；Deepseek背靠幻方，商业化压力小，引领国产模型架构创新；Kimi着重构建长上下文能力和侧重Agent智能体集群协同能力。**

报告摘要

- **国产模型的另一派崛起力量来自各个科技大厂。相比起初创公司更加轻盈的组织架构迭代优势，我们观察到各个头部大厂同样积极进行组织调整，提升创新活力与组织调动效率。**从25年开始腾讯、字节以技术突破为核心进行组织调整，26年3月阿里巴巴以Token整合上下游帮助组织快速协调。今年下半年各家发布新模型值得关注，阿里、字节等在加速提升模型的编程能力研究外，有望继续在多模态方向突破；结合最新技术负责人背景，腾讯在Agent能力方向有望快速提升。

- **大厂在25年重点探索多模态方向，26年补齐Agent方向能力后，有望在多模态+Agent方向拥有优势。**长期看，世界模型等是通向泛化AGI必经之路。国内大厂模型在多模态方向拥有优势，比如Qwen3.5实现原生多模态；字节在视觉理解方向拥有优势；腾讯得益于专有数据在3D生成、世界建模有优势。但当前受限于计算资源与数据，聚焦特定领域(如代码)可快速突破能力上限。未来Agent+多模态能力的组合能够帮助：①前端开发全链路提效；②视觉智能体:操作手机/电脑的GUI自动化，空间推理，助力教育科研多模态等。26年4月，阿里发布Qwen3.6-Plus，达到SOTA水平，显著增强了模型的智能体（Agent）编程能力。在前端网页开发，复杂的代码仓库级问题求解表现优秀。

- **投资建议：密切关注国产模型厂商进展，包括阿里巴巴（9988.HK）、智谱（2513.HK）、Minimax（0100.HK）、腾讯控股(0700.HK)。**

- 风险提示：宏观经济波动风险、下游需求不及预期风险、核心技术水平升级不及预期的风险、AI快速迭代平权化下竞争加剧等。

主要模型公司总结

表：核心模型架构、参数量、公司概况

厂商	最新核心模型	总参数量（Est.）	激活参数（Active）	上下文窗口-预测（tokens）	API 价格（1M Token）	公司/技术团队人数	组织情况	特点/路径	最新方向	26年收入
OpenAI	GPT-5.4	/	2000亿+	1.05M	基础版：$2.50（入）/ $15.00（出）	~4000		通过2C场景打造核心壁垒，发力企业与多模态		3月ARR 250亿美元
Anthropic	Claude 4.5	/	1000亿+	1M	$3.00（入）/ $15.00（出）	~4000	Dario Amodei： 前OpenAI研究副总裁，领导了 GPT-2/3 开发	坚持2B路线和Coding场景，在产品布局上相对克制		4月ARR 300亿美元
DeepSeek	DeepSeek-V3.2	6710亿	370亿	128K	$0.27（入）/$0.42（出）	150	组织不超过两层，研究实验室，轻松反内卷	算法架构上积极创新，开源为国产模型提供基石	提高算力利用率：mHC、Engram、国产芯片适配等	
智谱 AI	GLM-5	7440亿	400亿	200K	$1.00(入)/ $3.20（出）	883	脱胎于清华实验室，产学研融合体	学术背景强、全栈自研、幻觉率，拥抱国产算力	聚焦安装配置、代码开发、信息搜集、数据分析、内容创作	3月API平台ARR2.5亿美元
月之暗面	Kimi-K2.5	1万亿	328.6亿	262K	$0.6(入) / $3(出)	300	“共识驱动”，少数精英的共识能引领方向	长下文能力是优势，拥有智能体群	迈向原生多模态与Agent集群	2月ARR 1亿美元
MiniMax	M2.7	2300亿	100亿	205K	$0.3(入) / $1.20（出）	385	商汤背景，“系统理性”，将公司视为可设计和优化的函数	质价比突出，积极拥抱全球模型变化	看好编程、办公、多模态领域	2月ARR达1.5亿美元
小米	MiMo-V2Pro	1万亿	420亿	1M	小于256K：$1(入)/ $3（出）	AI实验室250（24年底）		与终端产品融合	上下文拓展、多模态、底层硬件	
阿里巴巴	Qwen3.5	3970亿	170亿	可扩展至1M	$0.6(入) / $3.6(出)	Qwen 100+，通义实验室600+	26年商业落地期，以Token整合上下游帮助组织快速协调	开源全家桶，架构创新驱动，在国内首先实现原生多模态	提升Coding能力、端到端世界模型	26自然年预计~2100亿元云收入，~500亿AI云（包括MaaS等）
腾讯	Tencent HY2.0	4060 亿	320 亿	256K	￥4.505（入）/ ￥11.13（出）	500+	25H2开始，各业务线的AI研发力量统一整合至混元团队	游戏等积累，在3D生成、图片、世界建模多模态方面有优势	4月对外推出新版本推理和 agent 能力有显著提升	26年随着GPU采购，云收入增速提升
字节	豆包2.0	/	/	256K	Pro版：￥3.20(入) / ￥16.0（出） Lite版：￥0.60（入）/ ￥3.6（出）	1000+	25年转型，技术突破期，研产分离，扁平设计强化基础研究地位	多模态、视觉理解功能性有优势，深度定制工程栈通过性价比抢份额	编程有望今年年中能力快速提升；年底多模态模型	26年预计500亿+云收入，MaaS收入100亿+

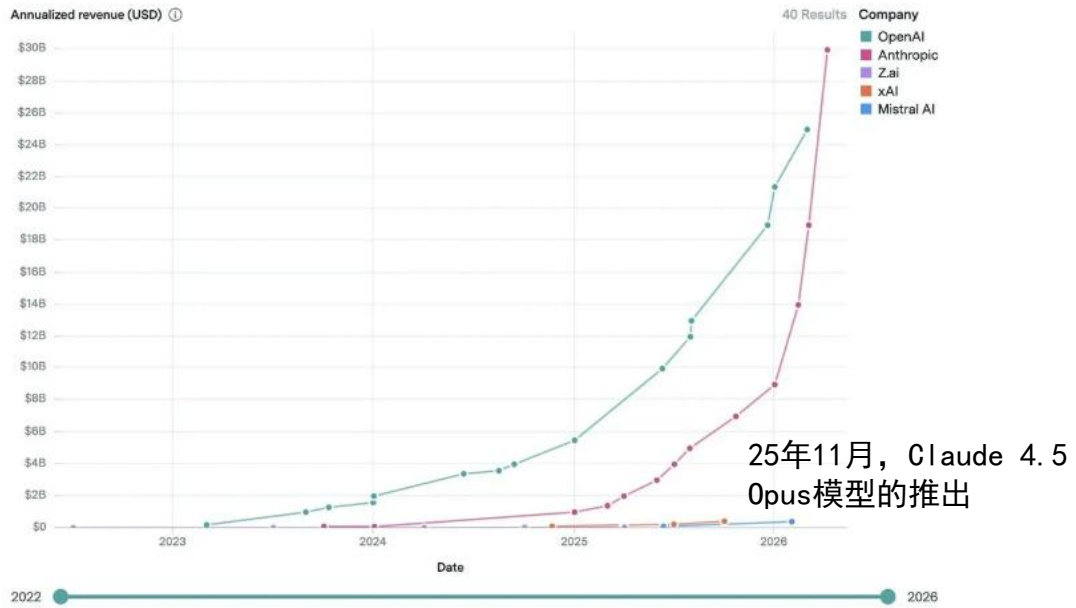
资料来源：官网、Artificial Analysis、晚点、路透社、The Information、国信证券经济研究所整理预测

目录

大模型迭代：26年大模型正式进入Agent时代，从对话走向智能体交互

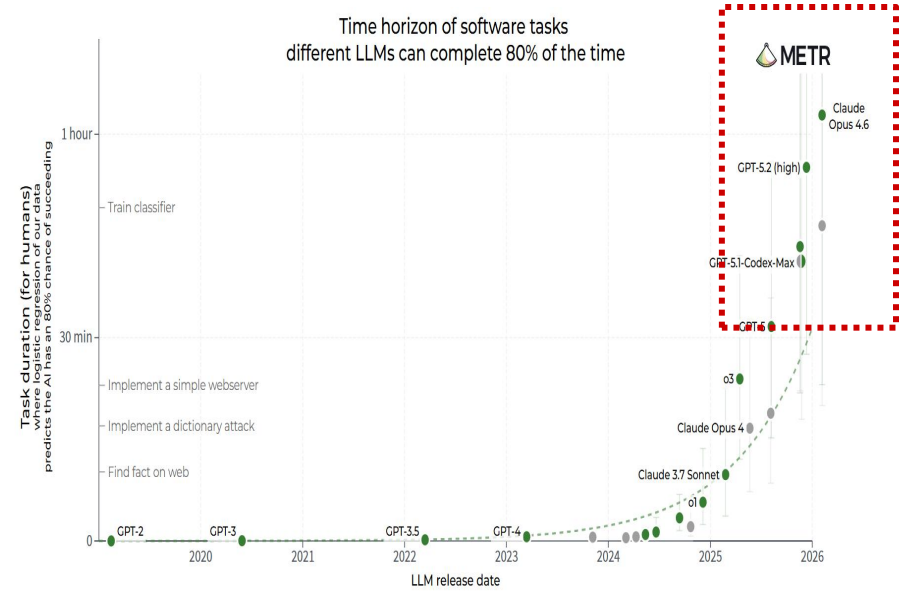
- 根据Semi-Analysis，26年4月，Anthropic ARR超过OpenAI达到300亿美元，成为全球AI收入规模增长最快的大模型公司。
- **编程能力被验证是支持Agent智能体发展的核心能力，26年全球正式进入Agent自主完成任务、调用工具的时代。** Agent将大模型包装起来，加上了记忆（Memory）、规划（Planning）和工具（Tool Use）。Agent能处理的任务时长，正以“每7个月翻倍”的速度指数增长。

图：通用前沿模型智能体能够以 80% 的可靠性自主完成的任务长度



资料来源：Anthropic、The Information、The AI Corner、国信证券经济研究所整理

图：通用前沿模型智能体能够以 80% 的可靠性自主完成的任务长度



资料来源：METR、国信证券经济研究所整理

Token需求侧：任务复杂性提升推动Token量非线性增长

- **随着技术成熟，任务复杂度、Agent占比、多模态任务提升推动Token量非线性增长。**A技术迭代使得大模型向“智能体执行”和全模态技术跃迁，任务复杂度从对话 → 编程 → 智能体 → 视频，Token消耗呈指数级增长。1）智能体：Anthropic实测数据显示单任务Token消耗达普通对话的4倍，多Agent协作则高达15倍。根据IDC预测，活跃Agent数量将从2025年的约2860万快速增长至2030年的22.16亿，这意味着单一用户的调用强度将持续放大。2）多模态：根据火山引擎，生成一段15秒纯生视频就消耗30.9万Token，测算是普通对话21倍。

图：全球企业活跃Agent关键数据预测

资料来源：IDC，国信证券经济研究所整理

图：不同任务类型Token消耗量

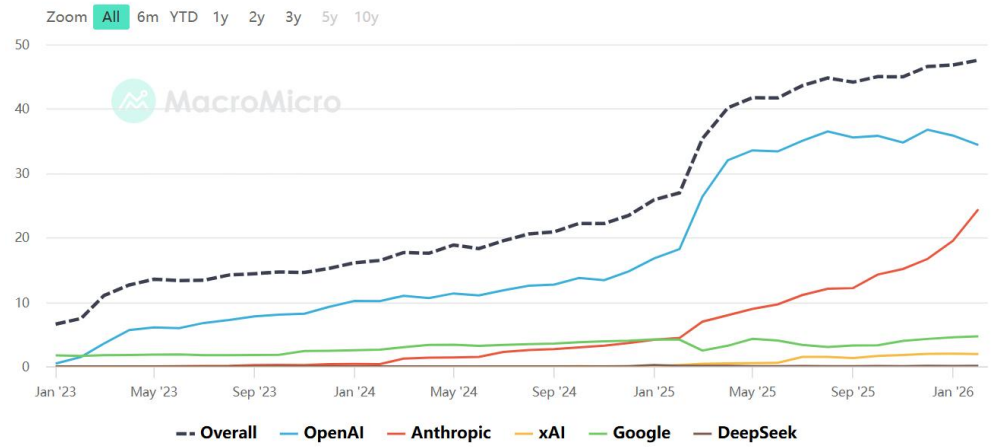
任务类型	典型Token消耗量
普通文本对话	500 – 5,000 Token • 简单问答：~300 Token • 5轮客服对话：1,500-5,000 Token
代码生成/编程	~15,000 Token（生成500行代码）
智能体（Agent）	单Agent：~6,000 Token 多Agent协作：~22,500 Token
视频生成	~308,880 Token（生成15秒标准视频）

数据来源：Anthropic、阿里云、字节跳动、国信证券经济研究所整理

全球企业AI采用率自25年起快速提升

- **OpenAI、Anthropic、谷歌在全球B端大模型中占优。**根据Ramp AI，美国企业AI采用量总体48%，其中OpenAI34%，Anthropic24%。
- **千问、豆包、Deepseek在国内B端市场占优。**根据Frost & Sullivan，在中国大模型B端市场，2025H2，千问（Qwen）系列模型的日均Token调用量优势扩大，占比32.1%，字节豆包为21.3%、DeepSeek为18.4%。，根据国家统计局，截至2026年3月，中国国内日均Token（词元）调用量已突破140万亿，增长超1400倍。

图：美国企业AI采用率：Anthropic赶超OpenAI



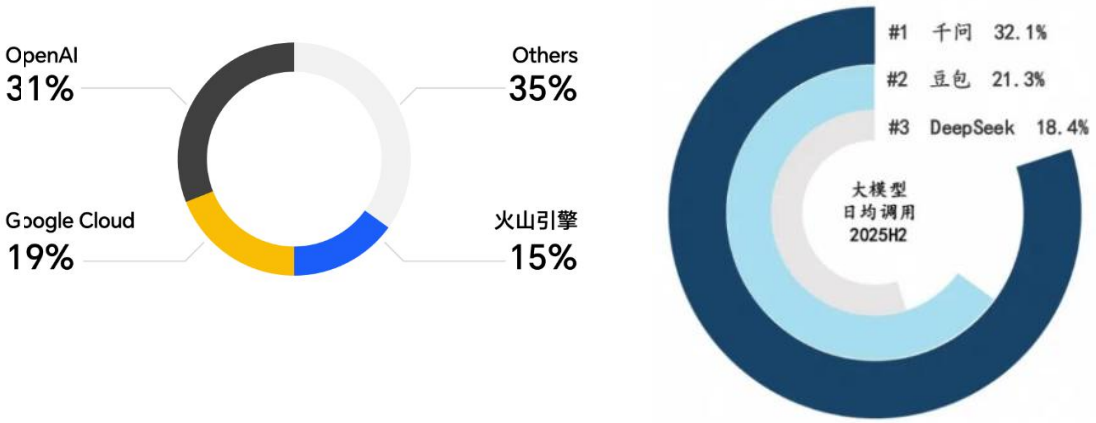
资料来源：Ramp AI Index，国信证券经济研究所整理

图：不同主体日均Token调用量（万亿）

	2025年9月	2025年12月	2026年3月
阿里巴巴			30-40
字节跳动		63	120
国内总体（统计局）		100	140
谷歌	43		
OpenAI			22（仅API）

资料来源：字节跳动、谷歌、OpenAI、国家统计局、国信证券经济研究所整理

图：2025全球企业级MaaS市场占比（左）、2025H2中国企业级大模型日均调用量



资料来源：Omdia、沙利文、国信证券经济研究所整理

Agent能力提升：模型架构、工程化、数据构建多维优势

国信证券 GUOSEN SECURITIES

- 在Transformer架构未发生颠覆性变革的背景下，架构红利边际递减。全球大模型架构基本收敛到 Transformer / MoE。
- **在国内Deepeek、千问开源基础上，国产模型着重提升工程化、数据能力，来弥补算力限制。**在国内算力紧缺背景下，单纯堆叠参数（如从千亿到万亿）带来的能力提升成本过高。工程化能力与高质量数据（思考过程等）是国产大模型提升能力和商业化价值的重要因素。

表：Agent时代大模型的关键能力

核心能力	核心决定因素	关键技术与行业趋势
记忆能力	底层大模型的架构设计+工程设计	短期即时记忆：来自模型原生的上下文窗口能力（内置记忆）。取决于能否发明新的学习算法。①注意力机制类型：Transformer的注意力机制是上下文记忆的物理基础；MoE架构可能影响不同“专家”对记忆的访问；②活跃参数量配比。 长期持久记忆：来自工程化的外部记忆架构 + 工具调用能力（外置记忆）。系统设计决定了实战效果。①外部存储设计 + 检索增强（RAG）能力；②训练优化层：针对性的语料注入 + 强化学习。训练时喂入大量长上下文对话日志、任务流程历史、工具调用 - 记忆检索的交互数据，让模型学习 “如何关联历史信息、何时需要调取记忆、如何精准描述记忆需求”。
规划能力	架构是重要因素，数据决定精度	架构决定了规划能力的 “可能性”。比如使用线性思维链（CoT）or思维树（ToT）+马尔可夫决策过程（MDP）。 规划能力本质上是模型对“因果逻辑”和“状态转移”的理解。这种理解不是天生的，而是从高质量的数据中提取的。①过程监督（Process Supervision）： 仅仅给模型看代码结果是不够的，必须给它看思考过程（Chain of Thought）。②合成数据的质量： 当人类数据耗尽，大模型开始通过“自我博弈”产生数据。 数据清洗和蒸馏的技术也是规划能力提升的重要因素。
工具调用	工程化是短期决定因素，数据决定精度	在工具调用中，失败往往不是模型不懂，而是环境不适配。协议标准化（MCP）： 24年11月Anthropic 推出的 Model Context Protocol（MCP）是工程化突破。它通过标准化的接口，让模型无需针对每个工具重新学习，而是像 U 盘一样即插即用。 数据决定了模型是否能区分细微的参数差别。①预训练阶段：底层认知的基础铺垫。代码与 API 数据：预训练中的大量代码、API 文档、脚本数据，让模型学会了结构化指令、参数传递、函数调用的底层逻辑。②监督微调（SFT）阶段：高质量的工具调用闭环标注数据集。③强化学习阶段：RLVR（可验证强化学习）——Anthropic 的核心训练方式，通过工具执行的客观结果（调用成功 / 失败、参数是否正确、任务是否完成）自动给模型反馈。

数据来源：《2025 全球 AI 训练数据服务市场指南》、Semi-Analysis AI、 国信证券经济研究所整理

测评集：场景洞察与评测集构建的意义在增强

- **传统通用榜单已无法准确衡量 AI 时代的真实生产力价值，模型厂商正转向构建以业务价值为核心的私有化、场景化自定义评测体系。**姚顺雨（OpenAI 前研究科学家）于2025年4月 GitHub 平台发布的一篇博客文章表示：AI正处在“中场休息”阶段，上半场是训练大于评估，下半场将是评估大于训练，重心转向 "定义问题" 。根据晚点访谈Minimax闫俊杰“中国跟美国模型的一个区别，就是缺少内部定义的 benchmark，一些自己的底层思考和设计，更多是在对齐 O1 等模型的输出。”

