全球AI应用产品梳理：模型能力持续迭代，智能体推动商业化进程

行业研究 · 行业专题

计算机 · 人工智能

投资评级：优于大市（维持评级）

证券分析师：熊莉
021-61761067
xiongli1@guosen.com.cn
S0980519030002

联系人：侯睿
hourui3@guosen.com.cn

摘要

国信证券 GUOSEN SECURITIES

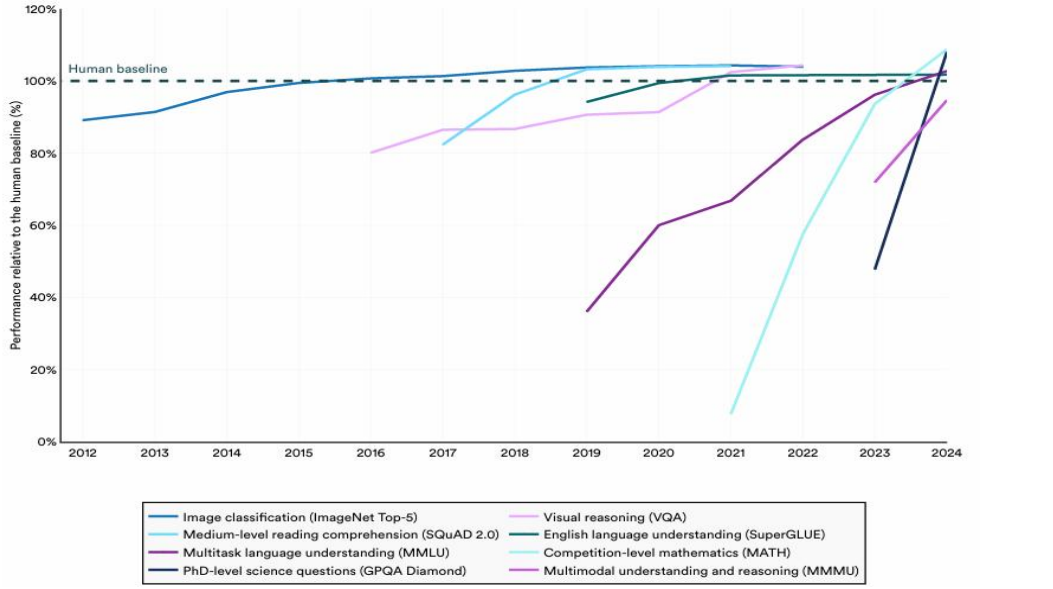
- **模型层：能力迅速提升，开源推动成本降低。**近年来全球AI持续发展，大模型在知识问答、数学、编程等能力上达到新高度，多种任务上表现超过人类水平，在各领域的可用性及准确度快速提升。Scaling Law目前正在从预训练扩展到后训练和推理阶段，随着大模型行业发展逐步成熟，厂商之间开始出现明显的价格竞争与市场份额抢占，大模型的推理成本有了显著下降。同时，得益于Llama 3.1以及DeepSeek R1等高性能开源模型的推出，开源与闭源之间模型差异快速缩小。
- **智能体：技术逐步完善，新产品密集发布。**人工智能体是一种能够感知环境、进行决策和执行动作的智能实体。通过支持添加MCP，Agent可以访问和利用各种外部工具和服务，丰富了Agent的功能范围。Google正式发布A2A协议，整合不同领域Agent的优势，完成跨系统复杂任务。随着应用效果提高，海内外智能体产品密集发布。
- **商业化：用量持续增长，国产模型表现亮眼。**中国与美国顶尖模型之间的差距正在迅速缩小，国产模型依靠开源走出自身生态。当前全球AI模型流量持续上涨，为应用侧发展提供基础。数据显示，各家云厂商推理芯片租赁价格均有不同程度上涨，API调用量亦呈现快速增长趋势。
- **C端应用：借助AI赋能业务，重塑流量入口。**AI应用有望重塑流量入口，各个厂商积极卡位。传统互联网巨头在AI领域具备先发优势，可利用专有数据和用户参与度将AI功能集成到现有的应用当中，在AI应用渗透领域具备先发优势。当前，编程成为人机协同的主要领域，办公类任务AI占比较低。
- **B端应用：开源提升投入意愿，推动企业上云。**开发工具和生态的繁荣大幅降低行业应用门槛，加速产业智能化落地进程。AI技术和解决方案已深入到传媒、医疗、机器人、制造等多个行业，通过创新产品和服务、优化生产流程来推动行业的智能化转型。随着智能化推进，AI应用有望进一步提升企业上云意愿。
- **风险提示：**AI应用落地不及预期、市场需求不及预期、行业竞争加剧、宏观经济波动、新技术研发不及预等。

