仅供内部参考，请勿外传！

渤海证券 Bohai Securities

智能体拉动 Token 需求，大模型商业化持续验证

——AI 大模型行业专题

专题研究小组成员： 连天龙

2026 年 10 月 13 日

专题研究小组
证券分析师
连天龙
SAC NO：
S1150525070002
liantl@bhzq.com
022-23861563

投资要点：

- **大模型迭代持续深化，Agent 能力助推商业化放量**

当前大模型在传统基准已接近人类水平，强 Agent 实操任务成为当前迭代重点。2025-2026 年间前沿 AI 系统在 OSWorld 智能体计算机操作这类强 Agent 属性的实操任务方面迭代提升速度显著加快。2025 年国内模型即服务市场 Token 调用规模已达到 1941 万亿 Token。展望 2026-2030 年，Agent 智能体应用的持续渗透将成为驱动 Token 消耗规模扩张的核心动力。

- **海内外模型厂商探索多元变现，商业化模式持续验证**

当前国内大模型厂商已演化出 MaaS 公有云 API 服务、私有化/本地化部署、行业定制解决方案、会员订阅服务四类主流变现路径。海外闭源大模型 Anthropic 大部分营收由 B 端企业 API 调用业务贡献。长上下文处理与复杂 Agent 任务带来更高的单任务 Token 消耗，持续推动 Anthropic 营收规模快速上行。国内开源大模型智谱主营业务收入结构在 2026 年上半年完成显著置换，公有云 MaaS（以 API 接口调用、Token 按量计费）业务逐步成长为核心增长曲线。

- **RSI 驱动递归自我改进，大模型迈向自进化体系**

RSI 是一类具备自我进化潜力的 AI 系统。AI 在 Agent 执行真实业务的场景里，自主利用任务成败、交互结果中提取可供学习的信号完成迭代进化，以此降低模型迭代对人工标注资源的依赖。海内外头部大模型厂商均已经围绕 RSI 开展对应的工程实践工作。大模型目前已经可以借助 Agent 执行任务产生的各类反馈辅助模型迭代优化。

- **大模型持续拓展应用边界，物理 AI 与 AI4S 迎来发展机遇**

物理 AI 是人工智能与现实物理系统深度融合的前沿技术赛道，具备实体经济落地价值与广阔成长空间。英伟达推出全球首款全开源全模态物理 AI 大模型 Cosmos 3，有望推动物理 AI 技术发展提速。AI for Science 正加速落地生命科学领域，AI 大模型正在重塑药物研发全流程。跨国药企持续加码 AI 药物研发，当前落地价值集中在科研流程自动化。

- **风险提示**

宏观经济风险；行业竞争加剧的风险；供应链波动加大的风险；AI 技术迭代速度不及预期的风险；AI 应用落地效果不及预期的风险。

仅供内部参考，请勿外传！

目 录

1. 大模型迭代持续深化，Agent 能力助推商业化放量 4
1.1 Agent 重塑模型能力，大模型厂商兼顾性能与成本 4
1.2 Agent 拉动 Token 需求增长，国产模型全球化落地提速 6
1.3 Token 经济持续发展，MaaS 行业竞争转向综合能力 7
2. 海内外模型厂商探索多元变现，商业化模式持续验证 9
2.1 国内大模型商业化模式分化，四类变现路径并行发展 9
2.2 AI 释放降本增收潜力，各职能收益体现不同 10
2.3 算力布局路径不同，厂商资本开支呈现分化 11
2.4 海外闭源大模型：B 端企业 API 驱动 ARR 高速增长 12
2.5 国内开源大模型：从私有化项目向公有云 MaaS 快速转型 13
3. RSI 驱动递归自我改进，大模型迈向自进化体系 14
3.1 RSI 革新模型迭代范式，依托场景反馈实现自优化 14
3.2 RSI 能力持续演进，海内外厂商积极布局 15
4. 大模型持续拓展应用边界，物理 AI 与 AI4S 迎来发展机遇 16
4.1 物理 AI 衔接虚拟与现实，软硬件协同赋能具身智能 16
4.2 AI4S 重塑药物研发链条，海内外药企加速布局 17
5. 相关公司基本情况介绍 18
5.1 智谱 18
5.2 MINIMAX 19
6. 风险提示 20

仅供内部参考，请勿外传！

图 目 录

图 1： 人工智能指数报告选取的部分技术性能比较基准与人类表现对比 4
图 2： Artificial Analysis 全球大模型智能综合指数排行榜（截至 2026 年 9 月 27 日） 5
图 3： Artificial Analysis 全球大模型单位任务平均输出 Token 量排行榜（单位：Token，截至 2026 年 9 月 27 日） 5
图 4： Artificial Analysis 全球大模型单位智能任务加权完成成本排行榜（单位：美元，截至 2026 年 9 月 27 日） 6
图 5： OpenRouter 周度大模型调用量情况（单位：万亿 Tokens） 7
图 6： OpenRouter 周度模型调用量排行榜（2026 年 9 月 1 日-9 月 27 日） 7
图 7： 2024-2030 中国模型及服务市场规模及预测情况（单位：万亿 Tokens） 8
图 8： 2025 年中国 MaaS 市场份额情况（按 tokens 份额） 8
图 9： 2025 年各职能人工智能使用的成本降低与营收增长情况 10
图 10： 2022-2025 年 Anthropic 与 OpenAI 年度算力支出情况（单位：十亿美元） 11
图 11： Anthropic 年化经常性收入情况 12
图 12： 智谱公有云 MaaS（以 API 接口调用、Token 按量计费）收入情况 13
图 13： 模型训练阶段自迭代示意图 14
图 14： RSI 能力 L0-L5 逐级演进框架图 15
图 15： 英伟达推出面向机器人、自动驾驶等使用场景的物理 AI 大模型 Cosmos 3 17
图 16： AI for Science 被视作继实验、理论、计算、数据科学之后的科研第五范式 17
图 17： 智谱营收情况 19
图 18： 智谱归母净利润情况 19
图 19： MINIMAX 营收情况 20
图 20： MINIMAX 归母净利润情况 20

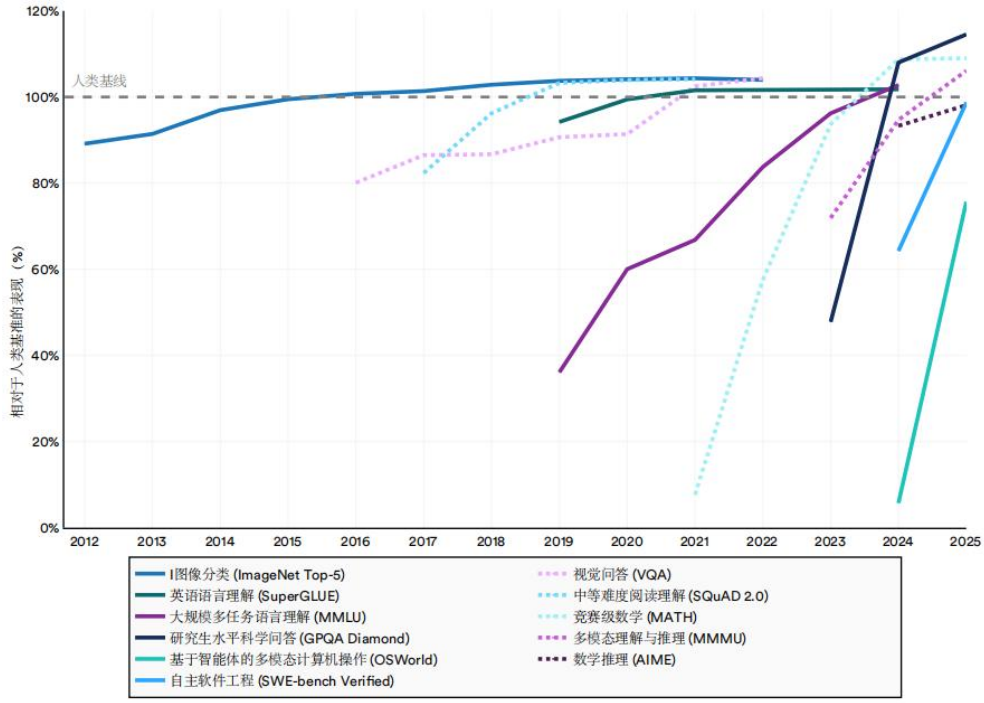
仅供内部参考，请勿外传！

1.大模型迭代持续深化，Agent 能力助推商业化放量

1.1Agent 重塑模型能力，大模型厂商兼顾性能与成本

当前大模型在传统基准已接近人类水平，强智能体（Agent，以下简称“Agent”）实操任务成为当前迭代重点。进入 2026 年，长上下文能力、多模态融合效果、推理稳定性、工具调用工程化水平以及真实业务场景下的任务完成率，共同构成模型迭代的核心主线。行业的模型评价体系也逐步向更加务实的工程落地指标迁移，可落地、可复用、能够自动化执行复杂任务的实用能力成为市场选型的优先考量。根据斯坦福大学人工智能研究院《人工智能指数报告 2026》的评测统计，2025-2026 年间前沿 AI 系统在自主软件工程（SWE-bench Verified，以下简称“SWE-bench Verified”）、基于智能体的多模态计算机操作（OSWorld）这类强 Agent 属性的实操任务上，模型整体表现仍低于人类基线，但迭代提升速度急剧加快，SWE-bench Verified 评测得分就从 2024 年约 60%提升至 2025 年接近 100%，侧面印证行业重心正在向真实复杂任务能力倾斜。