- 例如智谱发布龙虾场景端到端 Agent 评测基准ZClawBench。腾讯模型新基准 CL-bench，测试大模型“从上下文中学习”的能力。

表：全球主要测评集

能力维度	测评集名称
推理与通用能力	Humanity's Last Exam (HLE)、AIME 2025、HMMT 2025 (Feb)、IMO-AnswerBench、GPQA-Diamond、MMLU-Pro、SimpleQA Verified、AdvancedIF、LongBench v2
编程能力	SWE-Bench Verified、SWE-Bench Pro (public)、SWE-Bench Multilingual、Terminal Bench 2.0、PaperBench (CodeDev)、CyberGym、SciCode、OJBench (cpp)、LiveCodeBench (v6)
Agent 能力	BrowseComp、WideSearch、DeepSearchQA、FinSearchComp (T2&T3)、Seal-0、GDPVal
图像理解	数学推理：MMMU-Pro、MMMU (val)、CharXiv (RQ)、MathVision、MathVista (mini) 视觉知识：SimpleVQA、WorldVQA 感知：ZeroBench (含 / 不含工具)、BabyVision、BLINK、MMVP 文字识别与文档：OCRBench、OmniDocBench 1.5、InfoVQA
视频理解	VideoMMMU、MMVU、MotionBench、Video-MME (带字幕)、LongVideoBench、LVBench
计算机使用	OSWorld-Verified、WebArena

数据来源：Kimi官网、国信证券经济研究所整理

全球头部模型已进入自训练、自进化阶段

- **更好的模型可以导向更好的应用，但更好的应用和更多用户并不会导向更好的模型。**这个现象的底层原理是，在日常使用中，模型比大部分用户更聪明，大部分用户的 query（查询）没有模型自己模拟得好。必须训练好模型，才能够筛选出优质数据。

- **模型具有自进化能力有，全球头部模型已进入自训练阶段。**自进化（Self-Evolution）是指模型能够自主获取、精炼和学习自身生成的经验，形成一个持续自我优化的闭环系统。比如minimax的M2.7在内部测试中，模型可连续执行超过100轮“分析—改进—验证”的循环，自主调整采样参数、优化工作流策略，并在内部评测集中实现约30%的效果提升。公司在部分研发流程中，M2.7已可承担30%至50%的工作量。26年3月，根据小米MiMo大模型负责人罗福莉在谈及AI在未来一年最大的发展变化时表示，一年前大模型实现‘自进化’需要3—5年的历程，现在觉得1—2年就能完成。

表：OpenA VS Anthropic自进化核心路径

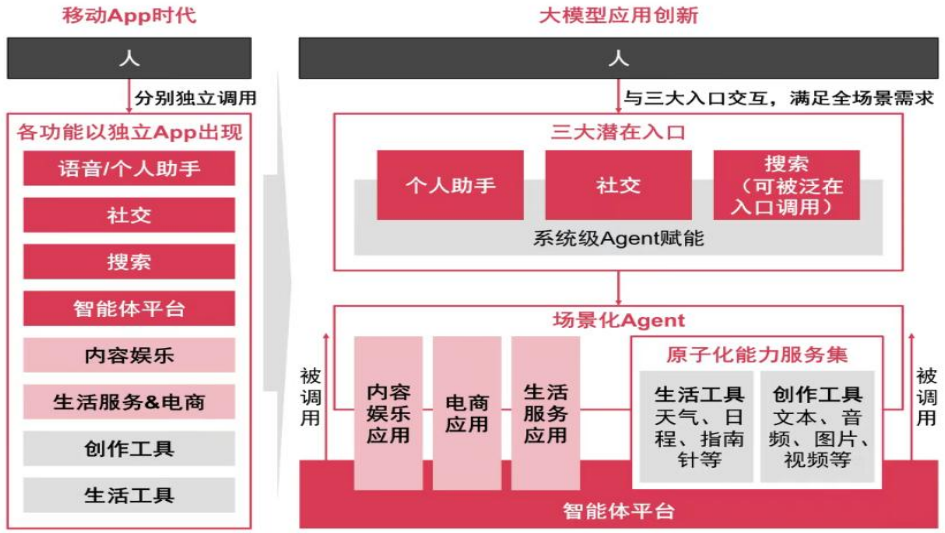
维度	OpenAI 路径	Anthropic 路径
进化驱动力	客观结果（可验证的奖励） 在数学、代码等有明确对错的世界里，通过海量“刷题”和自动批改来进化。	宪法原则（价值对齐） 依据一套成文的伦理与安全准则，不断进行自我批评和修正。
关键技术	RLVR：基于可验证奖励的强化学习。模型生成答案，由编译器/验证器自动判断对错，并用正确的样本训练自己。	宪法AI & RLAIF：模型根据“宪法”生成回答，另一个AI审查员根据同一“宪法”评估并改写回答，从而生成训练数据。
商业逻辑	追求极致能力，以解决更复杂的科学和工程问题为目标，吸引追求技术前沿的开发者与研究者。	将安全转化为信任和收入，通过极高的可靠性和合规性，赢得金融、医疗、法律等高端企业客户。

数据来源：OpenAI、Anthropic、国信证券经济研究所整理

AI Agent 时代：用户获取服务的方式或发生变化

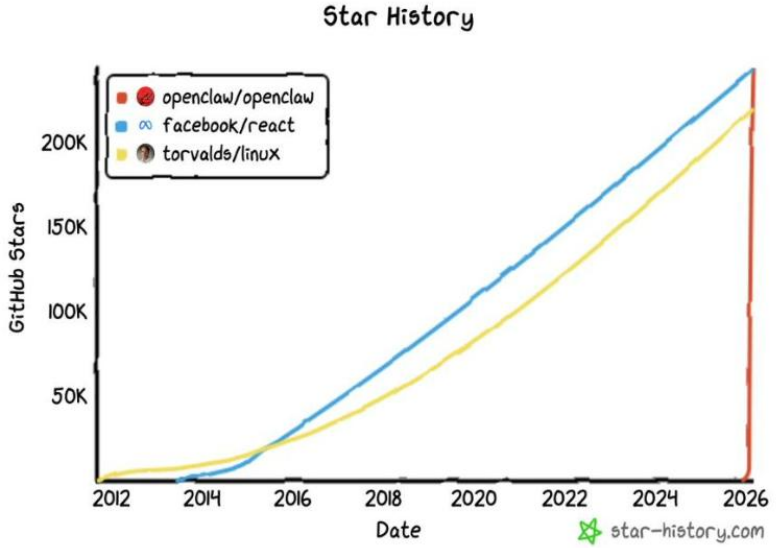
- OpenClaw是智能体计算机的操作系统，OpenClaw的爆火标志着AI从“对话助手”向“执行员工”的范式跃迁。
- **AI时代能力供给、分发模式去中心化。**与传统互联网不同的是，App Store 是「人找 App」，AI时代ClawHub （技能市场）更多是「AI 自主发现 / 调用技能」，用户感知更间接。我们认为一些简单的功能将会被Skill等内生化，沦为原子化服务集，一些电商、生活服务落地基建层价值仍存在。未来技能竞争焦点从争夺App Store曝光，变为：①能否被AI高效调用（接口标准化、可靠性）；②在垂直领域是否不可替代（专业度、数据壁垒、基建壁垒）等。

图：AI时代带来入口形态变迁



资料来源：普华永道、国信证券经济研究所整理

图：OpenClaw 的星数超过了 GitHub 上所有开源软件项目星数



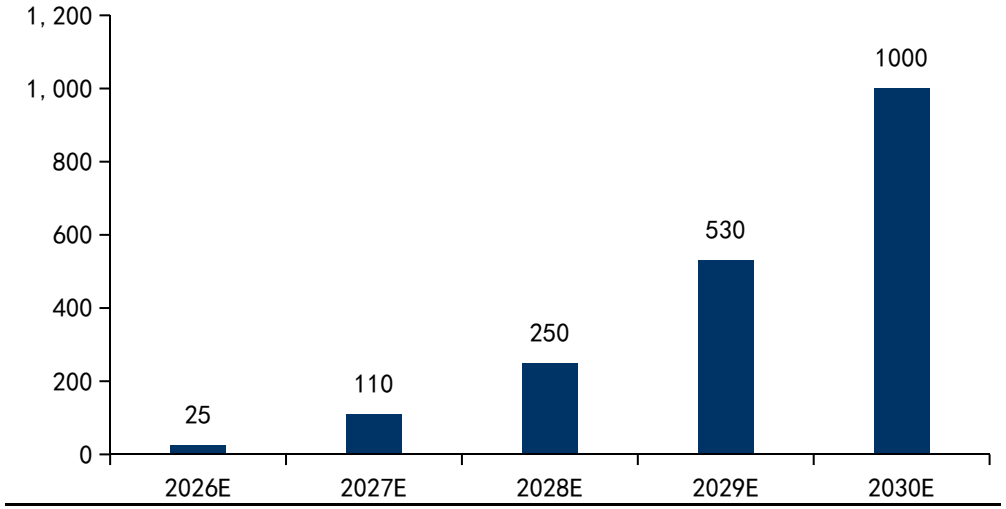
资料来源：Star-history.com，国信证券经济研究所整理

AI Agent 时代：模型巨头开启广告模式，需观察对传统广告市场影响

国信证券 GUOSEN SECURITIES

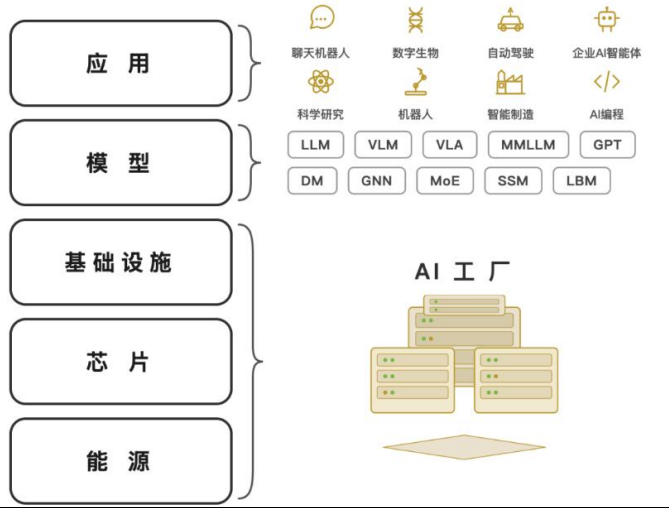
- **大模型基座（规则）：**根据姚顺雨在语言即世界的访谈，底座是少数强大的基础模型（单极），上层是围绕不同交互方式构建的海量、多元Agent应用生态。当前，大模型主要通过出售算力（Token）和模型能力调用许可来收费。比如根据TechCrunch，微软会将 Bing 和 Azure OpenAI 服务约 20% 的收入返还给 OpenAI。
- **AI 智能体/应用（决策）：差异包括Harness等。**①任务调度与编排能力；②私有数据与上下文记忆。③技能生态的集成与优化等。封装的OpenClaw，控制了用户与AI交互的首要入口和调度规则。曾经APP通过掌握中心化分发能力所取得的广告费用有可能受到影响。根据Axios，4月OpenAI计划，2030年广告收入达到1000亿美元，每次对话展示1-2条"Sponsored"推荐卡片，与AI回答物理隔离。

图：OpenAI广告收入预期（亿美元）



资料来源：OpenAI、AXIOS、新浪科技，国信证券经济研究所整理

图：AI是一个五层蛋糕：能源→芯片→基础设施→模型→应用



资料来源：英伟达官网、腾讯科技，国信证券经济研究所整理

Harness Engineering：模型外的“控制与支撑系统”

- **Harness Engineering 是提前把“上下文、约束和验证方式”设计成 AI 可以理解的结构，让模型在环境里自主运行。**包括工具调用、分层上下文工程、长记忆管理、工作流设计等系统工程手段等。
  - 案例：腾讯通过三个维度强化Harness与工程能力。①智能体开发平台（ADP）：通过RAG、知识库为智能体提供“图书馆”；②安全沙箱与Agent Runtime。③全面的安全方案。④ 记忆能力：腾讯云发布“龙虾”记忆服务——TencentDB Agent Memory，接入该服务后，OpenClaw的较原生记忆提升近59%。
- **大模型本身做Harness能够更好更快适配大模型短板。**Harness Engineering是一套动态的、旨在补偿AI模型短板的外部工程体系，本质是随模型能力变化的“补偿面”，未来的竞争力在于快速适应模型变化的能力。Claude Managed Agents 已于2026年4月8日正式开启公测，是一套可组合的API套件，专为在云端大规模构建和部署智能体而设计。

表：Agent 工程三次范式跃迁

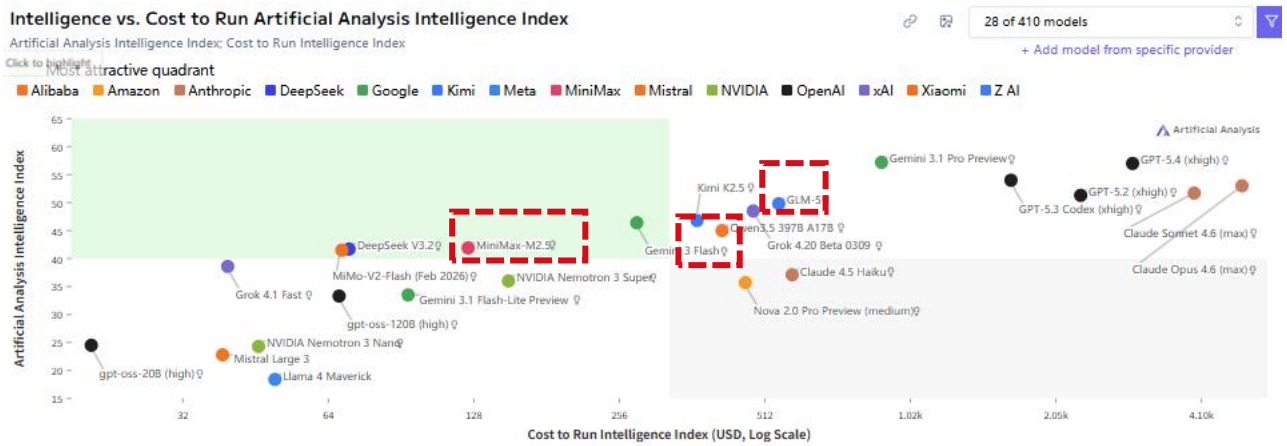
工程范式	兴起时间	核心解决问题	优化核心对象	核心关键技术
Prompt Engineering提示词工程	2022–2024模型早期落地	指令理解偏差、单次输出逻辑差、回答不规范、脑洞跑偏	模型输入文本 / 话术指令	CoT 思维链、Few-shot 少样本、角色设定、模板化 Prompt、输出格式强约束
Context Engineering上下文工程	2024–2025RAG / 长文本爆发	知识盲区、AI 幻觉、私有数据不联动、长对话失忆、专业信息失真	模型可视信息池（上下文、记忆、知识库）	RAG 检索增强、向量知识库、长短记忆管理、动态上下文裁剪、对话历史优化
Harness Engineering驾驭 / 管控工程	2026Agent 规模化时代	自主执行失控、流程不稳定、安全合规风险、故障难回滚、无法商业化落地	Agent全链路运行系统（环境、权限、流程、监控、容错）模型负责做事，Harness 负责让它在很长、很复杂、还容易跑偏的工作流程里，尽量一直做对事	包含以上技术，并关注工作流编排、沙箱权限管控、行为护栏、自检校验、故障回滚、全链路可观测、Agent 调度、合规审计、工具链管控等

数据来源：CSDN、国信证券经济研究所整理

国产模型：性能快速提升，同时质价比突出

- **国产模型在性能上快速提升。**根据Artificial Analysis，25年4月模型总体智能水平排序：Gemini 3.1 Pro>GPT-5.4>Claude Opus 4.6>Meta Muse Spark>GLM 5.1(智谱)>Qwen3.6 Plus>MiniMax-2.7>Grok 4.2>Mimo-V2。
- **国产模型质价比突出，其中最为突出的为Minimax、Deepseek。**从价格来说，国产模型输入输出价格为Claude/GPT的1/5-1/30。

图：通用模型智能水平和价格（截至2026/3/16）：绿色区域最具吸引力

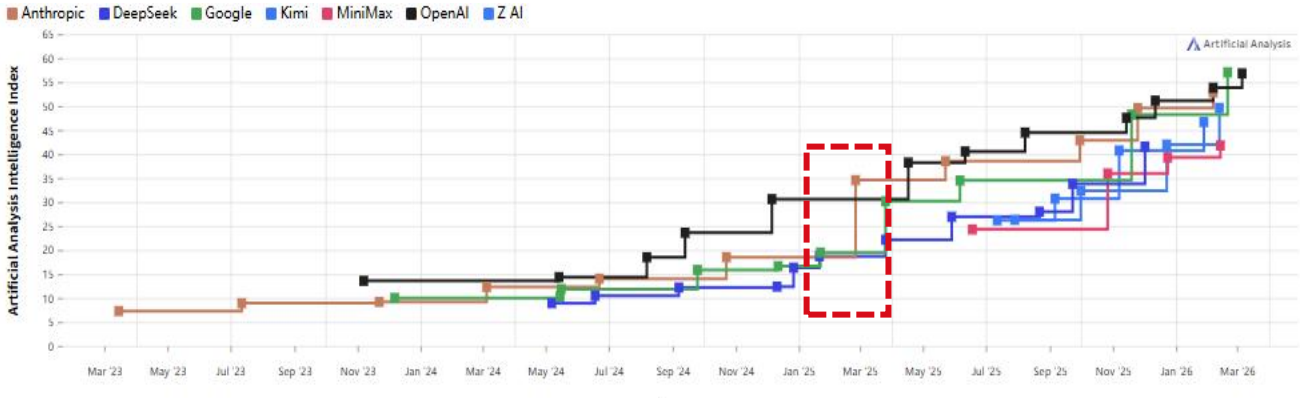


资料来源：Artificial Analysis、国信证券经济研究所整理

图：通用模型智能水平排序（26/3）

资料来源：Artificial Analysis、国信证券经济研究所整理

图：通用模型智能水平迭代轨迹：国产模型在25年底有了明显智能水平提升

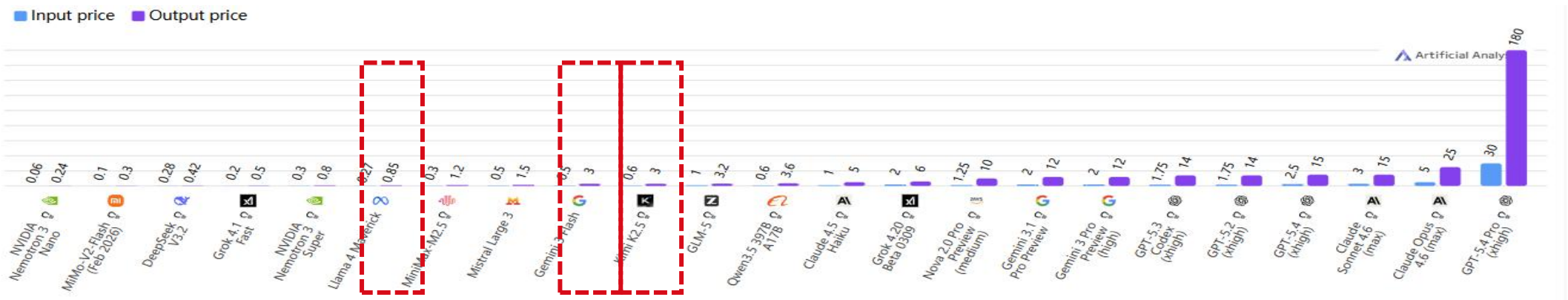


资料来源：Artificial Analysis、国信证券经济研究所整理

国产模型质价比来源：极致工程优化、AI Infra等有优势

- **技术层面的“降本增效”。**国外通过激活更多的参数来确保模型的“智力上限”。即便推理成本高，但只要模型表现出绝对统治力，就能通过商业订阅覆盖成本。国产模型通过更少参数量以及工程化设计让单次推理消耗的显存和算力更低。
  - 国内厂商针对 MoE、注意力机制做了原创性轻量化优化。如DeepSeek 独创 DeepSeekMoE 架构，总参数仅激活 5.5%，算力动态分配效率远超海外同规模 MoE 模型；配套 MLA 多头潜在注意力技术，将 KV Cache 显存占用降至传统架构的 1/10。
  - 长上下文长期探索和链路优化。国内厂商（如 Kimi、MiniMax）创始初期（23年左右），原生锚定长上下文路线，针对超长文本场景做了全链路成本优化。
- **算力平替与调度：** 国产厂商擅长将碎片化的算力资源（如国产芯片、老款芯片）通过集群调度发挥最大效能。
- **AI Infra：**国内电、人工成本更低，算力中心建设周期短，无过多电力限制等。