目录

AI技术快速发展，推动模型能力持续提升

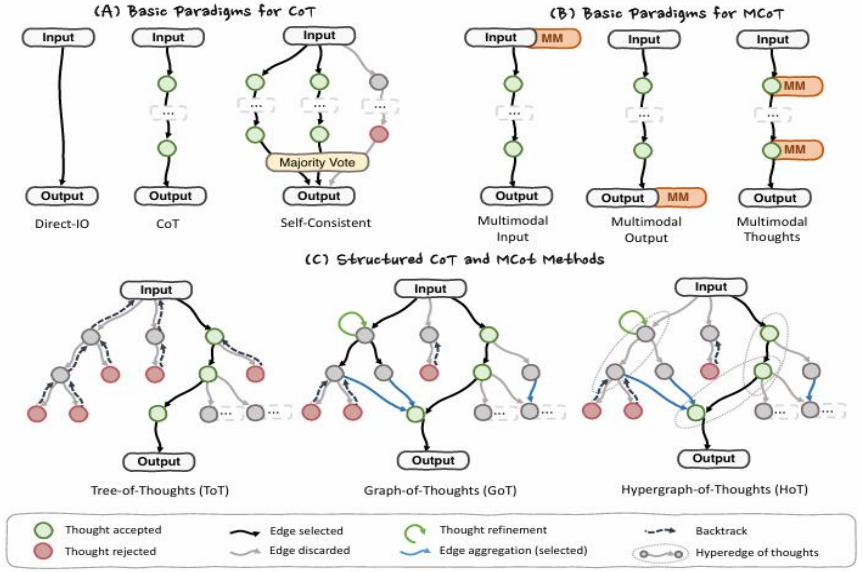
- **近年来全球AI持续发展，大模型在知识问答、数学、编程等能力上达到新高度，多种任务上表现超过人类水平，在各领域的可用性及准确度快速提升。**从模型技术来看：1）当前模型主流架构逐步转向MoE，通过将输入Token分配给不同的专家模型，让模型在处理复杂任务时展现出更强的能力，同时也能有效降低训练、推理所需的资源，DeepSeek-V3、Qwen3、Llama 4等模型均采用MoE架构取得了低成本的高性能表现；2）模型的多模态能力显著增强，通过跨模态对齐、异构数据融合等技术，模型能够处理图像、视频、音频等多种类型的数据，从而丰富了模型的应用场景，GPT-4o、Gemini 2.5 Pro等领先模型均采用多模态技术；3）模型开始采用思维链技术，将复杂问题逐步分解为多个简单步骤，并按照步骤推导最终答案，通过分步推理的方法，模型的回答不仅更加精确可靠，其思考过程也变得清晰可懂。2024年9月，OpenAI发布o1模型，首次将思维链技术运用在底层模型当中，大幅提高了模型在测试中的表现，后续DeepSeek-R1等模型均采用思维链技术，全球模型进入推理时代。除上述方面外，模型量化、超长上下文窗口、多种RAG变体、偏好微调等技术的发展亦共同推动了模型可用性的进步，为AI在垂直领域的应用奠定了基础。