图 1：人工智能指数报告选取的部分技术性能比较基准与人类表现对比



资料来源：《人工智能指数报告 2026》，渤海证券研究所

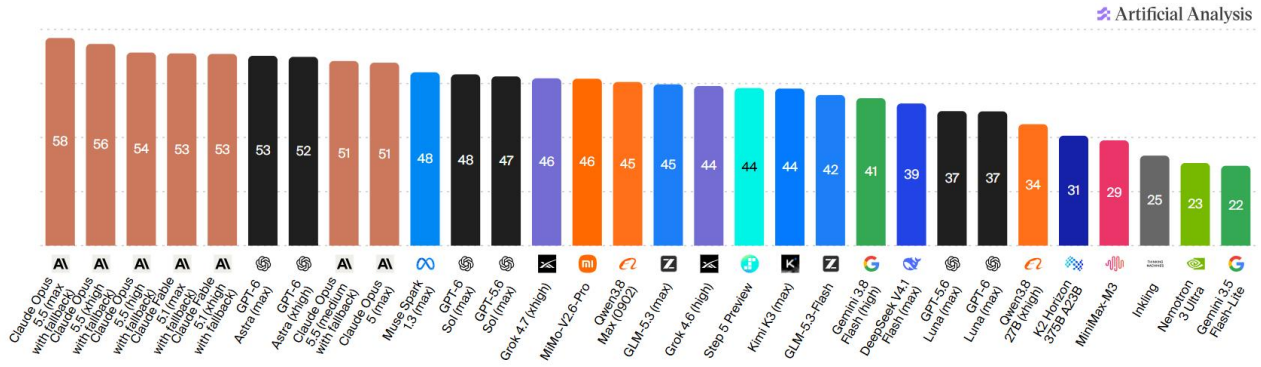
2026 年海内外头部厂商集中完成旗舰模型迭代升级，产品形态转向具备自主任务规划、多工具链式调用、多轮自我校验纠错能力的 Agent 产品体系。Agent 技术沿着清晰路径迭代演进，从早期仅支持单轮简单工具调用，发展至多步骤任务自主规划；当前头部模型已经具备循环纠错、异常处理、完整自主工作流执行能力。海

仅供内部参考，请勿外传！

外方面，Anthropic 推出 Claude Fable 5.1 与 Mythos 5.1 大模型，智能体能力得到系统性增强，适配高复杂度、长链路、多步骤持续性任务；OpenAI 发布 GPT-6 Astra 大模型，在计算机操作、网页智能浏览、软件工程开发等六大专业领域实现能力跃升。国内方面，智谱推出 GLM 5.3 大模型，通过后训练扩容显著提升长周期 Agent 任务稳定性；月之暗面发布 Kimi K3 大模型，面向长代码解析、全链路文档处理等高吞吐 Agent 作业场景完成能力适配。

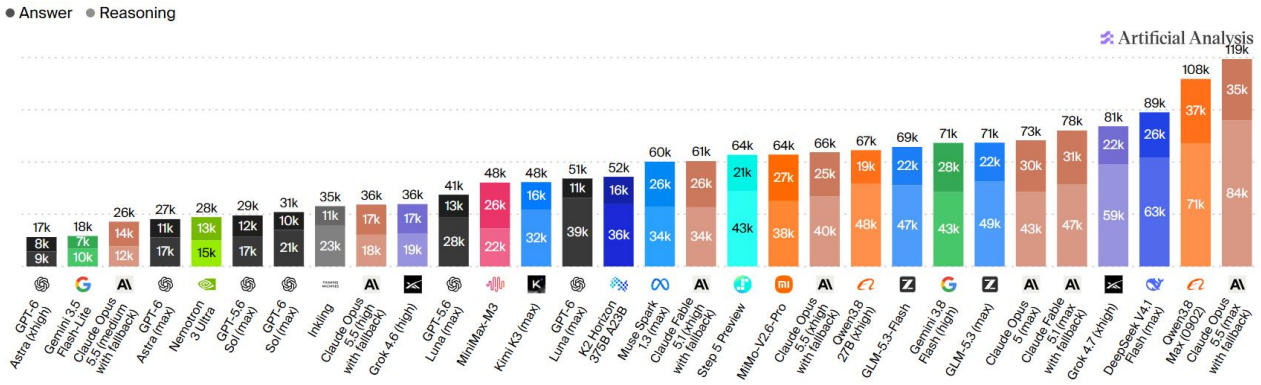
参考海外 AI 模型第三方评测机构 Artificial Analysis 横向多模型评测体系，智能综合指数得分（Intelligence Index Score）用来衡量模型整体智能上限，而单位任务平均输出词元（Token，以下简称“Token”）量（Output Tokens per Task）、单位智能任务加权完成成本（Cost per Intelligence Index Task）两项工程指标，则从资源开销与经济成本维度还原模型真实落地价值。从横向评测结果可以观察到行业一个关键特征：综合得分与 Token 消耗、单位任务成本并不呈简单线性正相关。部分头部模型智能综合指数得分处于 51-58 的第一梯队，但不同模型单任务输出 Token 规模差异较大，对应的单位智能任务加权成本分化进一步放大；同时也存在一批模型智能综合指数得分处于 44-48 区间，依托推理调度、后训练、缓存机制等工程优化，在接近的智能能力水平下实现低于第一梯队的 Token 开销与单位任务成本。这种“智能得分-Token 消耗-执行成本”的错位分化，意味着开发者选型不再只看跑分，能力、Token 开销、综合成本需要综合权衡。

图 2：Artificial Analysis 全球大模型智能综合指数排行榜（截至 2026 年 9 月 27 日）



资料来源：Artificial Analysis，渤海证券研究所

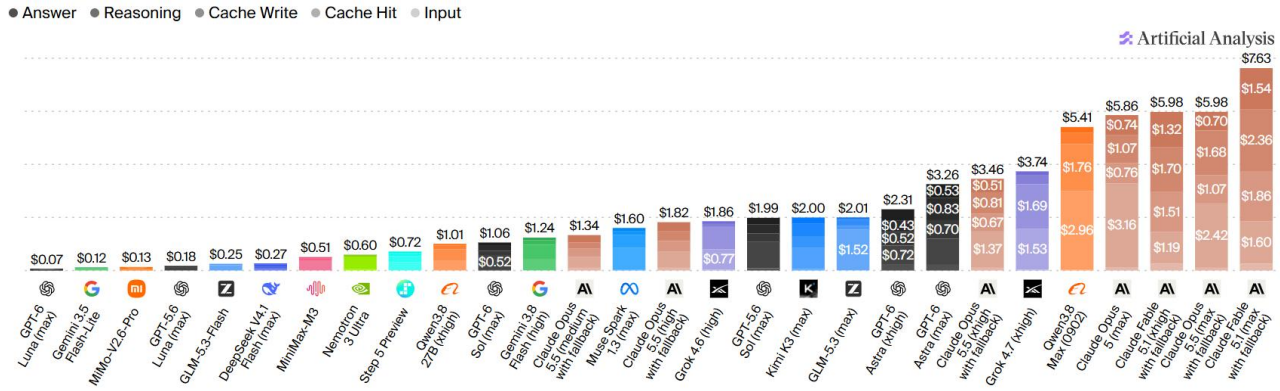
图 3：Artificial Analysis 全球大模型单位任务平均输出 Token 量排行榜（单位：Token，截至 2026 年 9 月 27 日）



资料来源：Artificial Analysis，渤海证券研究所

仅供内部参考，请勿外传！

图 4：Artificial Analysis 全球大模型单位智能任务加权完成成本排行榜（单位：美元，截至 2026 年 9 月 27 日）



资料来源：Artificial Analysis，渤海证券研究所

1.2Agent 拉动 Token 需求增长，国产模型全球化落地提速

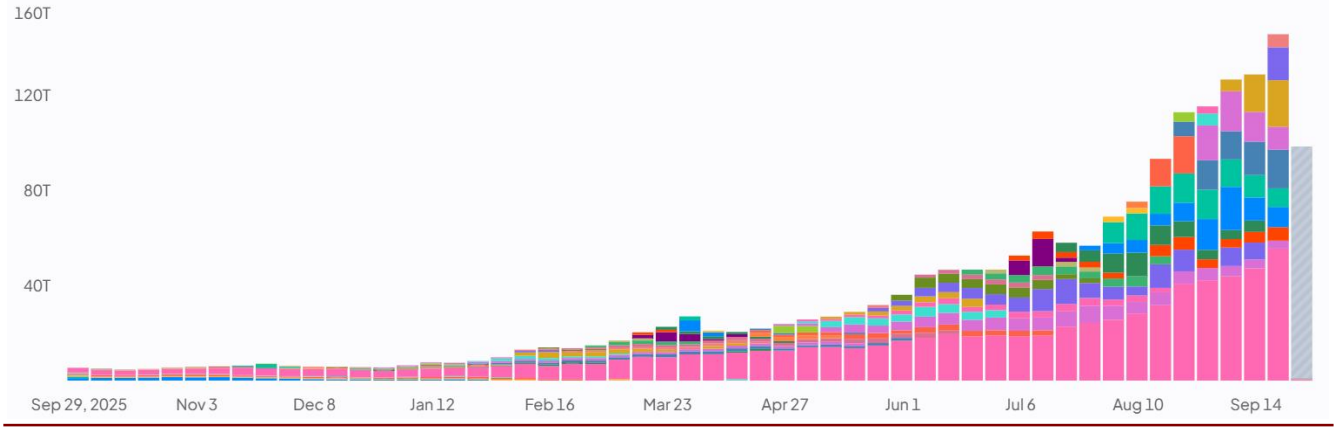
Agent 能力迭代带动 Token 消耗上行，成为模型即服务（Model-as-a-Service，以下简称“MaaS”）市场调用量与企业付费增长的核心增量。伴随 Agent 能力成为各大模型厂商迭代的核心重点并实现持续增强，长链路、多步骤的智能体任务会显著拉高 Token 消耗量，Agent 相关业务已经成为 MaaS 调用规模扩张、企业付费增长的核心增量来源。普通对话交互以短轮次问答为主要形式，交互链路简短，整体资源开销可控；而 Agent 在执行任务的过程中，需要持续读取历史上下文、多次发起外部工具调用、解析处理工具返回结果，并基于反馈开展循环迭代修正，整套工作流带来的整体 Token 开销会远高于传统聊天交互。下游智能体应用持续火热也进一步放大了这一效应，OpenClaw、Hermes Agent 等开源智能体应用产品快速破圈，获得行业高度关注，直接拉动全行业大模型调用规模高速扩容。相较于功能相对单一的传统对话式 AI 产品，智能体依托自主规划、多步骤协同执行的核心能力，可以承接工业、办公、政企服务等多元场景下复杂度更高的复合型业务需求，既拉长了整体用户交互链路，也有效提升用户使用黏性，持续推高全行业词元流量消耗，底层大模型的 API 调用需求随之同步大幅上行，并进一步带动大模型 API 接口服务、配套算力租赁等上下游产业链需求持续释放。