图：模型输入和输出价格比较（美元/百万token）



资料来源：Artificial Analysis、国信证券经济研究所整理

国产模型劣势：算力限制下推理深度较国外弱，数据集质量较低

国信证券 GUOSEN SECURITIES

- **国产模型弱于国外方面：**①在算力约束下，国产多采用思维链（CoT），线性单向推导的逻辑结构，推理深度和强度弱于海外思维数（ToT）。②数据集规模与基础质量较低，并且数据开源生态弱。

表：国内外数据集差别

对比维度	海外（以美欧为核心）	中国
数据基础质量	1、标注质量高，90% 以上数据集包含「需求 - 推理 - 调用 - 结果 - 反馈」全链路闭环；	1. 标注质量参差不齐，多数数据集仅覆盖「需求 - 调用指令」单环节，缺少完整闭环；
	2.动态工具调用（如 MCP 协议）相关数据集领先行业 1-2 年。	2. 真实用户交互数据积累时间短
标注体系与训练闭环	1. 形成了全行业统一的标注规范，标准化程度极高；	1. 各厂商有独立的 Function Call 规范，标注体系未统一，质量参差不齐；
	2. 异常处理、多轮协同的专项标注数据丰富，能支撑复杂长链路任务。	2. 异常处理、复杂多轮调用的专项数据占比不足，模型应对突发场景的能力弱。
开源生态与社区贡献	1. 开源生态极度活跃，标准先行，重“通用协议”。高校、研究机构、厂商、全球开发者共同贡献；	1. 中国厂商有大量不开源的垂直行业数据。例如，调用钉钉、飞书、或者国产 ERP 系统的复杂数据轨迹。这些数据因为涉及商业隐私，处于闭源状态。
	2. 数据集更新迭代快，能快速适配新技术、新协议（如 MCP）	2. 适配新技术、新协议的版本推出滞后，难以形成行业统一标准。

数据来源：Gartner、国信证券经济研究所整理

表：思维链和思维树的区别

维度	思维链（CoT）	思维树（ToT）
逻辑结构	线性：从 A 到 B 到 C 的单向推导	树状：多路径并行探索 + 回溯择优
技术隐喻	像“逻辑严密的演说家”	像“走一步看三步的棋手”
资源消耗	低：推理延迟小，适合大规模调用	极高：需要多次采样和打分，成本翻倍
中国采用度	极高（主流）	中低（特定场景）
美国采用度	高（已内生化）	高（作为底层搜索算法）
中国为何更偏好 CoT	1. 成本敏感：CoT 在 7B/32B 小模型上配合工程优化即可出效果，性价比极高。	1. 算力约束：ToT 极度消耗算力，在大规模公有云服务中成本压力大。
	2. 落地快：在金融、政务场景中，线性逻辑比复杂的发散搜索更易于审计和解释。	2. 复杂度：ToT 需要额外的“评估模型”，增加了系统链路的不稳定性。

数据来源：《Chain of Thought Prompting Elicits Reasoning in LLMs》、国信证券经济研究所整理

目录

Minimax创始人理念：技术驱动、成立初期开展全球化布局

国信证券 GUOSEN SECURITIES

- **企业愿景：**创业锚定三大核心原则 —— 直接服务用户、全球化布局、技术驱动，坚信AGI时代的核心产品是模型本身。
- **核心战略锚定：24年重新回归技术迭代，坚持第一性原理。**根据晚点2025年1月访谈闫俊杰，2024 年明确”技术迭代优先于短期收入“的核心路线。比如25年1月发布的开源模型MiniMax-01系列用全新“线性注意力”架构，本质是因为公司认为 long context（长上下文）很重要。
- **管理层具有前瞻性眼光，敢于实验新技术框架。**①根据36氪，在全行业死磕模型参数和性能的2023年，MiniMax已经有一款成绩在海外相当亮眼的AI应用—Talkie；②2023年夏天开始研发MoE混合专家架构，投入了80%的算力与研发资源，经历了两次失败才成功。2024年1月，MoE（混合专家）架构还远未成为技术主流，MiniMax又推出了国内首款MoE大模型abab 6。2024年4月，MiniMax开始钻研Linear Attention，并将其与MoE架构融合，成功研发出新一代的基于MoE+Linear Attention的模型。③2024Q2，多模态成了MiniMax布局的重点。

表：Miniamx创始人履历

姓名	职位	加入时间	学历背景	履历
闫俊杰	执行董事、主席、首席执行官兼首席技术官	2022年1月	东南大学数学学院（本科）、中科院自动化所（硕士、博士）、清华大学计算机系（博士后）	MiniMax创始人，曾在商汤集团股份有限公司任职超过六年，担任副总裁及研究院副院长等职位，2022年初创立MiniMax。
员烨祎	执行董事、首席运营官	2022年3月	约翰斯·霍普金斯大学（本科）	曾在商汤集团股份有限公司担任融资与战略投资部经理、首席执行官行政助理及战略部总监、创新业务部总监。
赵鹏宇	执行董事、大语言模型研究与工程负责人	2023年8月	北京大学（本科、硕士）	曾在北京葫芦科技有限公司担任研究级软件开发工程师。
周彧聪	执行董事、视觉模型研究与工程负责人	2022年3月	北京航空航天大学（本科、硕士）	曾任职于商汤集团股份有限公司、华为技术有限公司。

数据来源：MiniMax招股书，国信证券经济研究所整理

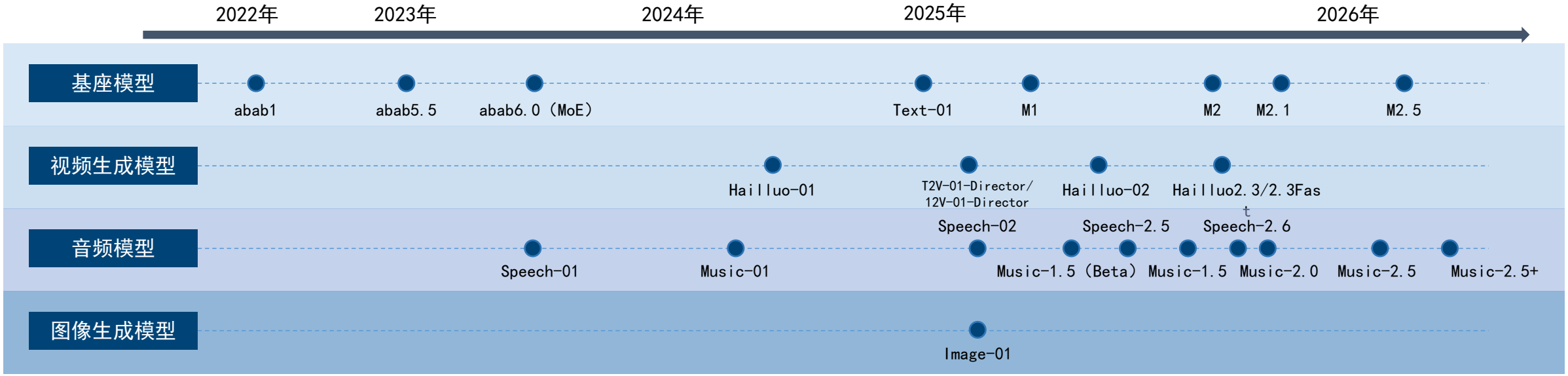
Minimax模型：多模态布局，模型商业化落地领先，迭代速度快

国信证券 GUOSEN SECURITIES

- **全模态布局：视频、音频能力在国内初创公司中领先，25年初发力Agent系列模型。**MiniMax 模型体系涵盖文本、语音、视频、图像与音乐五大方向。M 系列是 MiniMax 2025 年开启的全新模型线，从零开始为编程、Agent、企业级复杂生产力场景原生设计。走 “小步快跑、开源普惠” 的路线，108 天完成 3 次重大迭代。

- **模型核心技术壁垒：MoE 架构商业化领先，线性注意力率先规模化落地。**根据招股书，1）MiniMax是亚洲首家及全球首批实现MoE基础模型架构商业化的公司，这种结构性优势提升了可扩展性和效率，并直接转化为更少的计算需求及更低的推理成本。2）线性注意力机制：全球最早将 Linear Attention（线性注意力机制） 规模化落地的厂商之一。这一创新使公司模型在长文本处理方面表现尤其出色，进一步提升模型效率与可扩展性，亦助力开发更强大的AI agent。

图：Minimax模型推出时间



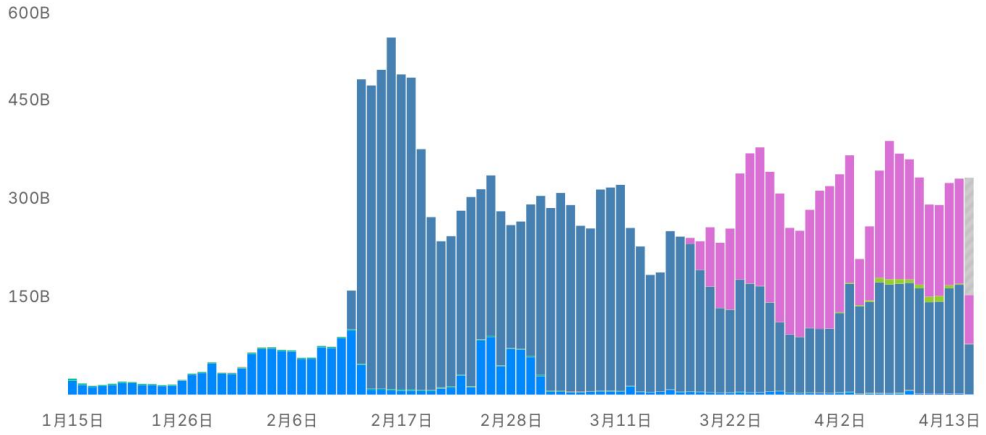
数据来源：Minimax招股书、公司官网，国信证券经济研究所整理

Minimax积极拥抱OpenClaw 浪潮，是极致性价比标杆

- **生态先发优势：OpenClaw 开源项目上线初期即完成底层技术接入，**工具调用能力表现突出，同等任务效果下成本仅为 Claude 的 5%，成为海外开发者首选的高性价比替代方案。

- **极致性价比标杆：**通过前期线性架构等探索以及激活参数量的权衡，M2.5 模型实现 Agent 经济可行性，1美金可支撑 Agent 连续工作1小时，1万美金可支撑4个Agent全年不间断运行，彻底降低AI Agent 商用门槛。从token价格来看，Minimax低于国内外头部模型厂商，百万toekn价格为$0.3（入）/ $1.20（出）；对比来看，Claude为$3.00（入）/ $15.00（出）；智谱为$1.00（入）/ $3.20（出）。截至2026/3/18，Minimax在Openroutertoken调用占比11.3%。根据经济观察报，Openrouter平台近47.17%的用户来自美国。

图：Minimax 调用量趋势（截至2026/4/13）



资料来源：OpenRouter、国信证券经济研究所整理

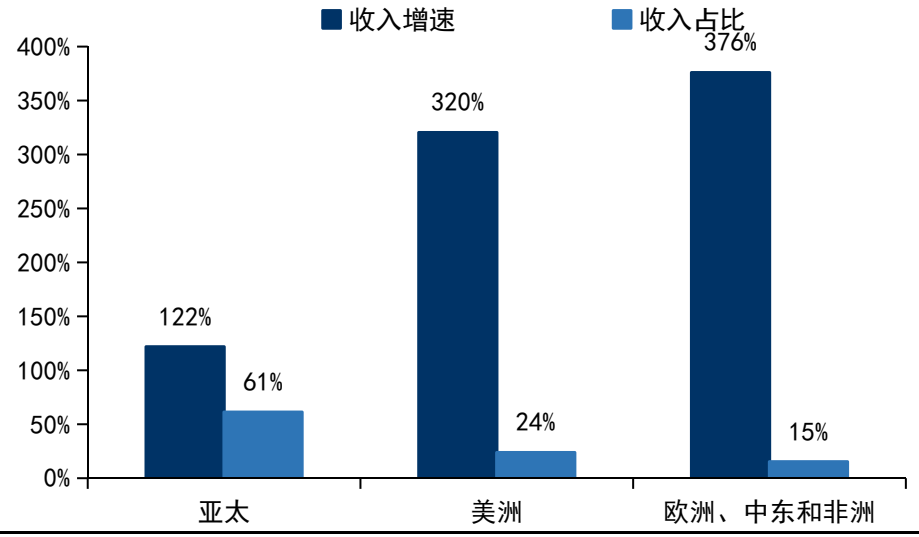
图：创建者 Peter Steinberger 1月 在社交媒体推荐 MiniMax 模型

数据来源：X，国信证券经济研究所整理

Minimax深度拥抱海外市场，积极拥抱全球模型变化

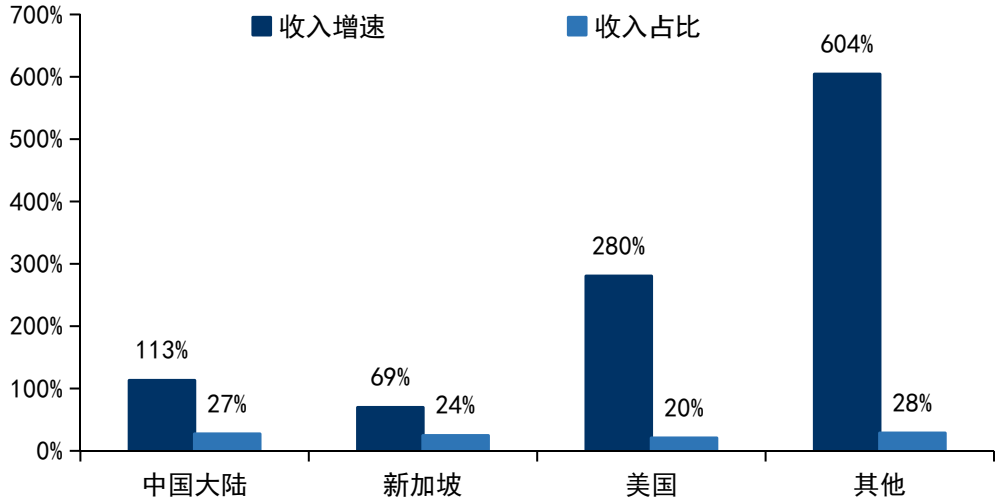
- **Minimax从创业初期，以全球化标准定义模型能力，积极响应国际开发生态。**模型层面在架构设计阶段就做了原生适配，也是国内首个深度兼容 MCP 协议的大模型厂商之一。在2024年底 Anthropic 发布 MCP 协议后，国内厂商大多处于观望状态。MiniMax 反应极快，在 2025 年初就率先官宣了其 MiniMax MCP Server 的上线。
- MiniMax海外收入占比73%。25H1海外市场占比总收入约73.1%。国家来看，新加坡、美国是主要海外市场。25M9，Minimax收入来自中国大陆、新加坡、美国的比例是27%、24%、20%。

图：MiniMax地区收入占比及增速（25H1）



资料来源：MiniMax招股书、国信证券经济研究所整理

图：MiniMax国家收入占比及增速（25H1）

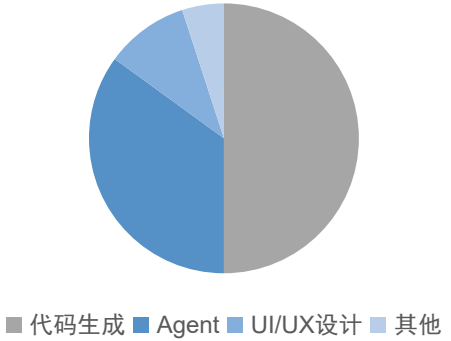


资料来源：MiniMax招股书、国信证券经济研究所整理

Minimax26年迭代方向：聚焦三大核心赛道，深耕生产力场景

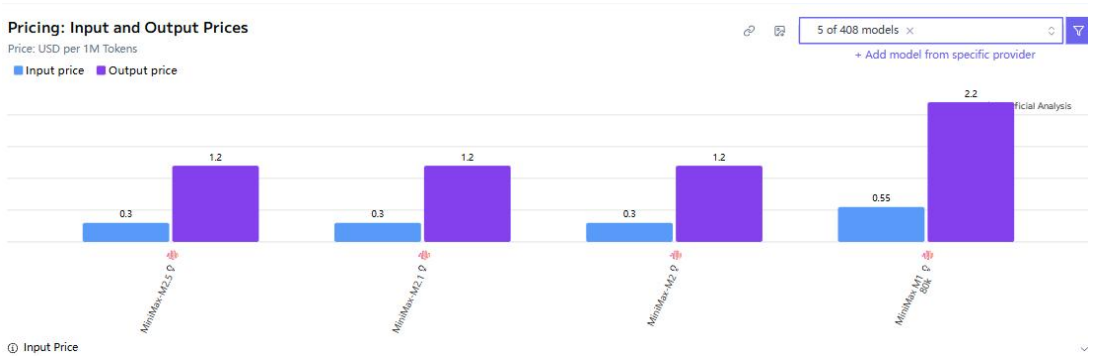
- **根据业绩会，公司2026年发力三大核心迭代赛道：**
  - 编程领域：从辅助 Copilot 向同事级协作升级，未来向 L4/L5 高阶智能突破，持续优化开放平台 API 与编程套餐业务。
  - 办公领域：判断办公领域会复刻编程领域的快速进阶速度，当前M2.7模型 已在 Word/PPT/Excel、金融建模等高阶办公场景实现能力跃升，打造企业级 Agent 核心产品。
  - 多模态领域：聚焦跨模态融合技术，实现可直接商用交付的生成内容，发力中长视频生成，升级海螺 AI 与多模态 API 业务。
- 目前较好的落地场景：①C端：聚焦创意生产、UIUX 设计、代码生成、自动化脚本开发场景；②B 端：聚焦企业多系统数据打通、软件自动化、金融法律流程自动化等企业级场景

图：Miniamx token使用场景



资料来源：Minimax官网、国信证券经济研究所整理

图：Minimax Token价格稳定



资料来源：Artificial Analysis、国信证券经济研究所整理

Minimax业务拆分：开放平台收入增长迅猛

- **Minimax Token调用量增长迅猛。**根据业绩会，2月M2系列文本模型平均单日Token消耗量较25年12月增长超6倍，Coding plan的Token消耗量同期增长超10倍；M2 系列模型每百万Token的推理算力成本下降超 50%。26年2月ARR超过1.5亿美金。
- **开放平台收入增长是主要驱动，预计28年收入占比可达66%。**公司除了开发平台外，还有C端原生产品Talkie（AI陪伴）、海螺（AI视频）等。①海螺AI视频性价比高，除此以外，海螺AI在精准执行用户指令、物理复杂交互方向表现出色。②海外版Talkie与国内版星野是AI陪伴产品，专注于实时人机交互体验。③开放平台：基于AI的企业服务收入，随着全球AI应用落地，Minimax顺应中小创业趋势，26年收入增长强劲。
- **经营杠杆释放带动亏损率收窄。**更高毛利率的开放平台业务增长驱动毛利率向上。

表：Minimax业务拆分(百万美元)——彭博一致预期

	2024	2025	2026E	2027E	2028E	2029E	2030E
总收入	31	79	226	677	1,615	4,252	9,136
yoy		159%	186%	199%	139%	163%	115%
AI原生产品	22	53	137	420	709	1,247	2,077
yoy	2777%	143%	157%	208%	69%	76%	67%
开放平台	9	26	86	292	991	3,005	7,059
yoy	223%	198%	233%	237%	240%	203%	135%
占比	29%	33%	38%	43%	61%	71%	77%
毛利率	12%	25%	30%	35%	42%	45%	49%
AI原生产品	3%	5%	0%	0%	0%	0%	0%
开放平台	63%	67%	59%	57%	62%	55%	55%
研发费用	189	253	463	644	1,015	1,702	2,194
研发费率	619%	320%	205%	95%	63%	40%	24%