图：AI在多种能力测试中超越人类水平



资料来源：Stanford University-《Artificial Intelligence Index Report 2025》-2025年-P93，国信证券经济研究所整理

图：CoT与MCoT的不同思维链范式



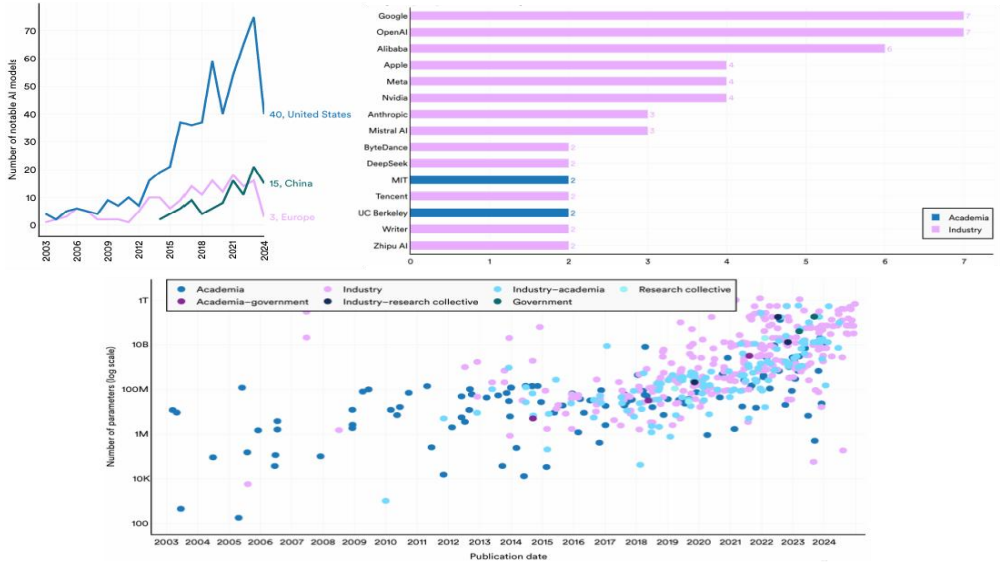
资料来源：Yaoting Wang等-《Multimodal Chain-of-Thought Reasoning: AComprehensive Survey》-arXiv-2025年-P8，国信证券经济研究所整理

模型训练竞赛趋缓，Scaling Law向推理侧迁移

- **据斯坦福大学数据，受训练规模不断增加、AI技术复杂性提升以及开发新模型方法面临更大挑战等因素影响，各地区2024年发布的模型总数同比均有所下降。**美国为2024年发布知名模型最多的地区，数量达40个，较2023年的61个同比下降34.43%。分机构看，2024年贡献知名模型最多的机构分别是OpenAI（7个）、谷歌（7个）和阿里巴巴（4个）。受MoE等新技术推动，2024年模型的参数数量保持快速上升趋势，规模扩大仍是模型性能提升的重要方式。