海外头部 AI 模型 API 聚合平台 OpenRouter 流量数据印证国产模型海外认可度提升，开发者选型兼顾能力、成本与实用性。OpenRouter 数据显示，自 2026 年年初开始，整体周度大模型调用规模呈现显著抬升态势；在 9 月 1 日-9 月 27 日的调用量 TOP10 榜单当中，GLM 5.3 Flash、Hy4 preview、DeepSeek V4 Flash 0731、MiMo-V2.5 等多款国产大模型跻身前列，印证国产大模型的综合能力与全球市场认可度正在快速提升。结合前文 Artificial Analysis 多维度评测结果可以发现一个值得关注的市场现象：榜单当中调用量靠前的国产模型，并不完全是智能综合指数得分排名最高的模型。部分国产模型依靠突出的单位任务成本优势，以及 Agent 工具调用环节扎实的工程实用性，收获了大量来自开发者的实际业务流量。这充分体现出产业端开发者已经形成“能力-成本-实用性”的综合选型逻辑，不再

仅供内部参考，请勿外传！

简单以实验室榜单跑分作为唯一评判标准。国产模型的高性价比优势，正在逐步转化为全球化落地的核心竞争力，也为国内 AI 模型、算力服务出海筑牢现实业务基础。

图 5：OpenRouter 周度大模型调用量情况（单位：万亿 Tokens）



资料来源：OpenRouter，渤海证券研究所

图 6：OpenRouter 周度模型调用量排行榜（2026 年 9 月 1 日-9 月 27 日）

1.	GLM 5.3 Flash by z-ai	57.5T tokens ↑>999%	6.	MiMo-V2.5 by xiaomi	21T tokens ↓29%
2.	Hy4 preview by tencent	56.4T tokens ↑>999%	7.	Nemotron 3 Ultra (free) by nvidia	18.7T tokens ↑27%
3.	GPT-5.6 Luna by openai	51.9T tokens ↑140%	8.	DeepSeek V4 Flash 0423 by deepseek	18T tokens ↓24%
4.	DeepSeek V4 Flash 0731 by deepseek	44T tokens ↑4%	9.	Hy3 by tencent	16.6T tokens ↓51%
5.	DeepSeek V4.1 Flash by deepseek	40.3T tokens new	10.	Space Bunny Alpha by stealth	13.9T tokens new

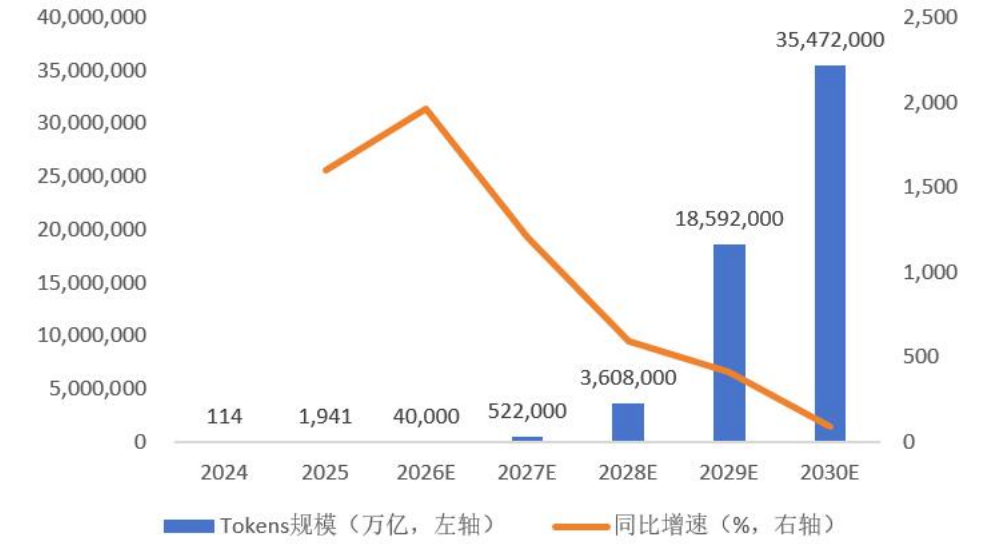
资料来源：OpenRouter，渤海证券研究所

1.3Token 经济持续发展，MaaS 行业竞争转向综合能力

国内市场方面，Token 调用规模的高速增长同样印证 MaaS 行业的规模化发展趋势。根据国际数据公司（International Data Corporation，以下简称“IDC”）测算数据，2025 年国内模型即服务市场的 Token 调用规模已达到 1941 万亿 Tokens。IDC 同时预计，2030 年 Token 调用规模有望达到 3547.2 兆 Tokens，未来数年国内 Token 总量将保持指数级扩张，市场体量将实现量级跃升。展望 2026-2030 年，Agent 智能体应用的持续渗透将成为驱动 Token 消耗规模扩张的核心动力。智能体业务所对应的复杂工作流、多轮工具调用、循环自我纠错机制，将持续抬升单任务的 Token 资源开销，进而带动国内整体 MaaS 市场需求持续上行。

仅供内部参考，请勿外传！

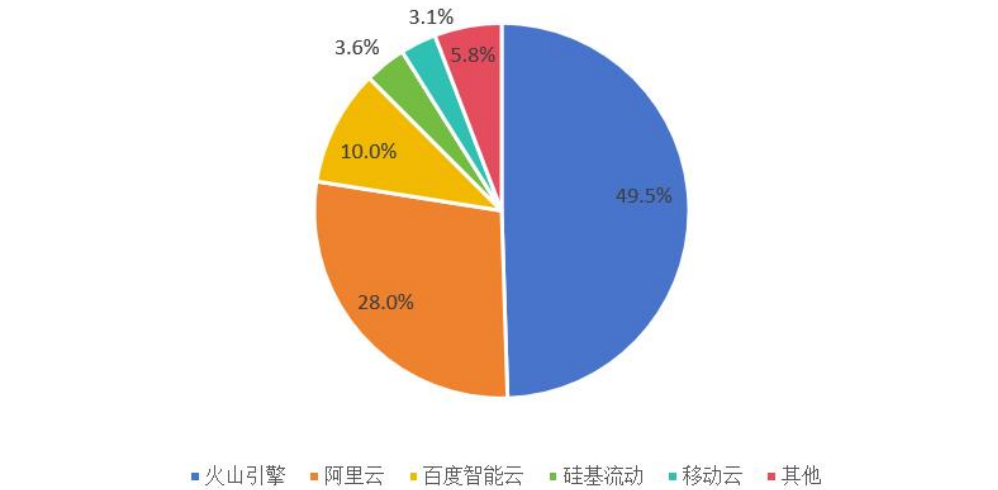
图 7：2024-2030 中国模型及服务市场规模及预测情况（单位：万亿 Tokens）



资料来源：IDC，渤海证券研究所

国内公有云 MaaS 市场呈现头部集中、多元主体差异化竞争的格局，不同厂商依托自身资源禀赋展开布局。从国内公有云 MaaS 市场竞争格局来看，根据 IDC 统计数据，以 Token 调用量口径统计，2025 年火山引擎凭借模型产品与算力调度优势占据近半数市场份额，位居行业首位；阿里云、百度智能云紧随其后，硅基流动、移动云同样具备一定市场规模。除此之外，腾讯云、华为云、商汤科技、天翼云等亦为赛道重要参与者，不同厂商依托自身资源禀赋实现差异化布局：云厂商普遍具备自有算力底座，兼顾模型能力与弹性服务供给；部分独立模型厂商依托开源模型优势抢占细分客户，国内 MaaS 市场已经形成多主体同台竞争的态势。

图 8：2025 年中国 MaaS 市场份额情况（按 tokens 份额）



资料来源：IDC，渤海证券研究所

伴随 Token 经济逐步成型、市场参与主体快速扩容，MaaS 行业的竞争逻辑也正在发生深刻转变。IDC 指出，未来 MaaS 赛道的核心竞争维度将聚焦成本优化、模型性能、工具链与应用开发支撑三大方向。成本层面，企业端 Token 日均消耗量持续攀升，Agent 类应用进一步放大整体资源消耗，市场竞争的重心不再局限于单纯