数据来源：公司财报，彭博、国信证券经济研究所整理预测

智谱创始人理念：脱胎于清华实验室，聚焦通用大模型长期突破

- 企业愿景：成为实现通用人工智能（AGI）的先行者与 “开路的人”。
- **核心团队：脱胎于清华大学知识工程实验室。**核心团队由国内顶尖 AI 学者与工程专家组成，首席科学家唐杰为清华大学计算机系教授、国内大模型与认知图谱领域权威，形成 “学术研究 - 工程落地 - 场景验证” 的完整闭环。
- **认知领先：对AGI 第一性原理的理解深度领先市场。**坚持长期技术投入，拒绝短期商业化倒逼研发的短视行为。根据26年3月雪球访谈张鹏，“举个例子，某个大厂也是比较早就开始做大模型，做了一段时间之后，它们内部也不是无限制投入，投入完之后被问怎么商业化，团队被逼商业化，结果失败了，团队被替换。”
- **路线差异化：更适配中国市场。**相较 OpenAI 的 “高风险、高投入、高回报” 路线，智谱追求技术的稳定性、可控性、可预期性。将开源视为生态核心，2025 年定为 “开源年”，设立3亿元 Z 基金支持全球AI开源社区，是国内开源大模型的核心标杆。

表：智谱创始人履历

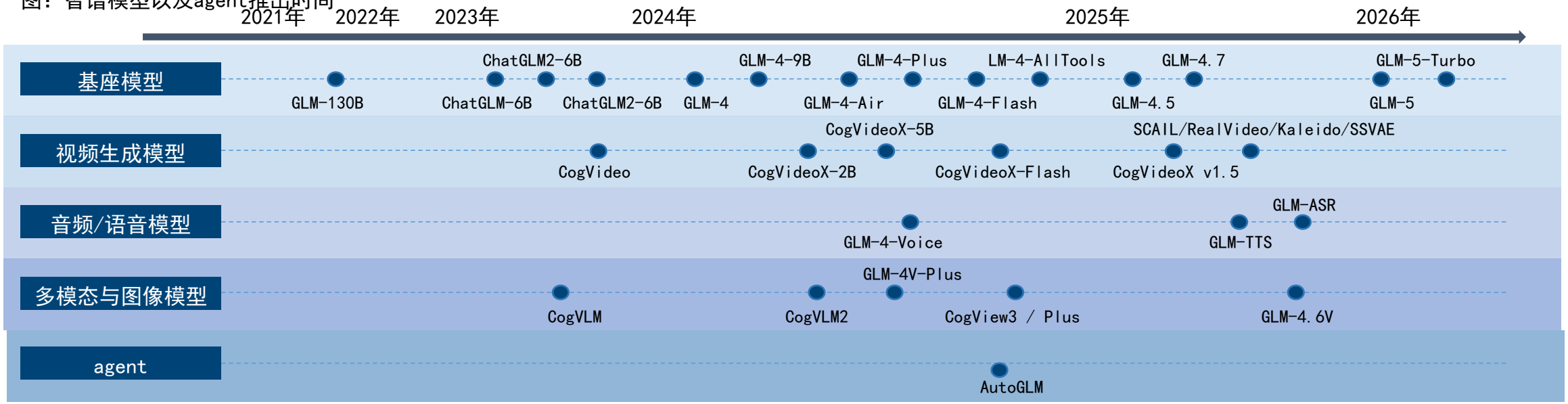
姓名	职位	加入时间	学历背景	履历
刘德兵	联合创始人、执行董事兼董事长	2019年6月	北京交通大学（本科）、中国科学院计算技术研究所（博士）	在计算技术行业拥有近18年经验，曾任职于特艺（中国）科技有限公司北京研究所，后担任清华大学高级工程师。
张鹏	联合创始人、执行董事、首席执行官兼总经理	2019年6月	清华大学（本科、硕士、博士）	在计算机科学领域拥有近20年经验，曾长期任职于清华大学。曾获王选新闻科学技术奖、中国电子学会科技进步奖一等奖、全国工业和信息化系统劳动模范称号。
张笑涵	执行董事	2021年10月	清华大学（本科、硕士）	自2022年7月起，一直担任智谱集团数据标签业务及核心产品“智谱清言”的核心经理。
唐杰	联合创始人、首席科学家	2019年6月	燕山大学（本科、硕士）、清华大学（博士）	清华大学计算机系长聘教授、前系副主任，国家杰出青年科学基金获得者，IEEE/ACM/AAAI Fellow。

数据来源：Wind，国信证券经济研究所整理

智谱模型：全栈自研技术底座，模型稳健幻觉率低

- **全模态模型矩阵：**核心迭代里程碑：2022 年发布国内首个千亿参数开源模型 GLM-130B → 2024 年发布智能体模型 AutoGLM，与全球顶尖厂商同步布局 Agent 赛道 → 2026 年发布 GLM-5 系列，实现代码与 Agent 能力跨越式突破。4月2日，GLM-5V-Turbo发布，面向视觉编程打造的多模态Coding基座模型。操作系统层面或在GLM6迭代。
- **独创架构创新：行业首创融合单向 + 双向注意力机制，**突破传统 GPT 自回归架构文本理解短板，同时解决 BERT 双向编码无法适配生成任务的痛点，在长文本理解、逻辑推理、低幻觉率上具备天然优势。
- **极低幻觉率：**斯坦福大学《2025 年AI指数》显示，GLM-4-9B 模型幻觉率低至1.3%，为全球顶级大模型中幻觉率最低的产品之一。

图：智谱模型以及agent推出时间

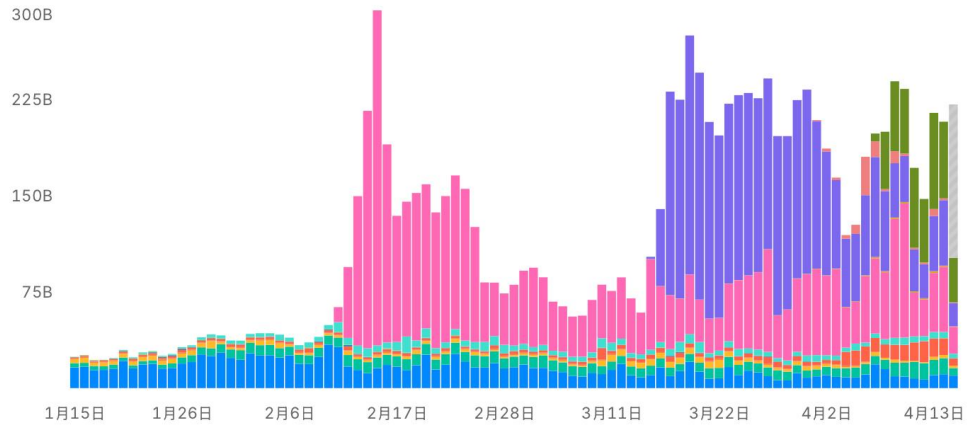


数据来源：公司官网，国信证券经济研究所整理

GLM做国内SOTA模型，为Agent生态打造场景化测评集

- **行业地位：国内开源模型智能水平第一。**截至4/16，Artificial Analysis榜单中GLM-5.1位居全球第六、国内模型第一。
  - GLM核心技术创新：①引入DeepSeek的DSA稀疏注意力机制，大幅降低训练与推理成本；②全新的异步Agent RL基础设施和算法：实现了“生成与训练”的深度解耦。这一算法针对动态环境下的规划与自我纠错能力进行了深度优化。
- **行业首发OpenClaw专属模型：**2026 年 3 月发布GLM-5-Turbo，全球首个面向 OpenClaw 场景深度优化的基座模型。
- **发布龙虾行业评测标准：**发布龙虾场景端到端 Agent 评测基准ZClawBench，填补行业场景化评测空白，通过场景洞察反哺模型能力迭代。任务类型覆盖安装配置、代码开发、信息搜集、数据分析、内容创作等。

图：智谱 调用量趋势（截至2026/4/13）



资料来源：OpenRouter、国信证券经济研究所整理

图：GLM-5-Turbo已接入软通动力旗下机械革命盒子

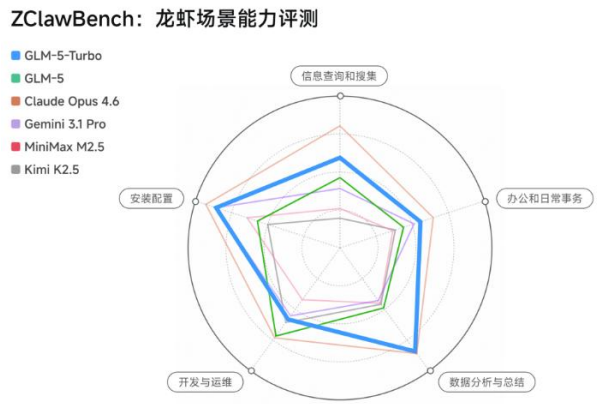
GLM-5-Turbo已接入软通动力旗下机械革命盒子中，面向全球首发接入GLM模型的机械革命“龙虾盒子”，打造原生AI Agent终端体验。

资料来源：公司官网、国信证券经济研究所整理

智谱：海外收入占比小，打造自主可控的AGI全链路生态，拥抱国产算力

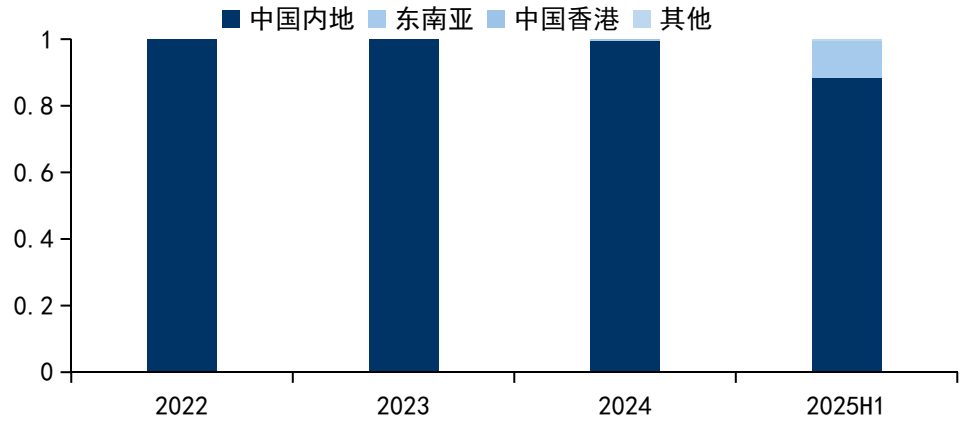
- **国产替代核心布局：受美国实体清单影响，全面推进算力供应链国产替代。**从模型发布伊始，GLM-5原生适配中国 GPU 生态。已完成从底层内核到上层推理框架的深度优化，全面兼容七大主流国产芯片平台：华为昇腾、摩尔线程、海光、寒武纪、昆仑芯、沐曦与燧原。根据业绩会，2026年中左右将发布与国产芯片协同设计的ASIC路线成果，解决算力长期问题。
- 技术布局：①异步长程任务的长度提升；②国产芯片极致优化与深度协同设计；③原生多模态：将VLM能力融入基模。
- 商业端侧，2024年起，智谱开始本地化部署服务海外客户，主要来自东南亚客户。25H1，东南亚客户在本地部署中占比11%。全球化思路：①主权大模型业务，聚焦中东、东南亚等，Q1将在中东达成新签约并推出新的海外主权大模型；②闭源模型与海外第三方算力平台的收入分成。

图：龙虾场景基准ZClawBench



资料来源：官网、国信证券经济研究所整理

图：智谱全球收入拆分

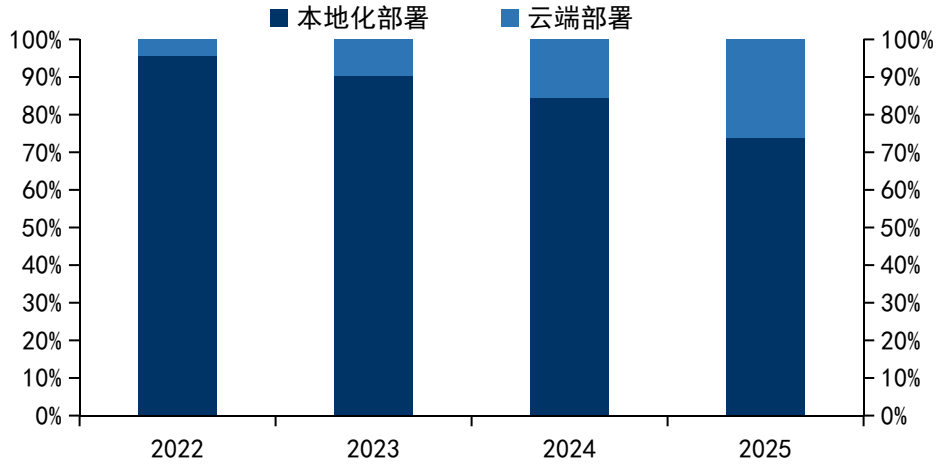


资料来源：官网、国信证券经济研究所整理

智谱：云端 MaaS 成核心增长引擎

- **核心商业模式：MaaS云收入增长迅猛。**2025H1本地化部署收入占比85%，云端部署占比15%。2025年整体云端部署占比26%。
- **国内公司模型服务于科技与互联网、公共服务、传统企业等。**智谱客户中互联网企业为第一大客户，25H1占比38%，公共服务与政企为第二大客户占比29%。在当前算力紧缺背景下，智谱优先满足互联网等大客户需求。前十大互联网公司有9家接入智谱模型，是字节调用量最大的独立第三方模型。
- 启动 “龙虾全国部署计划”，为国内政企客户提供上门安装服务，快速抢占 B 端市场。

图：智谱收入拆分



资料来源：官网、国信证券经济研究所整理

表：智谱下游收入拆分（25H1）

客户行业分类	收入占比	核心定位与说明
互联网与科技行业	38.3%	仍为第一大客群，覆盖头部互联网平台、科技厂商、软件与 SaaS 服务商
公共服务与政企类	29.4%	第二大客群，核心为政府机构、国央企、公共事业及国家级科技专项项目
电信行业	13.6%	传统行业第一大细分赛道，核心为国内三大运营商及通信服务商
传统企业	12%	传统行业核心赛道，覆盖金融、工业制造 / 能源等。核心场景为工业智能、生产调度、设备运维
消费电子	6%	用户运营、智能座舱、产品智能化升级等

数据来源：公司官网，国信证券经济研究所整理

智谱业务拆分：API用量在提价下持续爆发

国信证券 GUOSEN SECURITIES

- **API 用量爆发式增长。**26年3月，API平台业务ARR突破2.5亿美元，预计26年底ARR突破10亿美元。26年YTD提升6.4倍，YTD付费Token消耗提升4倍。GLM-5 系列模型能力跻身全球前列，API 调用量持续攀升，我预计API收入占比目标从2024年的 16%提升至26年42%，云端 MaaS 业务成为核心增长引擎。

- **产品提价助力毛利率提升。**2026年2月智谱对 Coding Plan 套餐整体涨价 30% 起，市场需求仍保持高景气，提价动作直接带动单 Token 价值提升，助力毛利率稳定。除此以外，公司通过模型量化技术和专用推理芯片（与国产算力链深度耦合），单次 Token 调用成本持续下降。25年公司毛利率41%，其中本地化部署/云端部署毛利率49%/19%。

图：智谱token价格变化

资料来源：Artificial Analysis、国信证券经济研究所整理

表：智谱业务拆分(百万元)——彭博一致预期

	2023	2024	2025	2026E	2027E	2028E	2029E	2030E
总收入	125	312	724	2,806	6,702	15,677	32,177	56,380
yoy		150%	132%	287%	139%	134%	105%	75%
本地化部署	113	264	534	1,163	2,151	3,732	7,519	11,363
yoy	90%	134%	102%	118%	85%	74%	101%	51%
云端部署	12	48	190	1,821	4,999	13,419	26,298	44,268
yoy		307%	293%	856%	175%	168%	96%	68%
占比		16%	26%	65%	75%	86%	82%	79%
毛利率	65%	56%	41%	36%	38%	42%	41%	43%
本地化部署	68%	66%	49%	53%	54%	55%	59%	59%
云端部署	31%	3%	19%	21%	25%	31%	32%	39%
研发费用	529	2,195	3,180	4,787	6,527	9,379	14,968	15,720
研发费率	423%	703%	439%	171%	97%	60%	47%	28%

数据来源：彭博，国信证券经济研究所整理

月之暗面创始人理念：起源于对AGI的本源好奇，重视长上下文能力

- **月之暗面创始源于AGI 探索的本源好奇，2026年Kimi会成为一个“与众不同”和“不被定义”的LLM。**根据25年底内部信，杨植麟表示：“Kimi的起点很简单，就是单纯的好奇，好奇AGI的上限在哪，好奇我们心中的理想模型是什么样。”**2025年初全面战略回归模型研发。**2024年，该公司在营销上的投入接近9亿元人民币。25年春节DeepSeek的成功让团队对之前的决策进行了“深刻乃至痛苦的反思”。停止市场营销与买量行为，集中全部资源聚焦基础算法与核心模型研发。
- **核心技术信仰：将长上下文类比为 “AI 时代的内存”。**更长的上下文是 AI 实现个性化交互、解决复杂任务的核心基础。
- **清晰业务边界：**根据晚点，月之暗面总裁张予彤表示，与资源更多的大公司竞争时，他们会刻意控制业务边界，比如不做生活娱乐方向、不做多模态生成业务等。“专注大模型层、逻辑层、Agent 层，以及深入研究、PPT、数据分析、网站开发这类偏生产力、偏复杂任务的链路”。

表：Kimi创始人履历

姓名	职位	加入时间	学历背景	履历
杨植麟	创始人、首席执行官（CEO）	2023年3月	清华大学计算机系（本科）、卡耐基梅隆大学（博士）	中国35岁以下NLP领域引用最高的研究者，Transformer-XL和XLNet论文第一作者。曾就职于Google Brain、FAIR，参与研发Google Gemini、Google Bard、盘古NLP、悟道等大模型。国家"万人计划"青年拔尖人才，智源青年科学家。
周昕宇	联合创始人	2023年3月	清华大学	与杨植麟、张宇韬为清华计算机系同班同学。毕业后加入旷视科技，从事算法量产工作。以共同一作身份发表ShuffleNet论文（CVPR），该成果影响了苹果3D人脸解锁等手机毫秒级人脸识别技术。持有月之暗面10%股份。
吴育昕	联合创始人	2023年3月	清华大学	与杨植麟、周昕宇同为清华背景，Google Scholar引用超万次，在大模型方面有丰富的工程和算法经验。曾参与研发 detectron2 视觉开源项目。部分资料显示其为早期联合创始人之一。
张宇韬	联合创始人、首席技术官	2023年3月	清华大学	研究方向为异构数据融合和知识图谱构建，在KDD、CIKM等计算机顶会发表多篇论文。曾任科技大数据分析平台AMiner技术负责人。持有月之暗面5%股份。