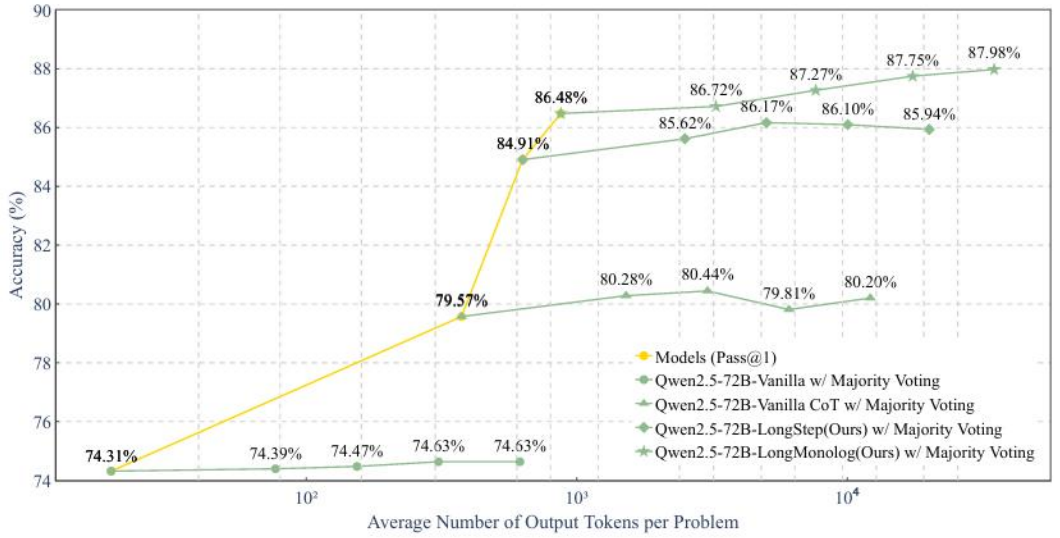
- **Scaling Law目前正在从预训练扩展到后训练和推理阶段，基于强化学习、思维链等技术在后训练和推理阶段投入更多的算力，可以大幅提升大模型的思考能力。**同时，随着强化学习时间和推理思考时间的增长，模型性能也将得到显著提升。据前OpenAI应用研究负责人Lilian Weng数据，s1实验中，通过强制延长思维链推理路径长度，以Token衡量的平均思维时间与下游评估准确率之间展现出明显的正相关关系。据上海交通大学研究表明，通过延长AI的推理时间，仅需500个样本训练，就能让模型在医疗诊断准确率上提升6%-11%，达到专业医生的诊断水准。随着模型推理能力快速提升，当前AI在各领域可用性、准确度不断提高，商业化前景被逐步打开。

图：全球新发模型数量同比下降



资料来源：Stanford University-《Artificial Intelligence Index Report 2025》-2025年-P46，国信证券经济研究所整理

图：模型在医学领域的准确性随思考时间增加而提升



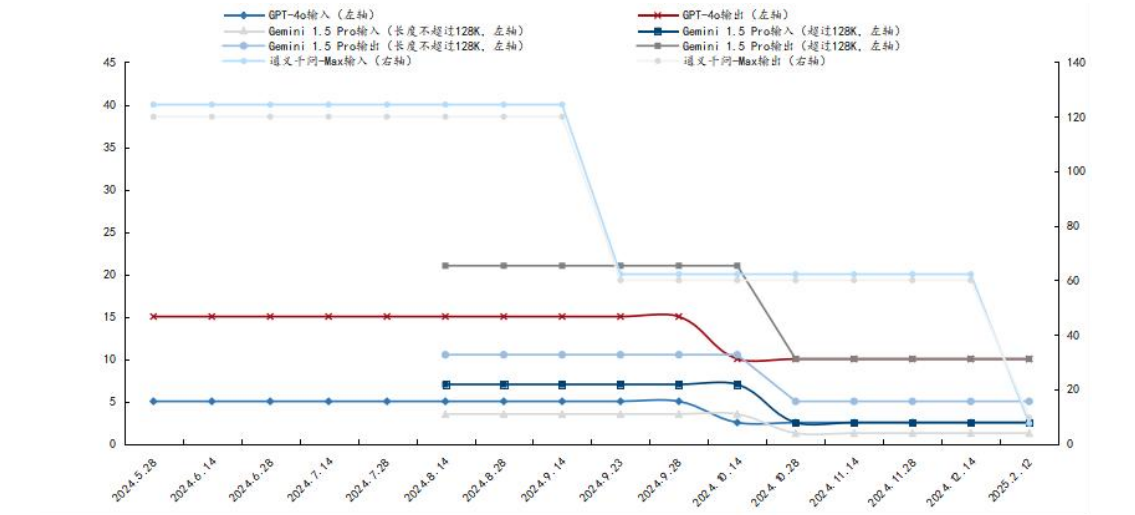
资料来源：Zhongzhen Huang等-《O1 Replication Journey–Part 3: Inference-time Scaling for Medical Reasoning》-arXiv-2025年-P6，国信证券经济研究所整理