仅供内部参考，请勿外传！

压低单 Token 报价，能否帮助客户降低完整业务链路的综合使用成本、保障高质量有效 Token 输出，将成为厂商核心竞争力。性能层面，大量产业落地场景对服务并发、吞吐能力提出严苛要求，算力供给约束短期仍将客观存在，依托底层算力调度、推理优化实现效率提升的服务商，将在客户选型过程中占据优势。工具链与开发支持层面，此前国内 MaaS 市场选型更多聚焦价格因素，受模型快速迭代影响，企业大多以直接调用 API 为主，较少开展二次调优；而随着垂直行业落地需求不断深化，智能体开发编排成为企业刚需，MaaS 平台配套的调试、评测、Agent 编排等工具链完备程度，在后续客户评估体系当中的权重将会持续抬升。

2.海内外模型厂商探索多元变现，商业化模式持续验证

2.1国内大模型商业化模式分化，四类变现路径并行发展

国内大模型演化四类主流变现路径，各类模式具备差异化特征。根据北京第四波科技智库、中关村天成创新研究中心联合发布的《中国 AI 大模型商业化报告（2026 年上）》，当前国内大模型厂商已演化出四类主流变现路径，各类模式在收入属性、规模化复制能力、目标客户结构方面存在较为显著的差异化特征，共同构成行业的商业收入底座。

（1）公有云 MaaS（以 API 接口调用、Token 按量计费）是当前独立大模型厂商增长最快的业务形态。该模式将模型能力封装为标准化 API 接口，下游开发者与企业按照实际 Token 调用量进行付费，业务具备较强的规模化复制潜力。行业整体正逐步向 Token 按量计费的标准化服务模式迁移。

（2）私有化及本地化部署采用项目制叠加年度运维服务的交付形态，厂商将模型部署至客户本地机房或者客户专属的隔离算力集群，向客户收取软件许可、项目实施交付以及后续持续运维服务费。该模式单项目客单价较高，但项目交付流程重，业务扩张高度依赖实施人力供给，政务、金融、能源等政企客户是该类方案的主要采购方。

（3）垂直行业定制解决方案立足于通用大模型底座，面向垂直行业完成模型微调、领域知识库构建、Agent 工作流编排开发，并配套完整项目实施交付。该业务需要投入大量垂直行业专家开展需求梳理与落地适配，业务规模化扩张会受到行业人力供给的约束，科大讯飞等厂商已经依靠该模式在教育、政务场景完成商业化落地验证。

（4）会员与席位订阅服务包含个人 C 端会员订阅以及海外市场较为成熟的企业席位订阅两类子形态。面向 C 端个人用户的会员订阅，通过付费解锁更高模型额度、高级功能，是国内大模型厂商探索较多的变现路径，但国内独立大模型厂商 C 端付费带来的营收占比普遍偏低，多数厂商仍然保持基础功能免费、增值功能付费的

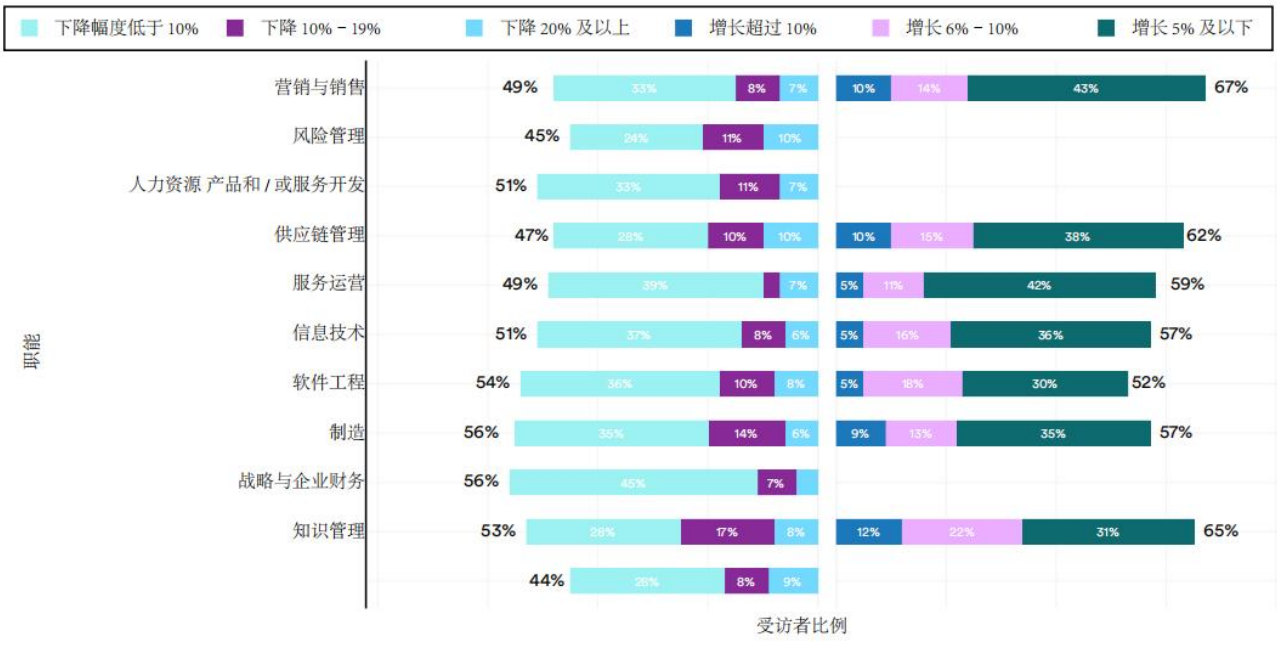
仅供内部参考，请勿外传！

产品策略，C 端订阅更多作为产品补充收入，尚难以覆盖模型推理与研发成本。面向企业内部员工、按账号席位收费的企业订阅模式，在海外厂商收入结构中占据一定比重；反观国内市场，企业现阶段更多优先选择 API 调用或者私有化交付，B 端席位制产品落地尚处在发展早期，业务收入贡献相对有限。

2.2AI 释放降本增收潜力，各职能收益体现不同

AI 大模型的产业落地，核心价值体现在助力政企主体实现降本与增效增收，不同业务职能板块的收益兑现呈现明显的结构性差异。从降本逻辑来看，AI 模型主要通过替代重复性知识类工作，压缩人力投入与信息化建设成本，企业层面覆盖合同审阅、知识库检索、工单处理、代码开发、文献挖掘等场景，依靠自动化文档处理、智能检索减少基础岗位工时消耗；政务端则落地于公文撰写辅助、政策检索、政务热线、申报材料预审等业务，部分地区依托集约化大模型平台减少软硬件重复采购，节约财政信息化预算。根据麦肯锡 2025 年调研数据，软件工程、制造职能板块的受访企业对 AI 带来成本节约的感知最为突出，对应受访占比均达到 56%，但多数职能的成本下降集中在 10%以内，实现大幅降本的企业占比有限。在营收增益维度，价值兑现更多集中在前端业务环节，营销与销售板块有 67%受访者观测到收入改善，知识管理、供应链管理分别有 65%、62%的受访企业实现营收层面正向反馈。整体来看，AI 在各职能部门的价值释放存在明显结构性分化，后台职能更多凸显成本节约效应，面向市场、产品的前端业务则更容易落地营收增量。

图 9：2025 年各职能人工智能使用的成本降低与营收增长情况



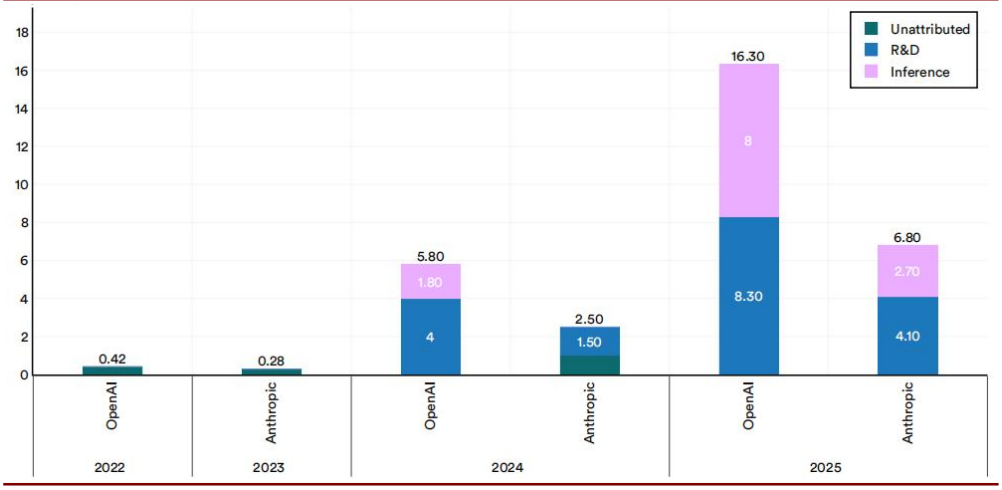
资料来源：麦肯锡，渤海证券研究所

仅供内部参考，请勿外传！

2.3算力布局路径不同，厂商资本开支呈现分化

AI 模型企业的算力投资是支撑模型训练迭代与 MaaS 业务 Token 供给的底层基础，算力投入可划分为训练、推理两大体系，不同类型市场主体的算力获取模式、资本开支结构呈现显著分化。根据 Epoch AI 统计，海外头部 AI 企业年度算力支出近年实现大幅扩张，OpenAI、Anthropic 算力投入均出现快速增长，算力成本伴随商业需求扩张持续抬升，前沿大模型产业的运行逻辑已经与大规模算力采购深度绑定。从算力资源获取方式来看，国内云厂商如百度、阿里、腾讯等以自建算力中心作为底座核心，同时采购外部算力资源承接业务峰值压力；而独立大模型创业企业受限于大额固定资产投入约束，训练环节普遍采用算力长约租赁模式，仅少数头部企业小规模搭建自有集群用于底座模型迭代。针对 MaaS 推理业务，由于 API 调用 Token 流量具备较强波动属性，行业普遍执行混合调度策略，基础稳态业务流量由自有或长约锁定算力承载，突发调用峰值则通过外部弹性算力完成补充。