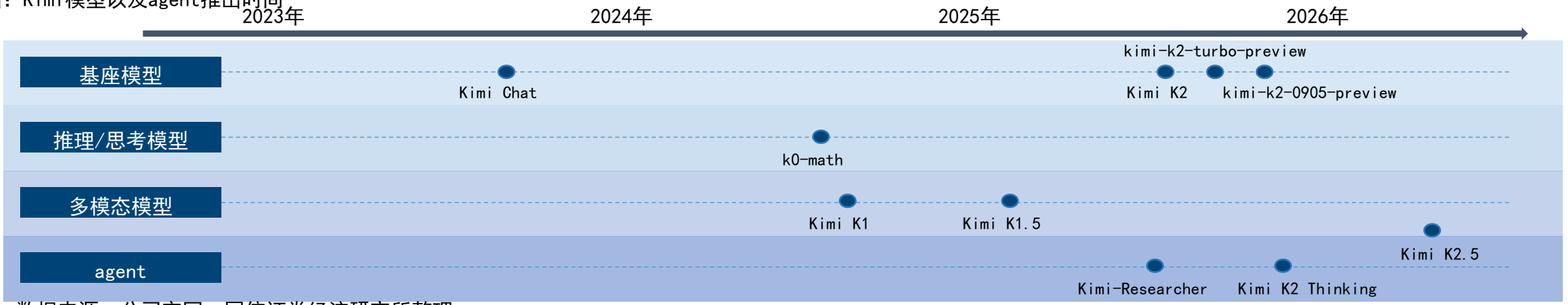
数据来源：公司官网、国信证券经济研究所整理

月之暗面模型：上下文原生先发壁垒，Agent 集群技术国内领先

国信证券 GUOSEN SECURITIES

- 2025年完成从“长文本工具”到“开源思考型智能体”的跨越式迭代。①2023-2024 年：Kimi Chat 上线，奠定超长上下文基座；②2025年：发布万亿参数MoE架构系列模型；③2026年：发布集大成者K2.5，原生多模态架构 + Agent 集群机制。
- **长上下文原生先发壁垒：**行业内极少数从成立之初就锚定 “长上下文原生” 路线的厂商，底层注意力机制、训练与推理框架全链路围绕长文本场景设计，而非后期补丁式扩展；是国内最早将长文本能力推向 200 万字量级的模型。
- **Agent Swarm 智能体集群技术：**K2.5 搭载国内领先的智能体集群机制，可自主将复杂目标拆解给最多 100 个子智能体并行执行，通过并行智能体强化学习（PARL）实现任务自动分工与调度，彻底解决长链路复杂任务的执行痛点。
- **多模态融合创新技术：**独创低比例早期融合训练策略，解决多模态模型常见的 “模态冲突” 问题；通过零视觉监督微调，突破人工标注数据瓶颈，大幅提升模型视觉理解与工具调用的泛化能力。

图：Kimi模型以及agent推出时间



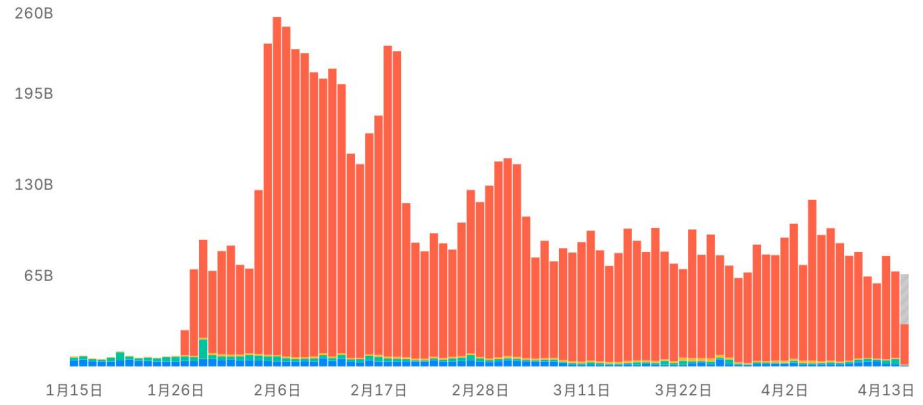
数据来源：公司官网，国信证券经济研究所整理

Kimi商业化：公司拥有C端产品Kimi对话助手，海外收入占比超过国内

国信证券 GUOSEN SECURITIES

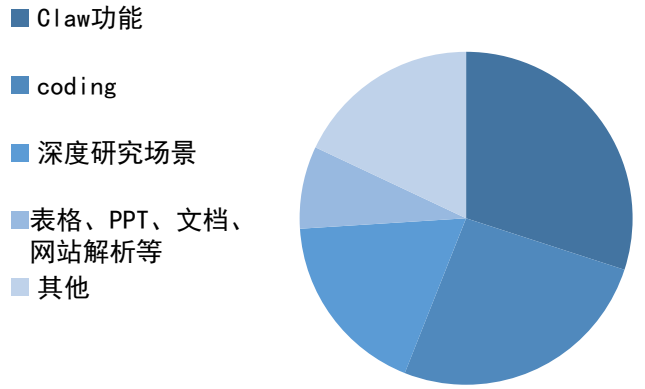
- **B端+C端双轮驱动。**根据每日经济新闻，受K2.5模型及Kimi Claw智能体上线推动，2026年1月底以来的约20天内，其累计收入已超过2025年全年总和。1）B 端：以 API 调用与企业定制化服务为核心，依托长上下文与 Agent 能力，服务互联网、金融、科研、制造等行业客户。2）C 端：以 Kimi 对话助手为核心载体，免费用户即可解锁核心长上下文功能，支持 100 + 文件、20 + 格式、200 万字内容一次性解读；搭建 49-699 元 / 月四档成熟会员体系，提供增值服务。
- **国内外：目前其海外收入已超过国内。**根据Readhub消息，Kimi K2.5模型发布（26/1/27）仅一个月，公司年度经常性收入（ARR）正式突破1亿美金。根据Similarweb、AI产品榜等，我们测算，国内流量占比62%，海外占比约38%。根据官网价格分析，我们推测海外收入贡献已超过国内，因为海外高阶订阅客单价、付费意愿高于国内。

图：Kimi 调用量趋势（截至2026/4/13）



资料来源：OpenRouter、国信证券经济研究所整理

图：Kimi token使用场景

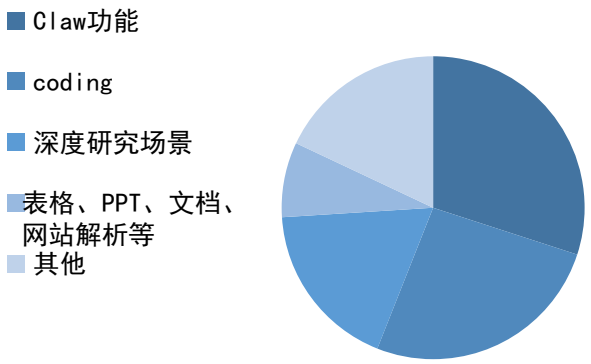


资料来源：官网、国信证券经济研究所整理

Kimi迭代方向：Token 效率、长上下文、智能体集群协同

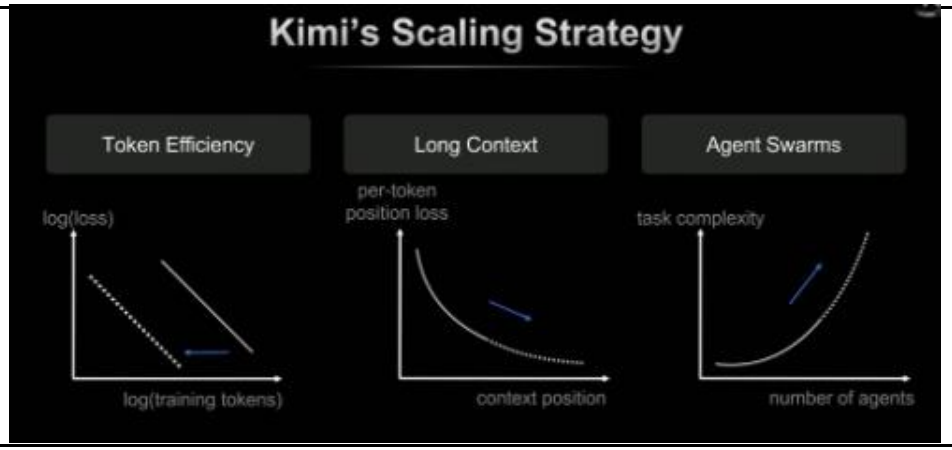
- **Kimi模型进化逻辑可以归纳为三个维度的共振：Token 效率、长上下文以及智能体集群（Agent Swarms）。**根据杨植麟GTC 2026演讲，Scaling不再是单纯的资源堆砌，而是要在计算效率、长程记忆和自动化协作上同时寻找规模效应。
  - 下一代框架：block attention residual注意力残差。根据钛媒体，传统的残差结构是通过对每一层的输出进行统一求和来实现信息传递，新的框架几乎不损失精度的情况下，降低计算和通信开销。允许模型在每一层选择性地关注此前各层的输出，而不是简单地进行求和。报告显示，经过改进的48B模型训练效率提升了1.25倍。
- **场景深度渗透：持续优化编程、办公自动化、深度科研三大核心生产力场景。**①编程，尤其是前端开发：K2.5能根据自然语言指令生成具备专业设计感的完整前端界面，并支持通过截图或录屏分析交互逻辑，直接复现代码。②办公自动化：深度集成Word、Excel、PPT等办公软件，用户通过自然语言指令即可完成复杂的数据处理、文档转换和演示文稿制作。

图：Kimi token使用场景



资料来源：官网、国信证券经济研究所整理

图：Kimi三个维度的共振：Token 效率、长上下文以及智能体集群（Agent Swarms）



资料来源：GTC大会、国信证券经济研究所整理

Deepseek创始人理念：商业化压力小，目前重视前沿基础研究

国信证券 GUOSEN SECURITIES

- **商业化压力小，早期重视前沿研究探索。**Deepseek依赖母公司幻方量化的利润输血，幻方作为顶级量化私募，能提供稳定且充足的研发预算。根据暗涌24年7月访谈，梁文峰表示Deepseek专注于做基础研究。从长远来看，希望建立一个生态系统，让行业直接使用Deepseek技术和成果，其他公司基于Deepseek模型开发B2B/B2C服务。“未来的世界很可能是一个高度分工协作的世界。基础AI模型需要持续创新，而大公司也有自身的局限，并不一定最适合承担这一角色。”

- **Deepseek真正的护城河在于团队的成长——积累技术Know-how，培养创新文化。**开源和发表论文不会带来重大损失。对于技术人员来说，被同行追随本身就是一种成就。
  - MLA创新的诞生：扁平化的组织结构有关。根据晚点，Deepseek研究团队只有梁文锋和其他研究员两个层级。根据暗涌，MLA创新最初是某个年轻研究员的个人兴趣，在总结主流Attention架构的关键演化规律后，突然灵感迸发，设计出了一种新的替代方案。但从想法到现实，是一个漫长的过程。Deepseek组建了团队，花了几个月时间验证可行性。
  - 根据晚点，2025年秋天起，梁文锋也开始更多提产品化和商业化。DeepSeek 已有小数十人的产品团队，但尚未涉足 AI 编程、通用 Agent 等热门应用方向。

表：Deepseek创始人履历

姓名	职位	加入时间	学历背景	履历
梁文锋	创始人	2023年7月	浙江大学	2008年起探索AI量化交易，2015年创立幻方量化，管理规模曾超千亿。2019年起投入巨资建设“萤火一号/二号”万卡A100超算集群。年科学家。

数据来源：证券时报、国信证券经济研究所整理

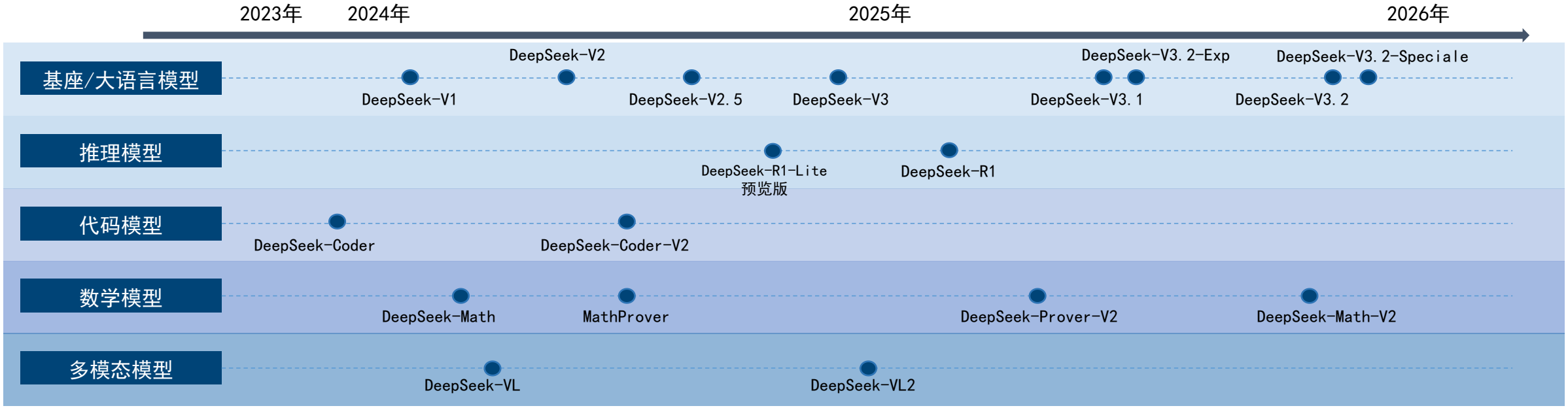
Deepseek模型：全栈自研模型矩阵，持续迭代领跑开源赛道

国信证券 GUOSEN SECURITIES

- **核心迭代里程碑：**2023 年 V1 模型上线→2024 年发布 V2/V3，完成 MoE 架构与核心技术突破→2025 年 R1 推理模型发布，能力对标 OpenAI o1 系列→2026 年 V3.2 持续优化。

- **DeepSeek近期正重点发力原生多模态融合与国产算力深度适配。**新一代旗舰DeepSeek V4预计于2026年4月发布：①将原生支持图像、视频和文本的联合理解与生成，并引入长期记忆机制。②V4在硬件适配战略上首次优先向华为昇腾、寒武纪等国产芯片开放早期访问权限，旨在成为首个完全跑在国产算力生态上的核心大模型。

图：Deepseek模型推出时间



数据来源：公司官网，国信证券经济研究所整理

架构创新：国产大模型底层技术突破的先行者，大幅提升模型效率

国信证券 GUOSEN SECURITIES

- **区别与多数厂商在工程创新调优，Deepseek从底层架构做出了原创设计，开源推动了国内模型实现极致性价比。**
  - MLA 多头潜在注意力：重构了传统注意力机制的 KV Cache 计算逻辑，通过低秩压缩技术，从底层把长文本推理的显存占用降至传统架构的 1/10，是其极致成本优势的核心来源。
  - DeepSeek MoE 架构：原创细粒度专家划分与共享 - 路由专家分离架构，重新设计了 MoE 大模型的参数激活与算力分配逻辑，671B 参数的大模型单 token 仅激活约 37B 参数（占比 5.5%），从架构层面实现了算力资源的动态高效分配。
  - GRPO 群体相对策略优化：推理逻辑核心支柱，取消传统 PPO 的独立 Critic 模型，通过组内相对对比更新策略，提升模型推理与代码能力。
  - MTP 多令牌预测：从训练架构上改变了传统自回归模型的单 token 预测逻辑，通过级联预测模块同时预测后续多个 token，实现了训练与推理效率的架构级提升。

表：Deepseek模型创新（2025/2前）

技术创新	模型版本	发布时间
Deepseek MoE 架构	DeepSeek-MOE	2024 年 1 月
Group Relative Policy Optimization (GRPO, 群体相对策略优化)	DeepSeek-Math	2024 年 4 月
Multi-Head Latent Attention (MLA, 多头隐式注意力)	DeepSeek-V2Paper	2024 年 6 月
Multi-Token Prediction (MTP, 多令牌预测)	DeepSeek-V3	2024 年 12 月
AI Infra 相关（以训练加速为主，如 FP8 混合精度训练、DualPipe 等）	DeepSeek-V	2024 年 12 月

数据来源：智能体AI，国信证券经济研究所整理

表：Deepseek模型创新（2025/2前）

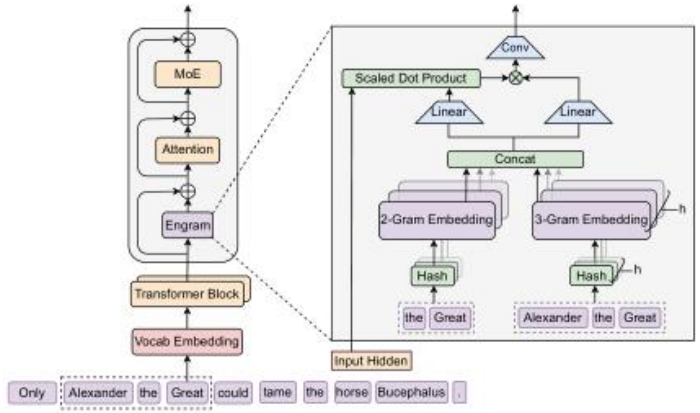
技术创新	模型版本	发布时间
AI Infra 相关（以训练加速为主，如 FP8 混合精度训练、DualPipe 等）	DeepSeek-V3：模型 / Paper	2024 年 12 月
通过强化学习显著提升模型推理能力，R1-Zero 在 AIME 2024 等推理基准测试中达到 OpenAI-o1-0912 的水平	DeepSeek-R1-Zero：模型 / Paper	2025 年 1 月
使用冷启动 - 强化学习（推理场景）-SFT - 强化学习（全场景）四阶段训练，R1 模型达到 OpenAI-o1-1217 的水平	DeepSeek-R1：模型 / Paper	2025 年 1 月
将 R1 推理能力蒸馏到小的稠密模型	DeepSeek-R1-Distill	2025 年 1 月

数据来源：智能体AI，国信证券经济研究所整理

Deepseek迭代方向：提高算力利用率，进行模型架构改进

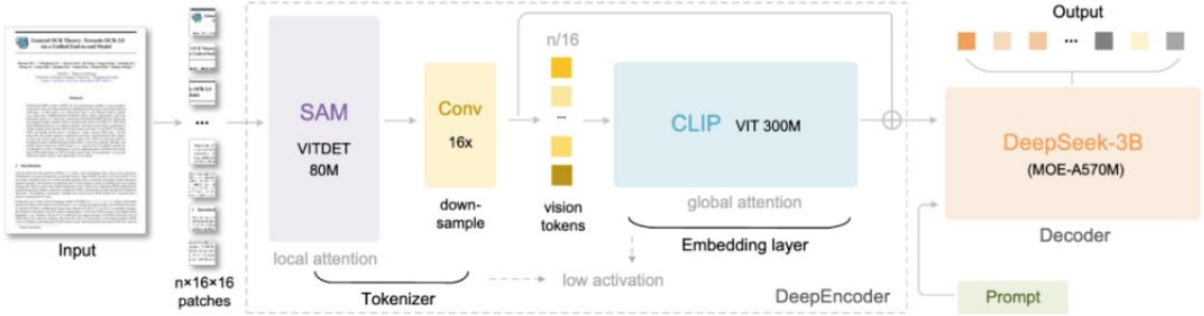
- **根据晚点，DeepSeek前期未重点投入多模态生成，因为梁文锋认为多模态生成不是智能的主线。**最新迭代方向包括：
  - **效率优化：极致压榨 GPU 算力，提高单位算力能产出的智能，并适配国产芯片。**1）DeepSeek在25年开源一整套训练与推理 Infra，涵盖推理 kernel、通信库、矩阵乘法库和数据处理框架。2）对 “注意力机制” 的持续改进：在不大幅增加算力的前提下处理更长的上下文。3）国产GPU适配：25年8月DeepSeek采用的 UE8M0 FP8，是一种数据压缩格式，针对下一代国产芯片。
  - **模型架构改进： mHC和Engram等。**1）26年初发布的 mHC（流行约束超连接），旨在提升大规模训练中的稳定性；2）Engram 条件记忆模块：CPU代替GPU成为知识容量载体。2026年1月12日，Engram提出“条件记忆”作为与MoE“条件计算”互补的新稀疏性维度，实现静态知识存储与动态计算推理的彻底分离。
  - **“非主流” 探索：**DeepSeek-OCR 帮助文档理解，助力金融、科研等核心领域。如把文本转成图片，再输入给模型，这个思路是让模型按更接近人类 “看文字” 的方式理解段落与层级，提升对复杂文档的理解力。

表：Deepseek Engram编码器结构：条件记忆给LLM装了「速查手册」



数据来源：《Conditional Memory via Scalable Lookup: A New Axis of Sparsity for Large Language Models》，国信证券经济研究所整理

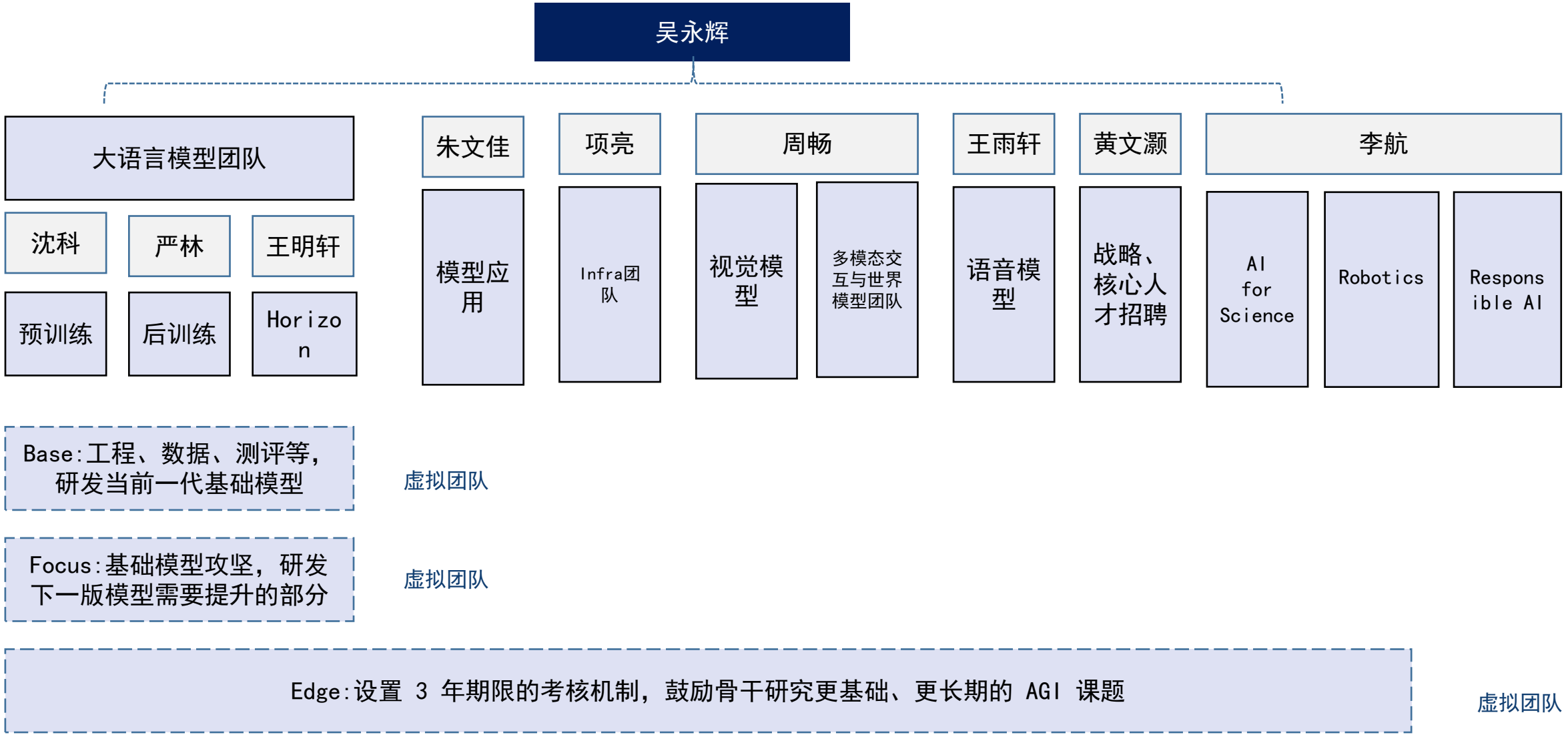
图：DeepSeek-OCR 整体架构图



数据来源：《DeepSeek-OCR: Contexts Optical Compression》，国信证券经济研究所整理

目录

字节跳动Seed组织架构



数据来源：晚点，国信证券经济研究所整理

字节Seed团队整合：技术突破期，研产分离，扁平设计强化基础研究地位

- 战略转型历程：①2023-2024 年：根据36Kr，走「突击队」路线，打快仗、抢热点，快速完成大模型从 0 到 1 的布局。

  ②2025 年：全面进入技术突破期，完成团队整合与一号位定型，核心战略从短期落地转向长期 AGI 基础研究。

  - 2025 年 1 月：启动 Seed Edge AGI 长期研究项目，设置 3 年周期考核，取消季度 OKR，鼓励长期基础研究。
  - 2025 年 2 月：前 Google DeepMind 研究副总裁吴永辉加入，任 Seed 团队 AI 基础研究负责人，直接向梁汝波汇报。
  - 2025 年 3-4 月：集团级核心研究部门 AI Lab 整体并入 Seed，完成大模型、基础 AI、多模态研发的决策权统一，结束此前多团队并行的分散格局。

图：字节Seed团队主要人物

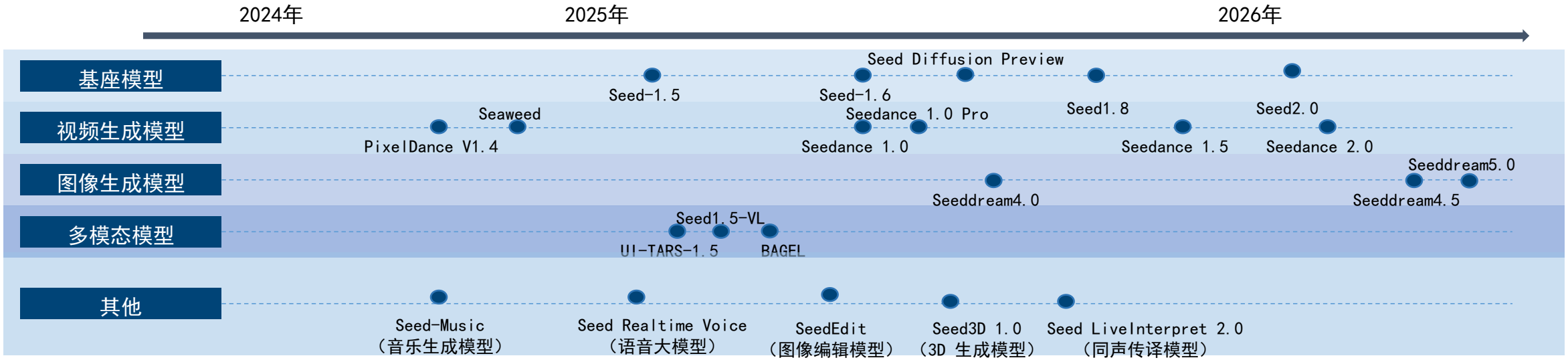
姓名	加入时间	学历背景	履历	理念
吴永辉	2025年2月	南京大学（本科）、加州大学河滨分校（博士）	前Google DeepMind研究副总裁、Google Fellow，在谷歌工作17年，参与GNMT神经机器翻译、Gemini大模型等。	理念强调长期主义。基础研究优先，探索智能上限。他明确提出团队目标是 把模型能力做到国内第一，与国际领先模型公司竞争”。
朱文佳	2015年	未公开	前百度搜索部主任架构师。2015年加入字节，历任今日头条CEO、TikTok产品技术负责人。2023年牵头组建字节大模型团队（Seed前身）。	贴着模型和用户需求做应用。作为Seed的创始负责人，其理念始终围绕 “探索与搜索、广告等下游业务的结合” 。
周畅	2024年8月	复旦大学（计算机本科）、北京大学（计算机博士）	曾担任阿里巴巴通义千问大模型的技术负责人，主导开发了2021年发布的M6多模态预训练模型。这是联合推出的中文语境下最大规模AI模型，被视为阿里大模型战略的重要里程碑。	

数据来源：晚点、36Kr、国信证券经济研究所整理

字节Seed模型：今年在编程、多模态方向发力

- 字节Seed已构建起覆盖通用语言、多模态理解和多模态生成的完整模型体系。
- **基模方向继续深耕多模态模型，复现Gemini能力。**根据央广网，字节模型以高效稀疏架构 + 深度推理为技术底座，主攻世界模型 / 具身智能、原生多模态、代码与科学 AI、Agent 化落地四大方向。
- **编程方向，有望在今年年中能力快速提升：组织新团队，着重数据准备。**数据颗粒度方面，字节正从第一档（海量代码训练+简单注释）提升至第二档（产品设计说明书、页面案例+注释+代码+bug修改记录）等。

图：字节Seed模型及agent推出时间



数据来源：公司官网，国信证券经济研究所整理

字节Seed模型优势：多模态、视觉理解有优势，工程化实现成本降低

国信证券 GUOSEN SECURITIES

- **结合视频生态发挥多模态优势。**字节Seed模型优势分两部分：一是多模态方向，字节投入最多、资源最好，生图与生视频方向有良好生态，模型能力直接受益；二是视觉理解相关功能性，在实时/离线视觉理解模型、OCR等方向投入领先且功能不错。但与谷歌性比，Seed 2.0非原生多模态，通过外挂视觉理解器。

- **字节模型性价比高快速抢占AI云份额：深度定制工程栈叠加区间定价。**1）独创的TRAE+Agent架构，豆包大模型2.0将推理成本降至行业标杆模型的1/10，实现了90%的成本降幅。根据36Kr，2026年2月14日，豆包大模型2.0针对“大规模生产环境”设计，创新点不在于某个单一算法，而在于对推理链条的整体重构。2）PD分离：工程优化。把处理长文本的预填充阶段，和生成文本的解码阶段拆开，用不同的硬件、不同的策略来处理，效率一下子就提升了好几倍。3）统一资源池与“方舟”调度系统：将内部GPU资源整合为统一弹性资源池，实现高效、市场化调度。

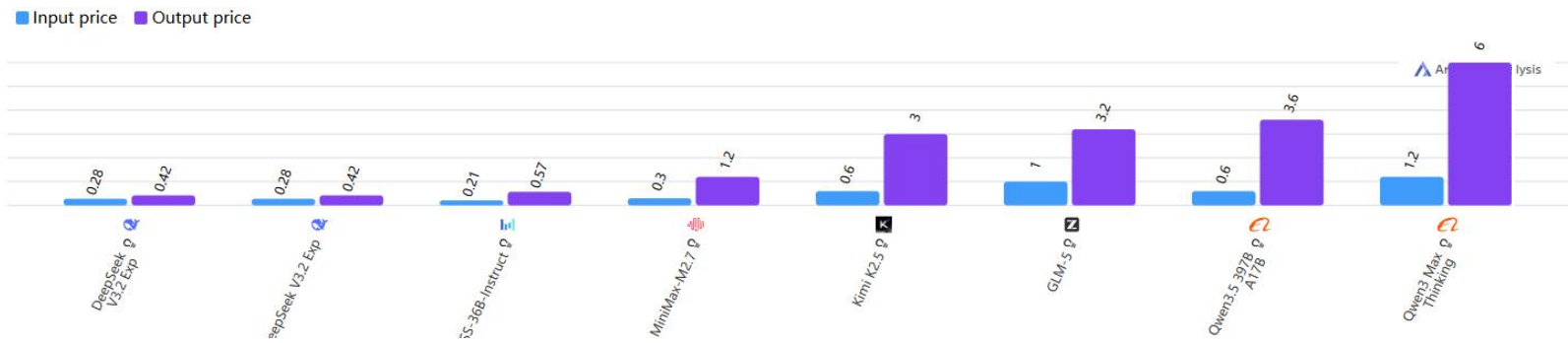
图：全球视觉模型盲测排行第五

Vision　　View →

Rank	Model	Score	Votes
1	gemini-3-pro	1290	13,906
2	gemini-3.1-pro-preview	1276	7,465
3	gpt-5.2-chat-latest-20260…	1275	4,212
4	gemini-3-flash	1274	14,159
5	dola-seed-2.0-preview	1261	4,120
6	gemini-3-flash (thinking-…	1258	11,942
7	gpt-5.2-high	1250	7,437
8	gpt-5.1-high	1248	9,824
9	gemini-2.5-pro	1247	83,351
10	kimi-k2.5-thinking	1246	7,605

数据来源：LM arena、国信证券经济研究所整理

图：国产模型输入和输出价格比较（美元/百万token）



资料来源：Artificial Analysis、国信证券经济研究所整理

字节产品矩阵：擅长C端产品，豆包凭借更接地气的产品定位突围

国信证券 GUOSEN SECURITIES

- **豆包成为国内AI个人助手用户数最多的工具。**2026年2月，豆包DAU峰值达到1.4亿。豆包成功主要来自：

1）短视频导流：字节拥有抖音这个超级流量入口。通过在抖音信息流中植入豆包广告，以极低成本触达下沉市场用户。

2）产品定位：更接地气的“全民助手”。①低门槛与拟人化： 豆包从一开始就没把自己定位成纯粹的“生产力工具”，而是更倾向于 “情绪价值 + 语音助手”。其默认的语音交互非常流畅，界面简洁，这让它在“银发族”和非技术用户中迅速破圈。2025年11月，豆包的语音对话功能，能识别20中方言。根据36Kr，为了做好体验，字节让同一批声优，用不同的方言录制一模一样的语料。后期的标注工作复杂度更是呈指数级上升。②垂直场景切入： 豆包内置了大量的“角色”和“插件”（如拍照答疑、豆包P图、周报助手），这种“开箱即用”的模式比 DeepSeek 这种硬核模型更吸引普通用户。

图：字节产品矩阵

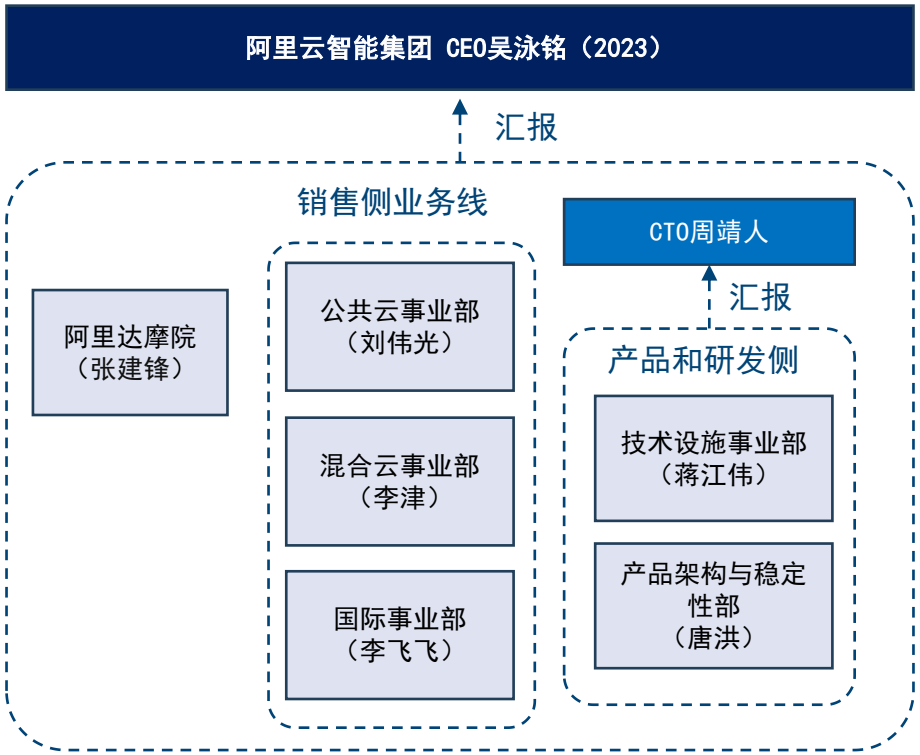
产品 / 平台	定位	面向用户
豆包 Doubao	通用 AI 助手入口	C 端大众用户
即梦 AI	AI 视频创作工具	C 端创作者、短视频用户
猫箱 Catbox	情感陪伴 / 角色互动	C 端年轻用户
TRAE	AI 编程助手	开发者、学生
豆包绘画 / Seedream	AI 图像生成	C 端普通用户、设计师
扣子 Coze	AI 智能体开发平台	开发者、企业、个人创作者
火山方舟	企业大模型服务平台	B 端企业、政企、机构
火山引擎 AI 服务	云 + AI 解决方案	B 端企业、开发者
飞书 AI	智能办公助手	B 端企业办公

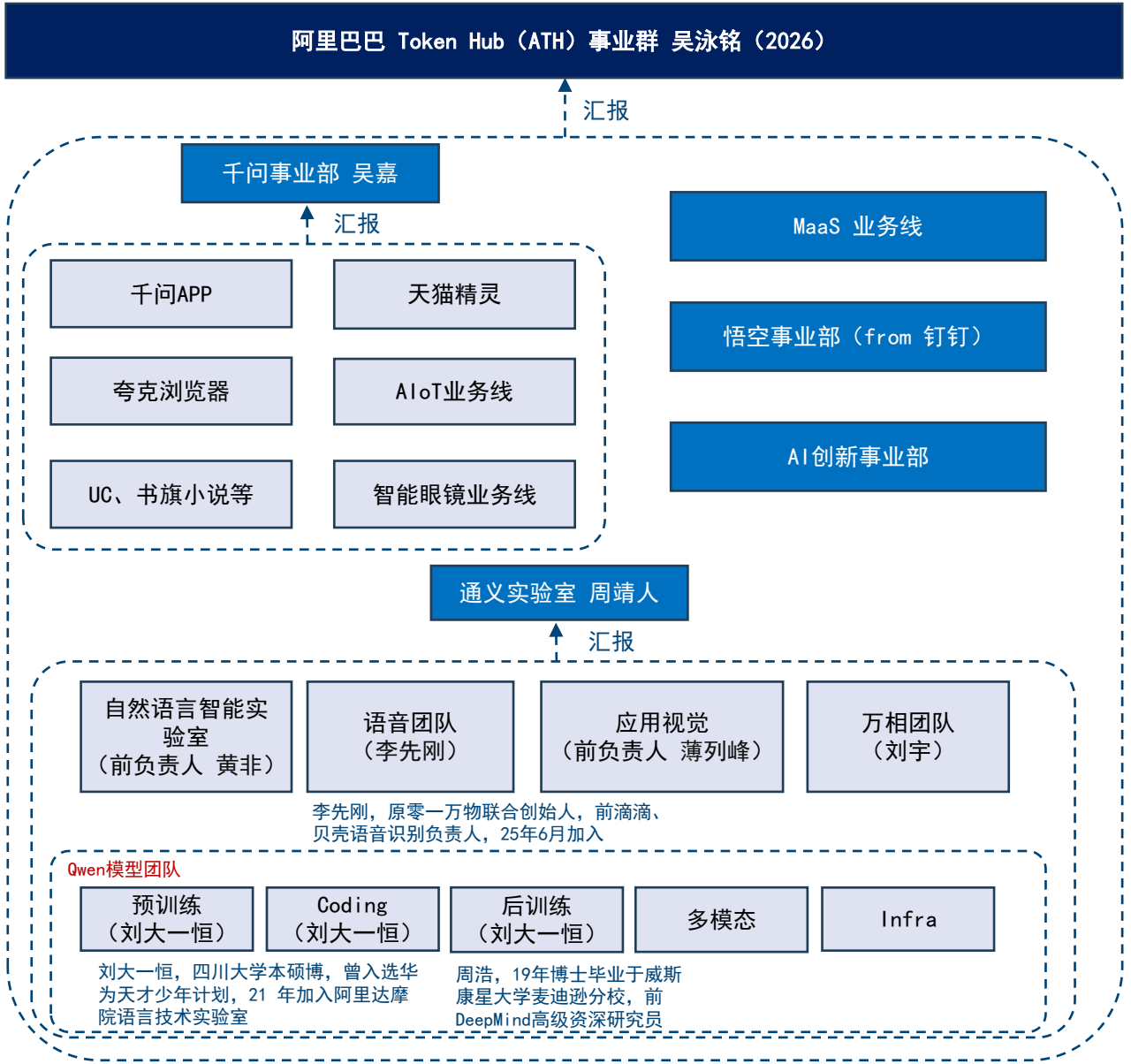
数据来源：公司官网，国信证券经济研究所整理

图：AI个人助手DAU用户数

数据来源：Questmobile，国信证券经济研究所整理

阿里云&AI组织架构





资料来源：晚点，国信证券经济研究所整理

阿里巴巴组织调整：商业落地期，以Token整合上下游

- **阿里巴巴26年3月垂直整合了与大模型相关的所有上下游业务。**根据新皮层，26年3月16日，吴泳铭通过内部信宣布了新事业群ATH的成立。ATH事业群由集团CEO吴泳铭直接负责，其核心使命是以Token为核心整合AI业务，围绕Token的创造、输送、应用构建完整的AI生产力生态。板块此前分别隶属于阿里云、千问C端事业群、钉钉、淘天等多个团队。阿里的逻辑在于，Agent时代的竞争烈度已经不允许缓慢的内部协调。
  - 上游：通义实验室（负责Qwen大模型的研发）：根据每日经济新闻，目标是提升单位Token的“智能密度”，为下游提供基础材料。
  - 中游：MaaS业务线（含阿里云百炼，负责模型服务与企业交付）；MaaS业务线则通过与云基础设施的打通，确保Token能高效触达开发者。
  - 下游：千问事业部（C端）、悟空事业部（B端）、AI创新事业部。将Token的消耗转化为实际的用户价值。