图 10：2022-2025 年 Anthropic 与 OpenAI 年度算力支出情况（单位：十亿美元）



资料来源：Epoch AI，渤海证券研究所

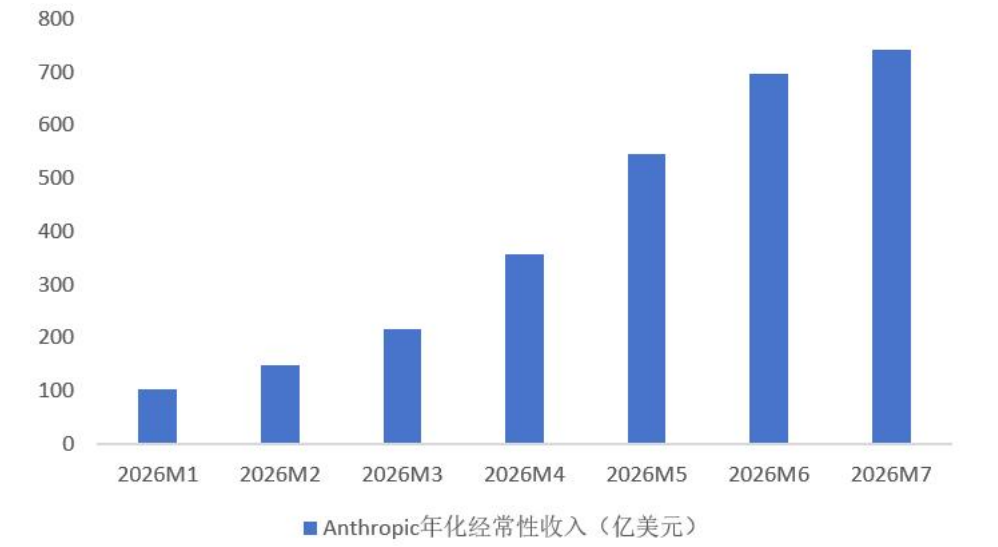
AI 行业资本开支更多向 GPU 服务器、机房配套、电力基础设施等硬件端倾斜，软件研发带来的资本投入占比相对有限，各类市场主体的投入在会计统计口径上存在客观差异。云厂商直接完成服务器硬件采购，相关支出纳入资本开支核算；独立模型企业账面资本开支整体相对克制，较少开展大规模机房自建，算力长约合约归属于承诺性运营支出（OPEX），不体现在当期资本开支当中，但会形成远期的支付义务。Token 消耗本质对应推理算力资源的耗用，伴随 Agent 长链路任务落地普及，推理算力需求持续走高，推理成本也逐步成为模型厂商成本端的主要压力。国内市场层面，云厂商持续加大 AI 算力中心建设力度，独立模型企业仍以算力租赁作为主流选择，仅有部分头部主体适度布局自建训练集群，实现固定资产投入的合理管控。

仅供内部参考，请勿外传！

2.4海外闭源大模型：B 端企业 API 驱动 ARR 高速增长

Anthropic 是全球范围内商业化推进速度领先的闭源大模型厂商，公司坚持闭源产品路线。公司 Claude 系列模型凭借超长上下文理解、Agent 工具调用能力构建起自身差异化竞争壁垒，业务收入主要来源于公有云 MaaS（以 API 接口调用、Token 按量计费）以及企业席位订阅，C 端订阅贡献占比较低，几乎不开展私有化及本地化部署项目交付。依托持续迭代的模型技术储备以及成熟的 B 端商业化服务体系，Anthropic 不断获得产业资本与海外政企客户的认可，商业化兑现的增长潜力充足。根据海外科技行业数据估算平台 TickerTrends 的市场估算数据，2025 年末公司年化经常性收入（Annual Recurring Revenue，ARR）约 90 亿美元，到 2026 年 5 月该数值攀升至 546 亿美元，2026 年 7 月公司年化经常性收入进一步达到 741 亿美元，大部分营收由 B 端企业 API 调用业务贡献，长上下文处理与复杂 Agent 任务带来更高的单任务 Token 消耗，持续推动营收规模快速上行。

图 11：Anthropic 年化经常性收入情况



资料来源：TickerTrends，渤海证券研究所，注：年化经常性收入为估算值，非官方披露数据

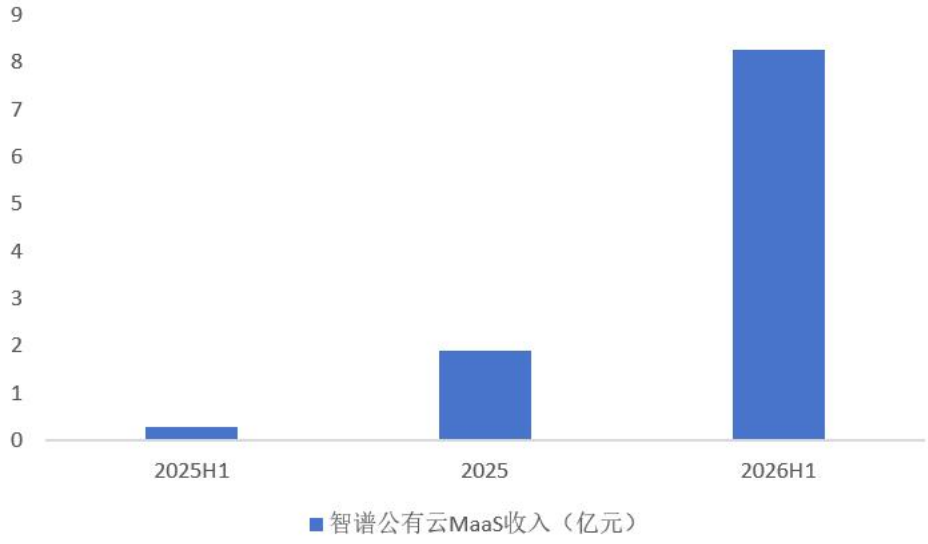
Anthropic 商业化稳步推进，模型能力优势持续兑现收入增长。公司的商业化进程呈现清晰的演进路径，发展前期重点打磨底层模型基础能力，依靠长文本处理优势切入大型企业客户群体，后续随着 Agent、工具调用能力持续迭代，进一步抬升单任务 Token 消耗，带动年化经常性收入持续走高。公司 Claude 系列模型在复杂 Agent 任务、超长上下文评测中处于全球第一梯队，模型层面的能力优势直接转化为 B 端客户付费的商业价值。2026 年 5 月 28 日 Anthropic 完成规模 650 亿美元的 H 轮私募融资，投后估值达到 9650 亿美元，充沛的资金储备将持续赋能算力建设、模型迭代以及全球客户拓展，有望推动公司营收增速继续提升，进一步打开长期商业化成长空间。

仅供内部参考，请勿外传！

2.5国内开源大模型：从私有化项目向公有云 MaaS 快速转型

智谱 AI 作为国内头部大模型厂商，底座模型开放权重，商业版本采用闭源模式，在产业落地过程中议价能力稳步提升，商业化兑现节奏持续提速。公司发展早期主要依靠政企私有化及本地化部署项目获取收入，下游智能体应用、政企数字化需求不断释放带动 GLM 系列产品市场需求持续放量，公有云 MaaS（以 API 接口调用、Token 按量计费）业务逐步成长为核心增长曲线。公司 2025 年公有云 MaaS 收入为 1.9 亿元，2026 年上半年该业务收入达到 8.25 亿元。伴随着模型能力的跃迁，公司主营业务收入结构在 2026 年上半年完成显著置换，公有云 MaaS 收入占比从 2025 全年的 26.3%提升至 86.5%，而模型私有化及本地化部署收入占比由 73.7%下降至 13.5%。

图 12：智谱公有云 MaaS（以 API 接口调用、Token 按量计费）收入情况



资料来源：智谱公司公告，渤海证券研究所

下游调用规模快速扩张、企业客户不断扩容的背景下，智谱 AI 多次对产品实施阶梯调价。2026 年 2 月智谱发布新一代旗舰底座 GLM 5，相比前代 GLM 4.7 上调价格 50%，同时上调 GLM Coding Plan 代码订阅套餐定价，套餐整体涨幅约 30%。调价过程中配套存量用户保护机制，保障老用户原有权益不变，仅针对新订单取消首购优惠、保留长周期订阅折扣，以此保障高负载场景服务稳定性。3 月面向 OpenClaw 智能体长链路任务推出专项优化的 GLM 5 Turbo，API 定价在 GLM 5 基础上上调 20%；4 月迭代发布 GLM 5.1，接口定价继续上调 10%；7 月底 GLM Coding Plan 切换积分峰谷计费模式，同样维持存量用户权益；8-9 月陆续推出 GLM 5.3 系列版本，新增高速档位 GLM 5.3 FlashX，依靠新品差异化定价承接高吞吐、低时延的 Agent 任务。多轮调价动作，既体现出下游大模型供给偏紧、Token 消耗持续上行的行业现状，也印证头部自研厂商议价能力正在稳步抬升，国内 AI 产业商业化落地与盈利兑现正在持续加快。业务层面，智谱 AI 正在完成由项目制向 Token 按量计费模式的转型，整体商业模式得到持续优化。