图：阿里千问团队主要人物

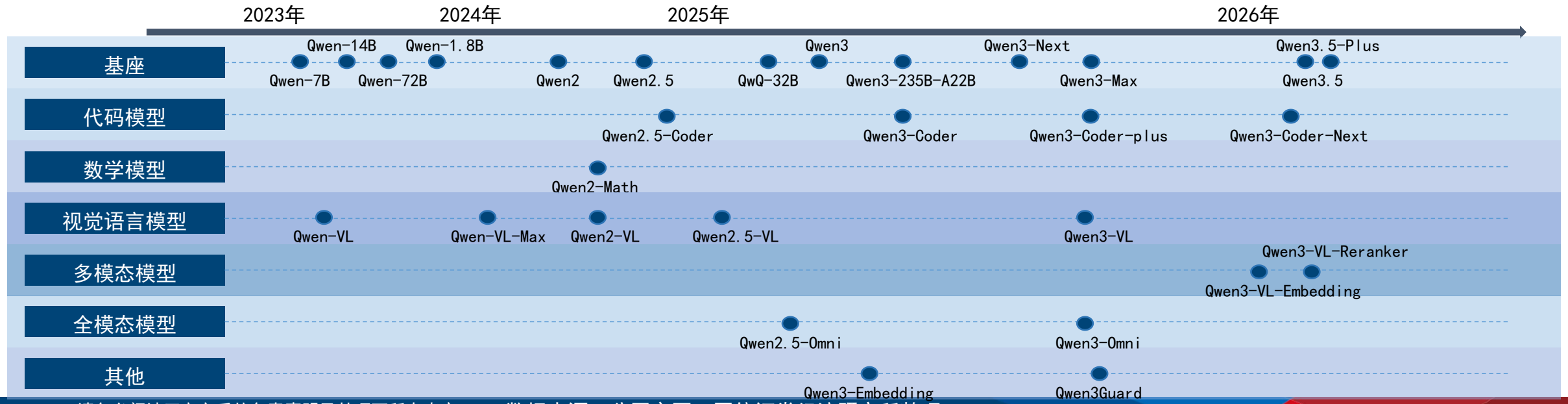
姓名	职位	加入时间	学历背景	履历	理念
周靖人	阿里云CTO、通义实验室负责人、代管Qwen模型一号位（2026年3月起）	2015年加入阿里巴巴	中国科学技术大学（本科）、哥伦比亚大学（博士）	2015年作为阿里云首席科学家加入，2022年出任阿里云CTO兼达摩院副院长，2025年晋升为阿里巴巴合伙人。自2022年底执掌通义实验室，坚定推动Qwen系列大模型的开源战略，并主导建设魔搭（ModelScope）模型开放平台。	“云+AI”深度融合，开源构建生态基础设施。公开强调“今天云的竞争，也是模型的竞争”，认为开源是技术普惠与生态繁荣的“最佳途径”，将大模型定位为云计算的新型基础设施。
林俊旸	前通义千问（Qwen）技术负责人（已离职）	2019年加入阿里达摩院	北京大学（本科、硕士）	2019年以应届生身份加入达摩院，参与M6、OFA等超大规模预训练模型开发。2022年底被任命为通义千问系列大模型的技术负责人，主导了Qwen系列模型的研发、开源与迭代，使其跻身全球顶级开源模型行列。2025年10月，他宣布组建了机器人和具身智能小型团队，探索AI从虚拟世界走向物理世界。于2026年3月4日宣布卸任离职。	

数据来源：硅谷AI见闻、36Kr、国信证券经济研究所整理

阿里巴巴模型：开源全家桶，架构创新驱动，实现原生多模态

- **架构上积极创新，原生多模态方向国内领先。**1）Qwen3 率先提出混合推理模型，将人类大脑 "快思考"（直觉反应）与 "慢思考"（深度推理）集成进同一模型，实现智能的算力分配。25年4月，通义千问模型Qwen3是国内首个“混合推理模型”。2）实现原生多模态架构（Qwen 3.5）。不同于其他厂商将视觉、语音、文本模型进行“后期缝合”，阿里 Qwen 3.5 实现了从预训练第一天起的全模态融合，探索统一多模态表征与排序。创新点：①Interleaved MRoPE（交错多模态旋转位置编码）；②DeepStack（多层视觉特征深度融合）；③Text-based Timestamp（基于文本的时间戳表征）。这些创新让Token承载了极高的空间位置信息，使得Qwen在处理复杂的 PDF、Excel 表格、手机屏幕截图时，坐标精度高。

图：阿里巴巴模型推出时间



阿里巴巴AI展望：B端MaaS增长强劲，针对Coding推出下一版本模型

- **区别于国内其他大厂，阿里巴巴拥有自研芯。**根据业绩会，平头哥自研的AI芯片已实现规模化的量产，60%以上的平头哥芯片服务于外部商业化客户。通过芯片、云基础设施和模型之间的协同优化，提升整体性价比。
  - **B 端：MaaS 业务增长强劲。**根据业绩会，截至26年3月三个月，阿里云百炼平台公共模型服务 API 市场的 Token 消耗规模，日均较12 月提升 6 倍；公司规划未来五年，包含 MaaS 在内的云和 AI 商业化收入突破 1000 亿美元。
  - **C 端：打通阿里生活服务生态。**春节期间打通淘宝、支付宝、飞猪、高德等阿里生态全业务。
- **迭代方向：26年4月发布Qwen3.6-Plus，显著增强了模型的智能体（Agent）编程能力。**围绕推理能力增强、指令模式实用性提升以及复杂任务执行能力拓展三个方向持续演进。在前端网页开发，复杂的代码仓库级问题求解表现优秀。依靠多模态优势，Qwen3.6-Plus 对世界的感知更加精准。

表：阿里巴巴C端产品布局

产品名称	定位	核心能力
千问 App	阿里 C 端 AI 主入口	个人 AI 助手、AI 购物、智能体、多模态交互
淘宝问问	电商 AI 导购	商品推荐、问答、文案、直播辅助、购物决策
夸克 AI	AI 搜索 + 超级助理	搜索、文档、学习、写作、简历、翻译、扫描
支付宝 AI	支付 / 理财 AI 助手	智能客服、理财分析、账单解读、便民服务
高德 AI	出行 AI 助手	智能导航、行程规划、路线推荐、语音交互
通义听悟	音视频 AI 工具	录音转文字、会议纪要、字幕生成、多语种转写
通义万相	AI 绘画 / 创作	文生图、海报、头像、设计生成

数据来源：公司官网，国信证券经济研究所整理

表：阿里巴巴B端产品布局

产品名称	定位	核心能力
悟空 Work	企业级工作平台	AI 工作流、智能体、任务执行、企业中枢
钉钉 AI	办公协同 AI	会议纪要、总结、代办、知识库、智能审批
阿里云百炼	大模型 MaaS 平台	API 调用、微调、部署、Agent 开发、企业私有化
1688 AI	批发电商商家 AI	选品、营销、客户管理、生意分析、企业查询
通义灵码	AI 编程助手	代码生成、补全、解释、调试、研发问答
通义点金	金融行业 AI	市场分析、财报解读、风险预警、投研辅助
通义仁心	医疗行业 AI	病历、问诊、医学文献、辅助诊断相关

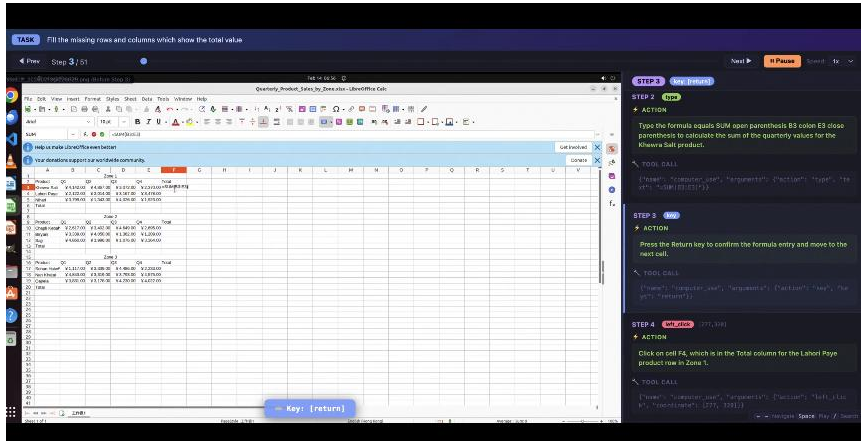
数据来源：公司官网，国信证券经济研究所整理

多模态+Agent提升了空间推理、自主操控的任务边界

国信证券 GUOSEN SECURITIES

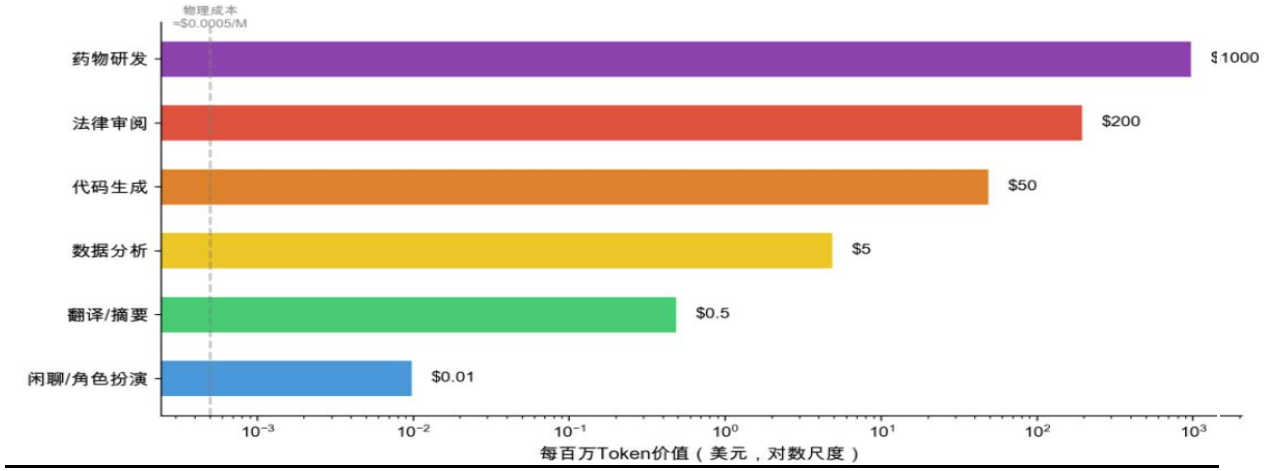
- **多模态的核心价值在于让 AI 拥有了“眼睛”，不仅能看懂画面，还能进行空间推理和自主操作。**Agent的任务边界被大幅拓宽，例如可以浏览网页和文档，生成图文并茂报告、PPT，还可以查询并解读K线图等复杂图表。以下是Qwen3.5案例。
  - **代码智能体：前端开发全链路提效。**1）核心优势：网页前端开发能力突出，可将自然语言指令直接转化为可运行代码，高效完成网站搭建、UI 设计等前端任务；2）拓展能力：支持项目实时迭代开发，可完成代码生成、创意视频生成等复合任务
  - **视觉智能体:操作手机/电脑的GUI自动化，空间推理，助力教育科研多模态等。**1）自动化操作：可作为 GUI 智能体，自主操作手机 / 电脑完成日常办公任务；2）视觉编程能力：实现手绘界面草图转前端代码、长视频内容提炼为结构化网页 / 可视化图表，大幅降低创意落地门槛；3）空间智能与推理：像素级图像建模，精准完成物体计数、空间关系判断、多步学科解题，为教育、科研等场景提供可靠的多模态 Agent 支撑等。

图：Qwen3.5GUI智能体自主操作手机与电脑完成日常任务



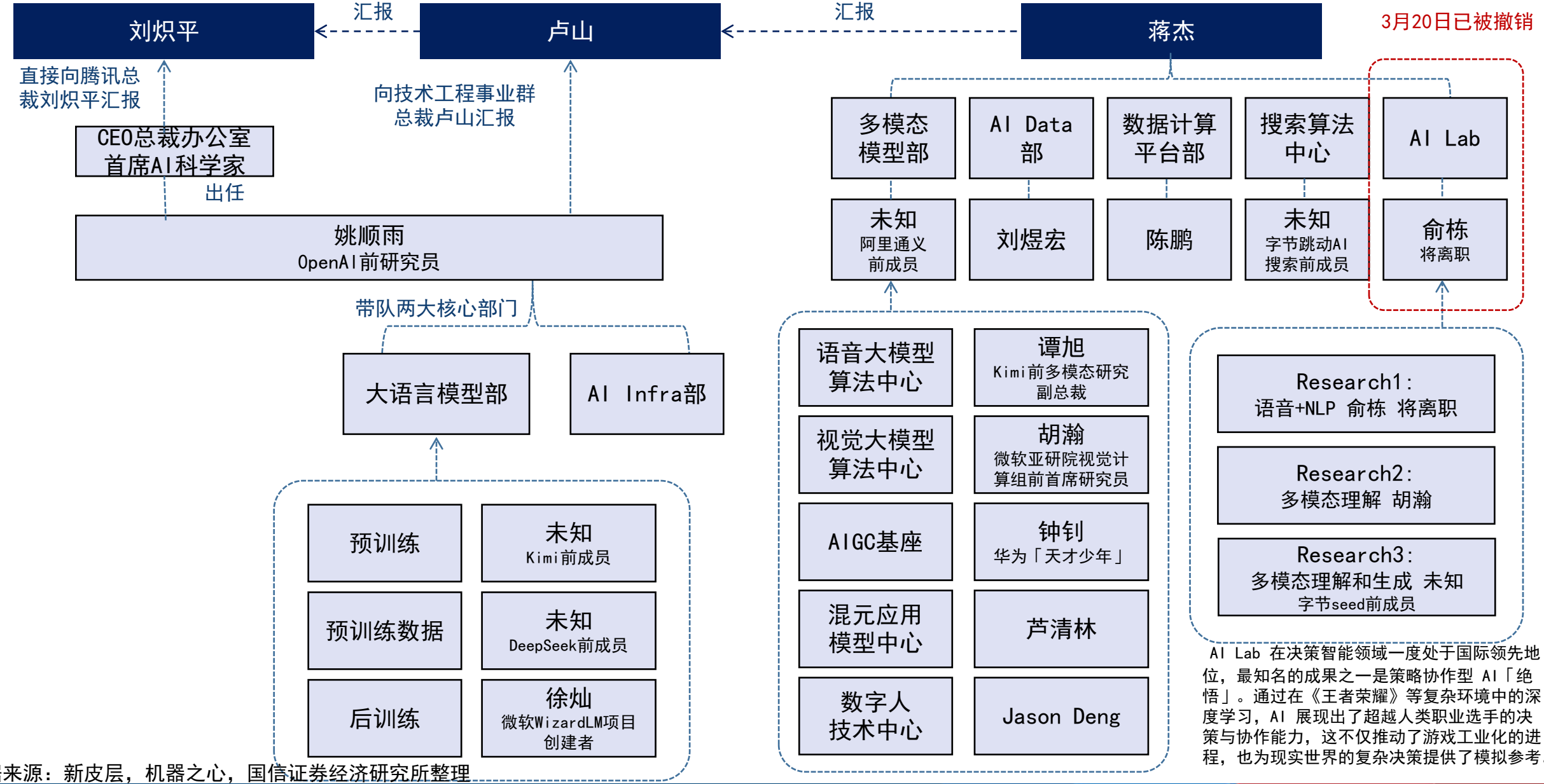
资料来源：公司官网、国信证券经济研究所整理

图：Token价值光谱—同一物理单元，经济价值差十万倍



资料来源：Litowitz et al. (2026)；Bergemann et al. (2025)，腾讯研究院，国信证券经济研究所整理

腾讯混元组织架构



AI Lab 在决策智能领域一度处于国际领先地位，最知名的成果之一是策略协作型 AI「绝悟」。通过在《王者荣耀》等复杂环境中的深度学习，AI 展现出了超越人类职业选手的决策与协作能力，这不仅推动了游戏工业化的进程，也为现实世界的复杂决策提供了模拟参考。

数据来源：新皮层，机器之心，国信证券经济研究所整理

腾讯混元组织调整：整合资源，“业务驱动”向“AI原生驱动”

国信证券 GUOSEN SECURITIES

- **将原本分散在各业务线的AI研发力量统一整合至混元团队，成立三大基础设施部门。**根据新皮层，2023年成立的混元是一个虚拟团队，其成员来自AI Lab、TEG的多个部门。2025年4月，腾讯从组织上配备了一个类似字节Seed那样的独立团队。

- **姚顺雨的双重汇报角色（既向总裁汇报战略，又向事业群总裁汇报执行），标志着AI业务被抬升至公司级战略高度。**25年12月官宣姚顺雨入职后，腾讯进一步调整组织架构，新成立了AI Infra部、AI Data部与数据计算平台部。通过构建统一的AI基建中台，大幅提升模型训练效率，降低内部业务使用成本。除此以外，腾讯加速人才扩张。根据新皮层，2025年的腾讯就是2024年的字节跳动。2024年，字节跳动加速基础模型的研发进度，第一个动作也是抢人才。以姚顺雨的加入为节点，腾讯招募了更多大语言模型（LLM）相关的人才，在此之前，新加入者主要研究方向都以多模态为主。

表：腾讯混元核心人员

姓名	职位	加入时间	学历背景	履历
姚顺雨	腾讯首席AI科学家、AI Infra部与大语言模型部负责人	2025年12月	清华大学（本科）、普林斯顿大学（计算机博士）	姚顺雨经验非常适合大模型下半场—Agent开发以及后训练。姚顺雨曾在OpenAI工作过一年，深度参与Operator、Deep Research等智能体项目。在普林斯顿博士期间，研究方向是自然语言处理与强化学习，研究成果包含①ToT（Tree of Thoughts，思维树）②ReAct（大模型可以边推理边行动）等开创性框架，是语言智能体方向的代表性学者。ToT和ReAct都可应用于模型的后训练阶段，提升模型的多步推理和动手操作能力。
薄列峰	腾讯混元大模型团队多模态方向负责人	2025年7月	西安电子科技大学（电气工程博士）	2025年加入腾讯。多模态专家，前阿里通义实验室应用视觉团队负责人，主导研发Animate Anyone、EMO等标杆视频生成技术。
卢山	腾讯集团高级执行副总裁、技术工程事业群（TEG）总裁	2000年	中国科学技术大学（计算机科学与技术本科）	腾讯资深高管，2000年加入。作为TEG总裁，是腾讯大模型研发体系的最高技术负责人之一，为混元提供底层算力与工程化支持。

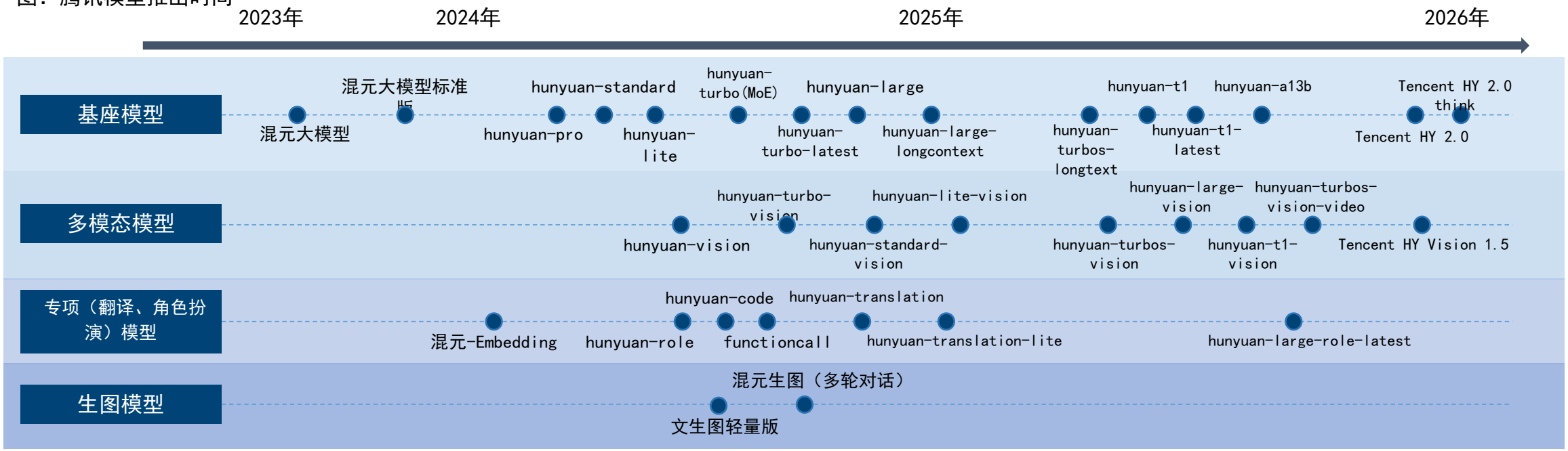
数据来源：36Kr、新皮层、国信证券经济研究所整理

腾讯模型：全栈全模态模型体系，世界模型等多模态方向拥有优势

国信证券 GUOSEN SECURITIES

- **受益于专有的数据及丰富的使用场景，混元模型在3D 生成、文生图和世界建模等多模态能力方面拥有优势。**腾讯庞大的产品矩阵（如社交、内容、游戏）每天产生海量的图文、音视频、3D内容交互数据。根据腾讯研究院，今年春节活动期间，混元图像3.0图生图模型，带动元宝AI生图日均调用量，增长了30倍；混元3D模型则继续保持行业领先，服务了拓竹科技、创想三维等3D打印企业，并开始向海外市场覆盖。26年3月，腾讯开源业界首个面向世界模型的强化学习后训练框架 WorldCompass。

图：腾讯模型推出时间



数据来源：公司官网，国信证券经济研究所整理

腾讯模型迭代：25H2提升数据质量，混元3.0聚焦Agent能力和参数优化

国信证券 GUOSEN SECURITIES

- **混元3.0模型关注MOE参数分配以及Agent能力。**25H2，腾讯对混元团队与研发流程进行了重构，聚焦提升数据质量，重建预训练与强化学习基础设施。根据业绩会，混元3.0 正在内部业务测试中，将于4月对外推出。根据腾讯研究院，即将发布的混元3.0，一方面激活参数大幅降低，体验更优，另一方面在复杂推理、长记忆、长文、多轮追问与Agent能力等多个维度，有明显的提升。
- 多模态也是混元与元宝的重点。同时，腾讯也非常关注适配端侧部署的小模型机会。比如混元的7B翻译模型，在2025国际机器翻译大赛31个单项中，斩获了30个第一名。
- **腾讯AI优势优势不在于单一的聊天入口，而在于将 AI 能力“隐形”化。**根据业绩会，智能体本身也具备跨设备、跨领域的特性。小程序这类去中心化体验也将迎来升级，具备比以往更强大的功能。
- **拥有独家全面的生态壁垒。**深度绑定微信 / QQ / 视频号 / 腾讯会议 / 游戏 / 音乐，国内独一无二的 C 端 + B 端全场景生态。

图：全球图像编辑模型盲测排行

Image Edit　　14 days ago

Rank	Model	Score	Votes
1	chatgpt-image-latest-high…	1402	243,541
2	gemini-3-pro-image-previe…	1392	229,951
3	gemini-3-pro-image-previe…	1391	521,159
4	gemini-3.1-flash-image-pr…	1388	43,471
5	gpt-image-1.5-high-fideli…	1381	262,006
6	grok-imagine-image	1339	10,161
7	grok-imagine-image-pro (2…	1319	136,785
8	grok-imagine-image (20260…	1315	141,512
9	hunyuan-image-3.0-instruct	1312	109,856
10	seedream-4.5	1310	443,277

数据来源：LM arena、国信证券经济研究所整理

腾讯 OpenClaw布局：布局积极，打通生态

- **腾讯在OpenClaw方面响应速度、产品矩阵、生态整合方面国内领先。**腾讯3月9日起密集发布产品，快速构建“自研龙虾、本地虾、云端虾、企业虾、云桌面虾” 完整产品矩阵，覆盖所有主流用户群体。除此以外，腾讯率先将"龙虾"与微信、QQ等十亿级用户平台打通。3月31日，腾讯workbuddy推出了小程序版，文档处理可以连接微信文件。4月2日，腾讯QQ 正式原生接入 OpenClaw 平台，成为国内首个被OpenClaw官方原生接入的社交平台。
- **腾讯还上线了专为国内用户优化的技能社区（SkillHub）。**提供国内镜像加速，告别插件下载卡顿的痛点。

表：腾讯龙虾矩阵

数据来源：腾讯科技，公司官网、国信证券经济研究所整理

图：腾讯AI产品汇总

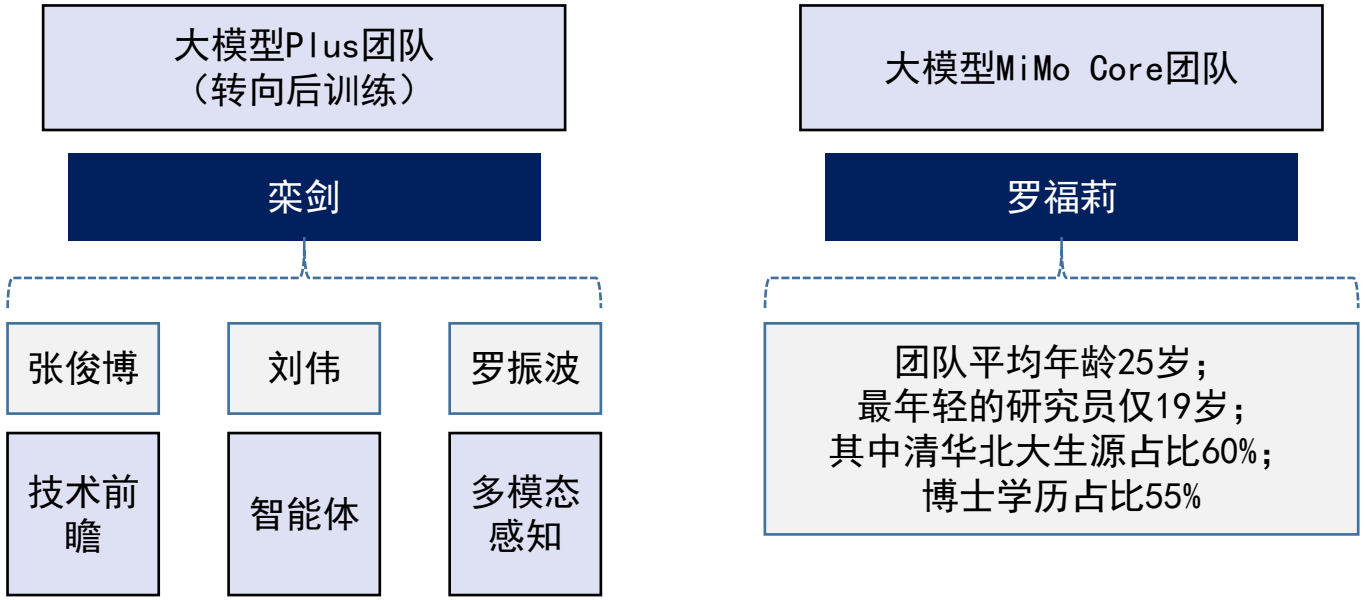
产品层次	核心产品/项目	定位与目标用户
个人应用层（C端）	腾讯元宝	独立的AI助手App，战略级产品。2026年春节后MAU突破1.14亿，已打通QQ音乐、腾讯视频、腾讯会议等数十个腾讯系应用，是腾讯AI的统一C端入口。
	微信AI智能体	嵌入微信的“绝密级”项目，被视为腾讯未来的“核武器”。目标是连接微信内数百万小程序，让AI像好友一样帮你完成叫车、点餐等任务，预计2026年第三季度推出。
	ima AI工作台	以知识库为核心的AI生产力工具，提供精准检索、高质量问答和文本创作辅助。月活超1300万，知识库文件超4.2亿。
	QQ浏览器AI	提供全场景“AI+”服务，如长文档总结、智能搜索等，累计服务用户超1.3亿。
	搜狗输入法AI	混元驱动的AI语音识别率达98%，移动端月活超6.7亿，稳居AI输入法行业第一。
企业应用层（B端）	腾讯云智能体开发平台（ADP）	面向中大型企业的一站式智能体开发平台，可秒级接入企业微信。已在金融、传媒、零售、医疗等20多个行业落地，传媒行业客户规模同比增长13倍。
	WorkBuddy	面向大众用户和中小团队的全场景桌面智能体，开箱即用，支持微信一键直连，内置超20种技能包。
	QClaw（本地虾）	面向个人开发者/极客的本地AI助手，可在电脑上一键部署，通过自然语言驱动复杂任务，主打安全隔离。
	腾讯云Lighthouse（云端虾）	面向开发者的云端智能体部署方案，支持7×24小时稳定在线，一个QQ号可部署多个智能体。
	腾讯云桌面（云桌面虾）	面向大型企业的分布式办公解决方案，支持多地多点远程接入。
	腾讯会议AI	提供“AI纪要”、“AI托管”等功能，AI用户量同比增长超150%。
	腾讯文档AI	企业版AI助手，支持智能续写、内容校对等。
	CodeBuddy	腾讯云代码助手，覆盖腾讯超90%工程师，整体编码时间缩短40%。
安全防护层	安全隔离虾房、云保安	为AI应用提供系统级安全保障。电脑管家可为AI开辟“隔离房”，云端有AI Agent安全中心实时拦截高风险指令。

数据来源：腾讯科技，公司官网、国信证券经济研究所整理

小米大模型组织架构

2023年4月，小米组建了大模型团队，由栾剑带队，战略重心是轻量化与本地部署。
24年底，罗福莉加入小米后，组建了大模型Core团队。小米大模型原负责人栾剑开始负责大模型Plus团队，与罗福莉的Core并列，但不再从事基座模型预训练方面的工作，而是转向后训练。

图：小米大模型团队组织架构



数据来源：新皮层、国信证券经济研究所整理

表：小米大模型团队核心成员履历

姓名	职位	加入时间	学历背景	履历
栾剑	AI实验室大模型团队负责人	2023年4月	未公开	业内知名AI领域专家，曾任微软小冰首席语音科学家，小米AI实验室NLP应用负责人。20多年来一直从事语音、语言方向的技术研究和应用，在声纹识别、语音合成、语音识别、歌唱合成、人机对话、机器翻译、大语言模型、多模态大模型等多个领域都有深入研究，发表论文100余篇。主导研发了业内首个AI歌唱合成软件X-Studio，小米超级拟人语音合成系统、小爱翻译离线字幕、小米自研大模型MiLM、MiMM等多个项目。
罗福莉	Xiaomi MiMo大模型负责人	2024年底	北京师范大学计算机学士；北京大学计算语言学硕士	2019年通过阿里星计划加入达摩院，主导多语言模型VECO（日均调用50亿次），推动AliceMind开源。硕士期间发表顶会论文超20篇，2019年ACL发文8篇（2篇一作）。2022年加入DeepSeek任研究员，为DeepSeek-V2关键开发者。

数据来源：新皮层、国信证券经济研究所整理

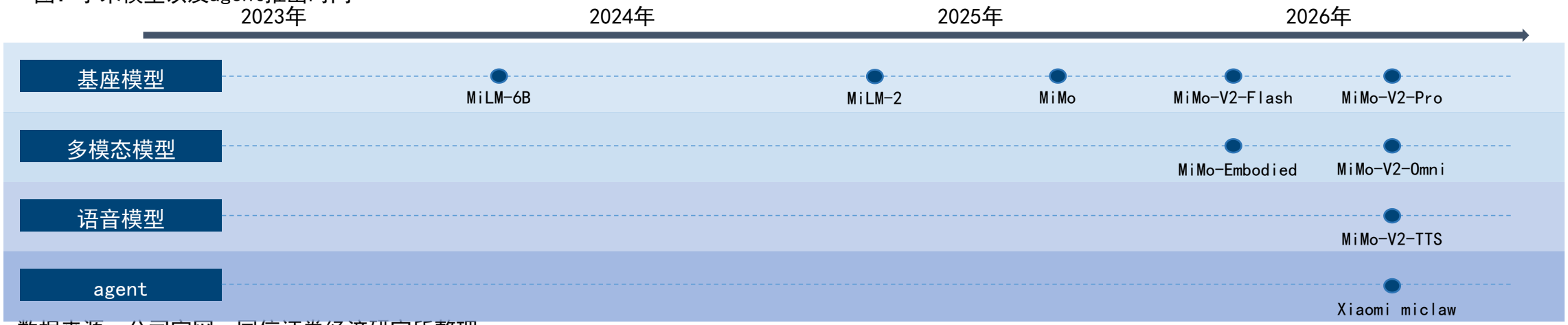
小米模型：26年发布万亿参数智能体模型

国信证券 GUOSEN SECURITIES

- 发展历程：25年4月，推出首个用于推理任务的大模型 MiMo-7B，模型系列名称正式从 MiLM 切换为 MiMo。2026年3月，旗舰模型MiMo-V2-Pro正式发布，总参数超1万亿，采用MoE架构并支持100万tokens超长上下文。该模型此前曾以匿名身份“Hunter Alpha”上线OpenRouter平台，在开发者不知情的情况下迅速登顶调用量榜首。

- **MiMo-V2-Pro模型优势在解决模型能力不均衡、强化真实场景应用、提升训练效率等。**根据钛媒体，MiMo-V2-Pro模型区别于同行的，是后训练阶段的三项技术。1）MOPD 多教师在线策略蒸馏，针对大模型后训练普遍存在的 “跷跷板效应”，包括 SFT、训练各领域极致专家教师、学生模型接受多教师 token 级实时监督的三阶段范式；2）真实环境的 agentic RL 训练，区别于行业内单轮、封闭的强化学习模式，构建了覆盖四大类场景、超 12 万个真实交互环境的训练体系；3）与北大联合研发的 ARL-Tangram 训练基础设施系统，针对传统 RL 框架静态资源预留导致的算力严重闲置问题。

图：小米模型以及agent推出时间



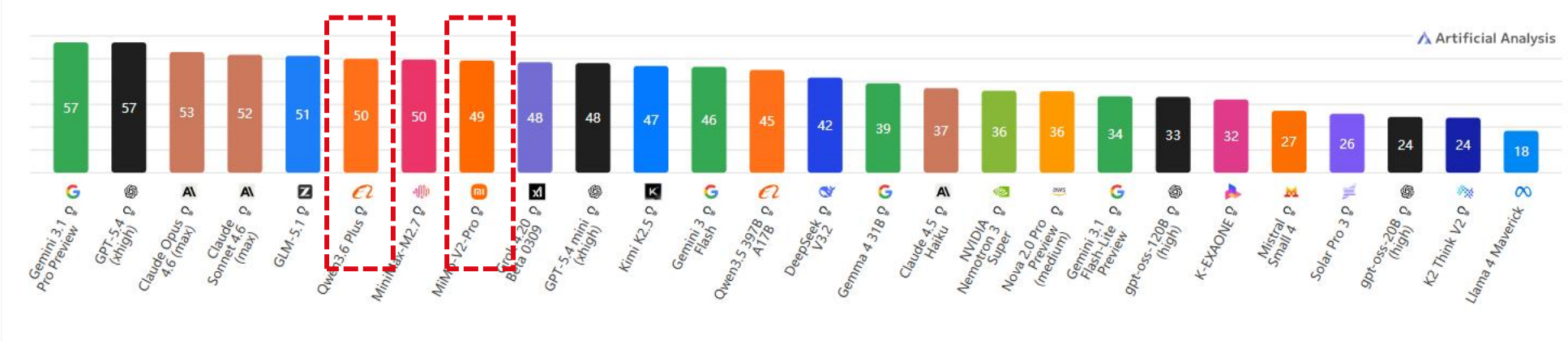
数据来源：公司官网，国信证券经济研究所整理

小米模型迭代：目标是更好服务客户和生态，模型将与终端产品融合

- **MiMo模型将在上下文拓展、原生多模态、底层硬件方向发力。**小米目标是服务好用户，在模型选择上开放。根据新皮层，小米大模型Plus团队的后训练不止基于小米自研模型，也可能直接拿Qwen等模型做后训练，再服务于小爱同学等。

- **小米模型将与终端产品融合。**2026年3月发布会上，雷军正式宣布小米首款AI原生手机“龙虾”Xiaomi miclaw已启动封测。这款产品通过搭载MiMo大模型，实现了与操作系统及人车家全生态的深度融合，用户无需复杂的指令说明，AI即可自主寻求最优方案并安全执行。除此以外，小米大模型能力将在在汽车（SU7的XLA认知大模型）和智能家居（Xiaomi Miloco方案）中全面落地。根据上海证券报，2026年小米有希望在一款终端产品上实现自研芯片、OS、AI大模型的“大会师”。雷军已宣布未来三年在AI领域投入超600亿元（2026年单年超160亿元）

图：通用模型智能水平排序（26/4/8）



资料来源：Artificial Analysis、国信证券经济研究所整理

风险提示

第一，宏观经济波动。若宏观经济波动，公司业务、产业变革及新技术的落地节奏或将受到影响。

第二，下游需求不及预期。若下游AI需求不及预期，相关的AI研发投入增长或慢于预期，致使行业增长不及预期。

第三，核心技术水平升级不及预期的风险。AI大模型研发进度落后，AIGC相关产业技术壁垒较高，核心技术难以突破，影响整体进度。

第四，AI快速迭代、平权化下竞争加剧，影响云业务利润率。

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。

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