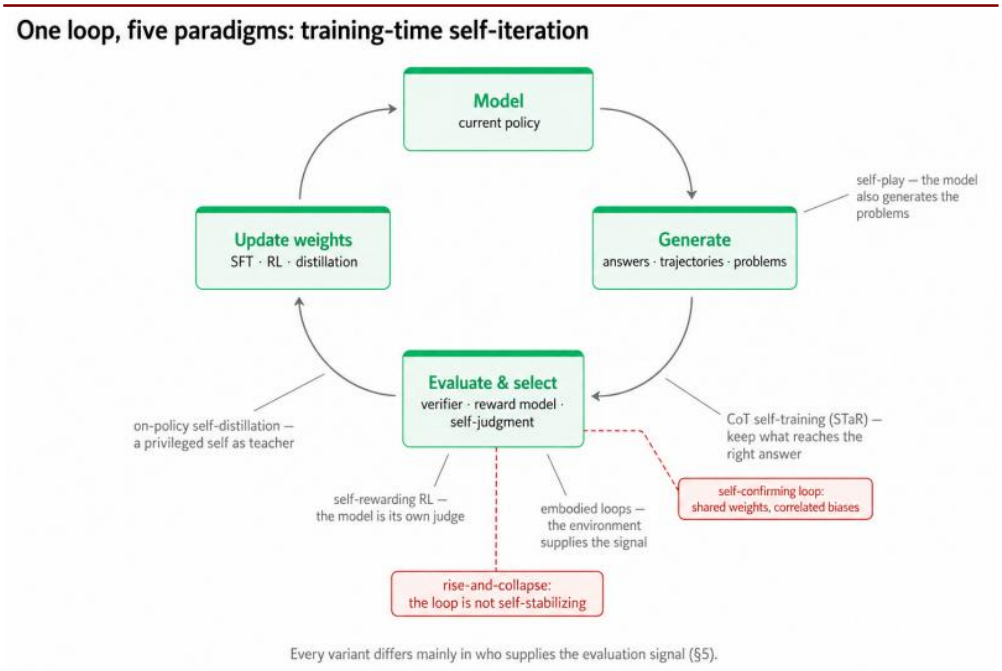
仅供内部参考，请勿外传！

3.RSI 驱动递归自我改进，大模型迈向自进化体系

3.1RSI 革新模型迭代范式，依托场景反馈实现自优化

递归自我改进（Recursive Self-Improvement，以下简称“RSI”）是一类具备自我进化潜力的 AI 系统。该系统在完成各类实际任务的过程中，会沉淀大量真实运行反馈，AI 可以利用这些反馈持续优化自身能力，更进一步还能迭代升级“用来改进 AI 的整套底层机制”，由此形成一套闭环的递归自我提升流程。大模型传统的能力优化手段主要包含有监督微调（Supervised Fine-Tuning，SFT）、人类反馈强化学习（Reinforcement Learning from Human Feedback，以下简称“RLHF”）。这类优化方式需要大量人工参与，由研究人员或者标注人员整理、输出高质量训练标注数据，模型基于人类产出的数据完成学习提升。与之形成区别的是 RSI 范式，它不再高度依靠人工制作标注数据集，AI 在 Agent 执行真实业务的场景里，自己从任务成败、交互结果中提取可供学习的信号，自主利用这些现实场景信号完成迭代进化，以此降低模型迭代对人工标注资源的依赖。

图 13：模型训练阶段自迭代示意图



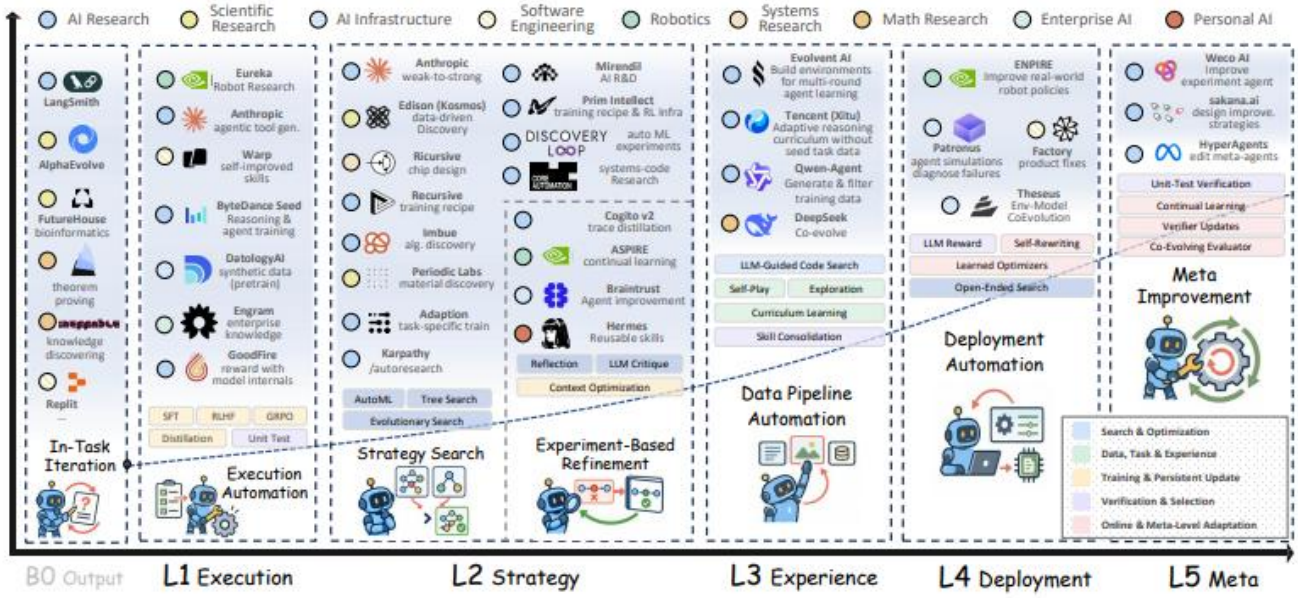
资料来源：arXiv，渤海证券研究所

基于发布于预印本学术开放存档平台 arXiv 的论文所提出的分级框架，RSI 能力可划分为从 L0 到 L5 的逐级演进体系。L0 层级为基础大模型，仅能够被动执行外部给定的任务，完全不具备任何自我改进的能力；L1 层级对应依托外部人类标注反馈开展微调，也就是行业广泛使用的 RLHF 方案，也是目前绝大多数主流大模型所处的能力阶段；进入 L2 层级之后，Agent 在执行业务任务的过程中，可以自

仅供内部参考，请勿外传！

动采集任务成功或者失败的反馈信息，并将这些信息回传用于模型的训练迭代；到 L3 层级，模型已经可以自主设计评估任务，自动生成对应的训练样本，以此驱动自身能力迭代升级；L4 层级下，模型具备优化自身整套训练、评测流程机制的能力；最高阶的 L5 属于元进化阶段，系统能够自主发明全新算法与网络架构，实现能力层面的跨越式跃迁，该层级目前仅属于理论层面的远期技术目标。

图 14：RSI 能力 L0-L5 逐级演进框架图



资料来源：arXiv，渤海证券研究所

3.2RSI 能力持续演进，海内外厂商积极布局

当前全球大模型行业整体的 RSI 能力正处在 L2-L3 的发展区间。大模型目前已经可以借助 Agent 执行任务产生的各类反馈辅助模型迭代优化。现阶段 RSI 还未达到元进化阶段的理想状态，行业在实际落地过程中主要分化出两条可行的工程实现路线。其中数据递归自我改进（Data-Recursive Self-Improvement，Data-RSI）路径聚焦于利用任务执行过程沉淀得到的反馈数据来迭代优化训练集，Agent 置身于真实业务环境中完成各类任务，完整留存任务输入、模型输出内容以及任务成功或失败的结果信号，经过筛选后将高质量样本补充进入训练数据集，以此实现真实业务场景的数据反向赋能底座模型迭代。另一条模型递归自我改进（Model-Recursive Self-Improvement，以下简称“Model-RSI”）路径则围绕智能体调度框架（Agent Harness，以下简称“Agent Harness”）展开，侧重实现自动评测与策略层面的迭代，通过搭建 Harness 自动化闭环框架，由 Agent 自主生成测试任务，完成任务执行、结果收集以及自动化优劣评判，再依据评测输出调整模型推理策略、提示策略与工具调用策略，该路径并不以重新更新模型权重作为必要条件。

海内外头部大模型厂商均已经围绕 RSI 两类路径开展对应的工程实践工作。海外的 Anthropic、OpenAI 将研发资源大量投入模型评测与 Agent 闭环体系的探索工作；国内 DeepSeek 也对外公开披露 Agent Harness 相关的技术实践成果，其余

仅供内部参考，请勿外传！

国内头部大模型企业同样在逐步搭建业务任务反馈闭环，将真实场景反馈数据用于底座模型的持续迭代优化，推动 RSI 技术从理论框架走向产业落地。

4.大模型持续拓展应用边界，物理 AI 与 AI4S 迎来发展机遇

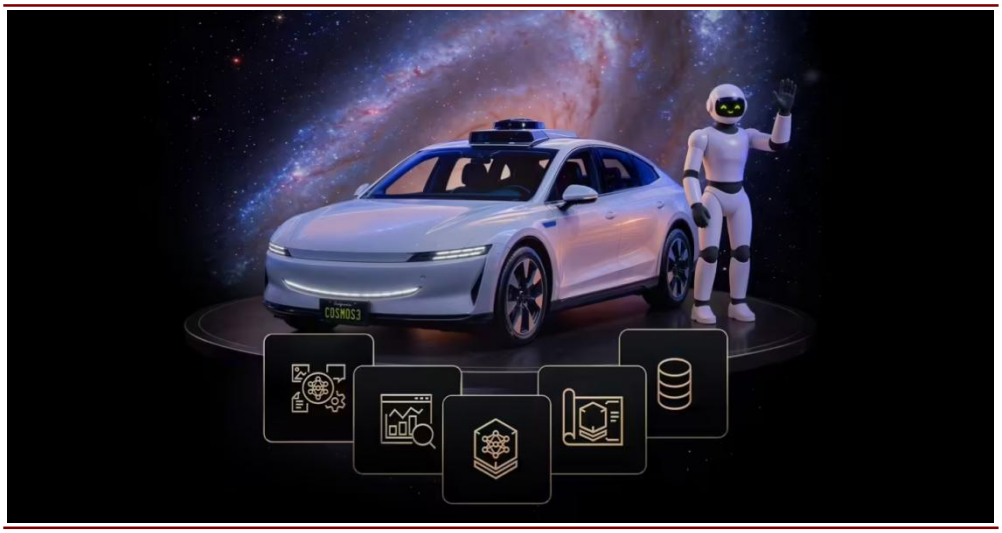
4.1物理 AI 衔接虚拟与现实，软硬件协同赋能具身智能

物理 AI 是人工智能与现实物理系统深度融合的前沿技术赛道，具备实体经济落地价值与广阔成长空间。区别于聚焦文本、图像等纯数字内容生成的通用生成式大模型，物理 AI 不再局限于线上虚拟内容创作，其核心是让 AI 掌握空间、力学、物体碰撞、运动演化等客观物理规律，依托世界模型生成具备真实物理逻辑的仿真数据，进而完成场景仿真、动作规划与实体交互，打通虚拟数字空间与现实世界，真正实现人工智能从感知理解数字信息，到落地现实物理空间执行决策、完成实体动作的跨越。从系统架构来看，物理 AI 以大模型作为智能“大脑”，接收摄像头、各类传感器输出的感知信号，经过理解、规划之后输出动作指令，直接控制硬件实体完成现实世界的各类操作，下游可广泛赋能人形机器人、自动驾驶、无人机、智能制造等领域。

英伟达推出全球首款全开源全模态物理 AI 大模型 Cosmos 3，有望推动物理 AI 技术发展提速。Cosmos 3 属于面向物理人工智能的开放世界基础大模型，在单一系统当中融合视觉推理、世界生成与动作预测能力，能够原生理解并生成文本、图像、视频、环境音效及动作内容，拥有业界领先的物理仿真精度。该模型采用混合 Transformer 架构，结合推理 Transformer 与生成类 Transformer，先解析物体交互、运动规律以及时空关联关系，再开展视频生成与动作轨迹预测。产品层面 Cosmos 3 形成分层版本矩阵，适配物理人工智能不同研发阶段的需求，其中 Cosmos 3 Super 主打超高物理仿真精度，面向机器人、自动驾驶领域的高精度模型微调场景；Cosmos 3 Nano 侧重轻量化推理，可快速完成高清视频解析与智能动作推理；Cosmos 3 Edge 则聚焦终端部署，满足边缘设备实时推理的行业刚需。分层的产品布局叠加英伟达完备的 AI 全栈生态，有望降低行业整体的物理 AI 研发门槛，持续助推物理 AI 技术的商业化落地进程。

仅供内部参考，请勿外传！

图 15：英伟达推出面向机器人、自动驾驶等使用场景的物理 AI 大模型 Cosmos 3



资料来源：英伟达官网，渤海证券研究所

4.2AI4S 重塑药物研发链条，海内外药企加速布局

人工智能驱动的科学研究（AI for Science，以下简称“AI4S”）正加速落地生命科学领域，AI 大模型正在重塑药物研发全流程。AI4S 被视作继实验、理论、计算、数据科学之后的科研第五范式，推动 AI 模型深度渗透生命科学的完整科研链条。在药物研发领域，AI 已经贯穿靶点挖掘验证、蛋白与分子结构预测、全新药物分子生成、合成路线设计、临床试验方案设计、海量科研文献挖掘等关键环节，能够有效压缩新药研发整体周期，降低传统模式下高昂的试错成本。

图 16：AI for Science 被视作继实验、理论、计算、数据科学之后的科研第五范式



资料来源：《科学计算与 AI4S 基础设施白皮书》，渤海证券研究所

全球 AI4S 在生命科学领域已落地多项标杆案例，海外大厂率先完成产业布局，国内厂商同步跟进技术迭代与场景落地。DeepMind 推出的模型 AlphaFold 实现蛋白质结构预测并斩获诺奖，充分验证了 AI 在生命科学硬核科研场景中的核心应用价值。海外企业层面，头部大模型厂商持续深耕生命科学赛道，依托专业科研大模型平台赋能医药研发全流程创新。其中 Anthropic 针对性打造科研专属大模型能力，

仅供内部参考，请勿外传！

聚焦生物医药领域核心科研难题开展技术攻坚；近期丹麦制药龙头诺和诺德与 Anthropic 达成深度合作，依托 Claude Science 科研大模型平台推进新药发现与开发工作，聚焦攻克高价值核心科学问题，同时借助大模型工具优化研发环节的软件开发效率。作为全球知名的代谢类药物研发生产企业，诺和诺德管理层明确，AI 技术不仅能够显著提升医药研发效率、缩短基础科研到产品商业化的落地周期，更能够开拓全新的科研探索路径，助力行业深化对人体生物学机理、药物作用机制的认知与推演。国内市场层面，本土 AI4S 企业依托自主技术平台持续深耕生物医药赛道，不断迭代优化核心技术能力，持续推进创新药物研发管线布局，加速 AI 技术在生命科学科研与产业场景的规模化落地。

跨国药企持续加码 AI 药物研发，当前落地价值集中在科研流程自动化。诺和诺德、礼来、默沙东、罗氏等多家跨国药企均在加大 AI 领域布局，希望借助技术提速药物发现和产业化进程。从行业现实来看，当前已有部分 AI 辅助挖掘得到的候选药物进入人体临床试验阶段，但完全由 AI 主导研发的药物尚未实现获批上市。现阶段 AI 给制药行业带来的最直接价值，更多体现在科研重复性工作的自动化，包括科研文献归集梳理、海量数据解析、临床试验方案设计等环节，AI4S 在医药研发领域仍处在持续验证、逐步兑现价值的发展阶段。

5.相关公司基本情况介绍

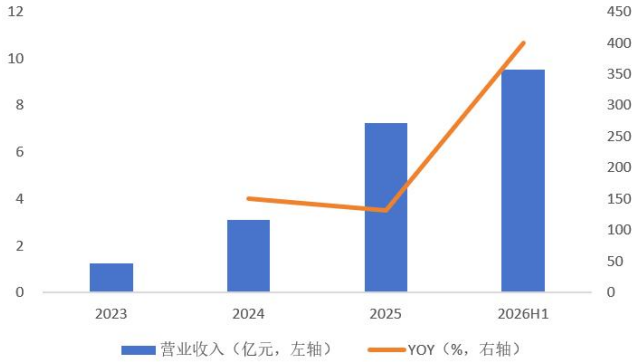
5.1智谱

北京智谱华章科技股份有限公司成立于 2019 年。公司主营业务为通用人工智能基座大模型研发，依托原创 GLM 预训练模型框架，提供模型即服务（MaaS），产品包含开放平台及 API、企业级通用大模型、代码模型、多模态模型与智能体相关服务。公司 GLM 系列模型形成完整基座模型矩阵，适配多款国产芯片，MaaS 平台面向政企机构客户、开发者及个人用户提供大模型调用、微调、智能体开发一体化能力，覆盖全球多个国家和地区，持续推进基座模型向工业级智能体工程方向落地。

公司收入高速增长，MaaS 开放平台及 API 业务成为增长核心。2026 年上半年，公司实现营业收入 9.54 亿元，同比增长 399.74%；实现归母净利润-20.71 亿元，上年同期亏损 23.49 亿元。收入增长主要源于模型能力迭代带动 API 调用规模大幅提升，开放平台及 API 收入占比提升至 86.5%。2026 年上半年，公司持续完善 GLM 系列模型能力，推进代码、安全领域智能体落地，搭建国产芯片规模化推理基础设施，单位词元推理成本显著下降；同时持续拓展全球开发者生态，MaaS 平台企业及开发者用户规模持续扩张

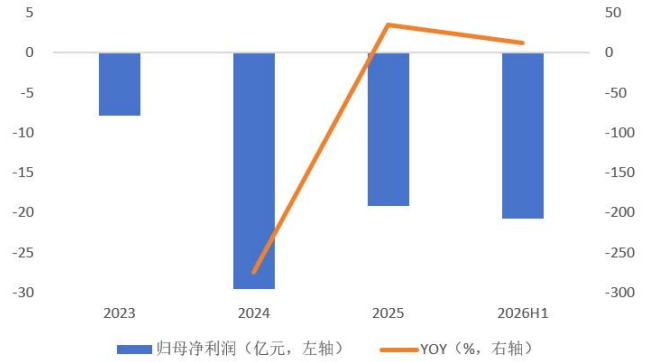
仅供内部参考，请勿外传！

图 17：智谱营收情况



资料来源：ifind，渤海证券研究所

图 18：智谱归母净利润情况



资料来源：ifind，渤海证券研究所

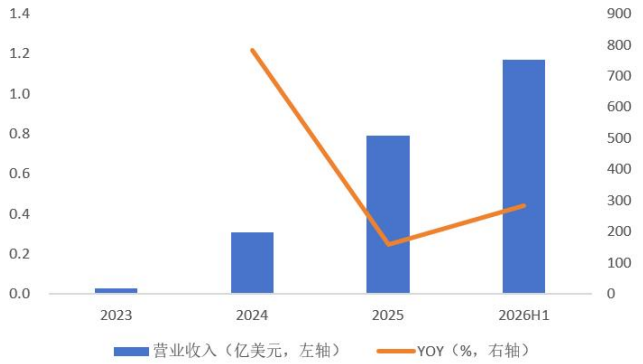
5.2MINIMAX

MiniMax Group Inc.成立于 2022 年。公司主营业务为全模态通用基础大模型研发，自主研发覆盖文本、语音、图像、视频及音乐的多模态模型体系，核心产品包含 M 系列文本大模型、Hailuo 系列视频生成模型、Speech 系列语音模型。依托自研基础模型，公司打造面向企业与开发者的开放平台，以及海螺 AI、MiniMax Agent 等 AI 原生产品。公司 M 系列大模型采用混合专家架构，具备百万级长上下文能力，在代码任务、智能体自主工作流场景具备突出性能，产品面向全球市场，覆盖 230 多个国家及地区的个人用户、开发者与企业客户。

公司收入高速增长，开放平台企业服务成为增长核心。2026 年上半年，公司实现营业收入 1.17 亿美元，同比增长 283.1%；实现归母净利润-3.58 亿美元，上年同期亏损 4.02 亿美元。收入增长主要得益于全球企业客户与开发者规模扩张，API 调用量持续提升，开放平台及企业 AI 服务收入占比提升至 63.4%。2026 年上半年，公司持续迭代全模态模型能力，推出 M3 模型强化代码与智能体任务性能，后续发布开源 H3 视频模型赋能商业内容创作；持续完善模型全栈推理效率，扩大全球用户生态，企业与开发者客户数量稳步增长。

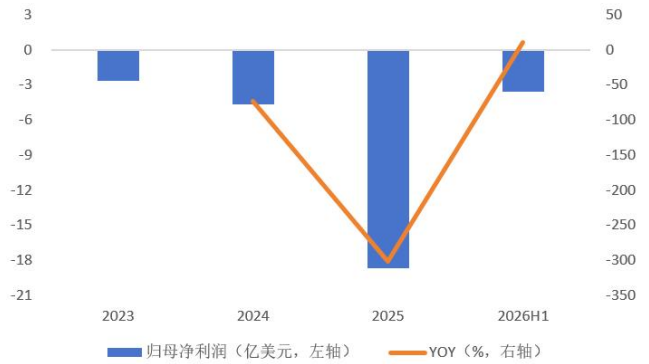
仅供内部参考，请勿外传！

图 19：MINIMAX 营收情况



资料来源：ifind，渤海证券研究所

图 20：MINIMAX 归母净利润情况



资料来源：ifind，渤海证券研究所

6.风险提示

宏观经济风险

全球经济增长放缓、通胀压力传导及地缘政治冲突升级，或将直接导致企业端缩减 AI 技术研发、数字化转型等资本开支，消费端智能终端需求亦将呈现疲软态势，对整体业绩表现构成潜在压力。

行业竞争加剧的风险

当前人工智能行业赛道热度高企，吸引多类市场主体涌入，行业竞争呈白热化态势。企业为维持竞争力需持续加大研发投入，市场端降价抢份额、客户争夺加剧等行为或将压制行业整体毛利率。

供应链波动加大的风险

受全球芯片供需失衡、地缘冲突扰动及极端事件影响，供应链稳定性面临严峻挑战。核心零部件短缺将直接导致企业生产交付周期延长、订单履约能力下降，而原材料价格上涨与物流成本攀升则进一步压缩企业利润空间。

AI 技术迭代速度不及预期的风险

AI 技术处于快速演进阶段，算法优化、算力突破等核心环节仍存技术瓶颈。若行业内企业因研发投入不足、技术判断偏差或人才储备短缺，导致 AI 技术迭代速度滞后于市场预期，将使其丧失核心技术优势，难以满足下游客户智能化需求，进而错失新兴应用场景机遇，前期研发投入回报效率亦将显著降低。

AI 应用落地效果不及预期的风险

AI 技术商业化落地进程受多重因素制约，一方面，AI 模型在实际场景中算法偏见

仅供内部参考，请勿外传！

等问题尚未完全解决；另一方面，客户对 AI 技术认知度有限及数据合规要求趋严，均制约应用渗透率提升。若落地效果不及预期，将直接影响企业市场口碑及订单获取能力，对行业商业化推进节奏形成拖累。

仅供内部参考，请勿外传！

分析师声明

本报告署名分析师在此声明：我们具有中国证券业协会授予的证券投资咨询执业资格并登记为证券分析师，以勤勉尽责的职业态度、专业审慎的研究方法，使用合法合规的数据和信息，独立、客观地出具本报告；本报告所表述的任何观点均精准地、如实地反映了研究人员的个人观点，结论不受任何第三方的授意或影响。我们所获取报酬的任何部分无论是在过去、现在及将来均不会与本报告中的具体投资分析意见或观点有直接或间接的联系。

风险提示及免责声明

投资者应自主作出投资决策并自行承担投资风险，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。

本报告由渤海证券股份有限公司

何合同或承诺的基础，不因接收者收到本报告而视其为本公司客户。本报告仅在相关法律许可的情况下发送，并仅为提供信息而发送，不构成任何广告。

本报告的信息来源于已公开的资料，本公司对该等信息的准确性、完整性或可靠性不作任何保证，不保证该信息未经任何更新。本报告所载的资料、意见及推测仅反映本公司于发布本报告当日的判断，本报告所指的证券或投资标的的价格、价值及投资收入可升可跌。过往表现不应作为日后表现的依据。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。本公司不保证本报告所含信息保持在最新状态。同时，本公司对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。

在任何情况下，本报告内容的全部或部分均不构成对任何人的投资建议。本报告中所指的投资及服务可能不适合个别客户，不构成客户私人咨询建议。在任何情况下，本公司、本公司员工或者关联机构不承诺投资者一定获利，不与投资者分享投资收益，也不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任。投资者务必注意，其据此做出的任何投资决策与本公司、本公司员工或者关联机构无关。在任何情况下，本公司、本公司员工或者关联机构不就本报告中的任何内容对任何投资做出任何形式的担保。

在所知情的范围内，本公司、本报告撰写人以及财产上的利害关系人与本报告所评价或作出明确估值和投资评级的证券无利害关系。本公司利用信息隔离墙控制内部一个或多个领域、部门或关联机构之间的信息流动。因此，投资者应注意，在法律许可的情况下，本公司及其所属关联机构可能持有并交易本报告中所提及公司的股份或其他财产权益，也可能为这些公司提供或者争取提供投资银行、财务顾问或者金融产品等相关服务。

市场有风险，投资需谨慎。投资者不应将本报告作为作出投资决策的唯一参考因素，亦不应认为本报告可以取代自己的判断。在决定投资前，如有需要，投资者务必向专业人士咨询并谨慎决策。

本报告版权为本公司所有 或个人不得以任何形式刊载

、发布或引用本报告全部或部分内容，亦不得从未经本公司书面授权的任何机构、个人或其运营的媒体平台接收、翻版、复制或引用本报告全部或部分内容。如征得本公司同意进行引用、刊载或转发的，需在允许的范围内使用，并注明出处为“渤海证券股份有限公司”且不得对本报告进行任何有悖原意的引用、删节和修改。若本公司以外的其他机构（以下简称“该机构”）发送本报告，则由该机构独自为此发送行为负责。通过此途径获得本报告的投资者应自行联系该机构以要求获悉更详细信息或进而交易本报告中提及的证券。本报告不构成本公司向该机构之客户提供的投资建议，本公司、本公司员工或者关联机构亦不为该机构之客户因使用本报告或报告所载内容引起的任何损失承担任何责任。

仅供内部参考，请勿外传！

投资评级说明

项目名称	投资评级	评级说明
公司评级标准	买入	未来 6 个月内相对沪深 300 指数涨幅超过 20%
	增持	未来 6 个月内相对沪深 300 指数涨幅介于 10%—20%之间
	中性	未来 6 个月内相对沪深 300 指数涨幅介于-10%—10%之间
	减持	未来 6 个月内相对沪深 300 指数跌幅超过 10%
行业评级标准	看好	未来 12 个月内相对于沪深 300 指数涨幅超过 10%
	中性	未来 12 个月内相对于沪深 300 指数涨幅介于-10%—10%之间
	看淡	未来 12 个月内相对于沪深 300 指数跌幅超过 10%

渤海证券研究所机构销售团队：

高级销售经理：朱艳君
座机：+86 22 2845 1995
手机：135 0204 0941
邮箱：zhuyanjun@bhzq.com

天津：
天津市南开区水上公园东路宁汇大厦 A 座写字楼
邮政编码：300381
电话：+86 22 2845 1888
传真：+86 22 2845 1615

北京：
北京市西城区西直门外大街甲 143 号凯旋大厦 A 座 2 层
邮政编码：100086
电话：+86 10 6810 4192
传真：+86 10 6810 4192

渤海证券股份有限公司网址：www.bhzq.com