模型推理成本显著下滑，利好应用端成本下降

- **随着大模型行业发展逐步成熟，厂商之间开始出现明显的价格竞争与市场份额抢占。**据OpenAI和谷歌官网数据，2024年双方主力模型API调用价格均出现大幅下降，其中GPT-4o输入API调用价格为2.5美元/百万Tokens（下降50%），输出API调用价格为10美元/百万Tokens（下降33%）；谷歌Gemini 1.5 Pro输入API调用价格为2.5美元/百万Tokens（下降64%，超过128k），Gemini 1.5 Pro输出API调用价格为10美元/百万Tokens（下降52%，超过128k）。国内方面，千问、Kimi、腾讯等主力模型价格均有不同程度下降，据千问官网数据，Qwen-Max输入API调用价格在2025年下降至2.5元/百万Tokens（下降88%），输出API调用价格下降至9.6元/百万Tokens（下降84%）大模型API调用价格下降利好AI应用厂商成本下降，进而传导至终端AI应用消费者费用的下降。

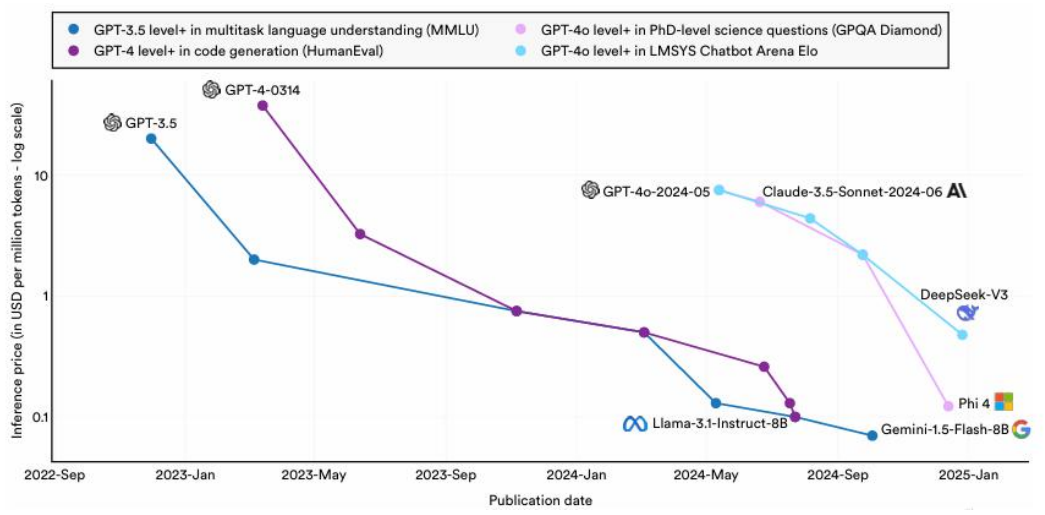
- **据斯坦福大学数据，在保持AI性能不变的前提下，近年来大模型的推理成本有了显著下降。**例如，在流行的MMLU基准测试中，达到GPT-3.5水平（得分64.8）的AI模型推理成本，从2022年11月的每百万Tokens 20美元，大幅下降至2024年10月的仅0.07美元（对应Gemini-1.5-Flash-8B），这意味着在大约1.5年的时间里，推理成本下降了超过280倍。在更具挑战性的基准GPQA上，对于性能评分超过50%的模型，其推理成本从2024年5月的每百万Tokens 15美元，下降到了2024年12月的0.12美元（对应Phi-4）。据Epoch AI的估算，根据推理任务的不同，大模型的推理成本每年都在以9到900倍的速度下降。

图：国内外主力模型API调用价格下降



资料来源：公司官网，国信证券经济研究所整理

图：全球大模型推理成本快速下降



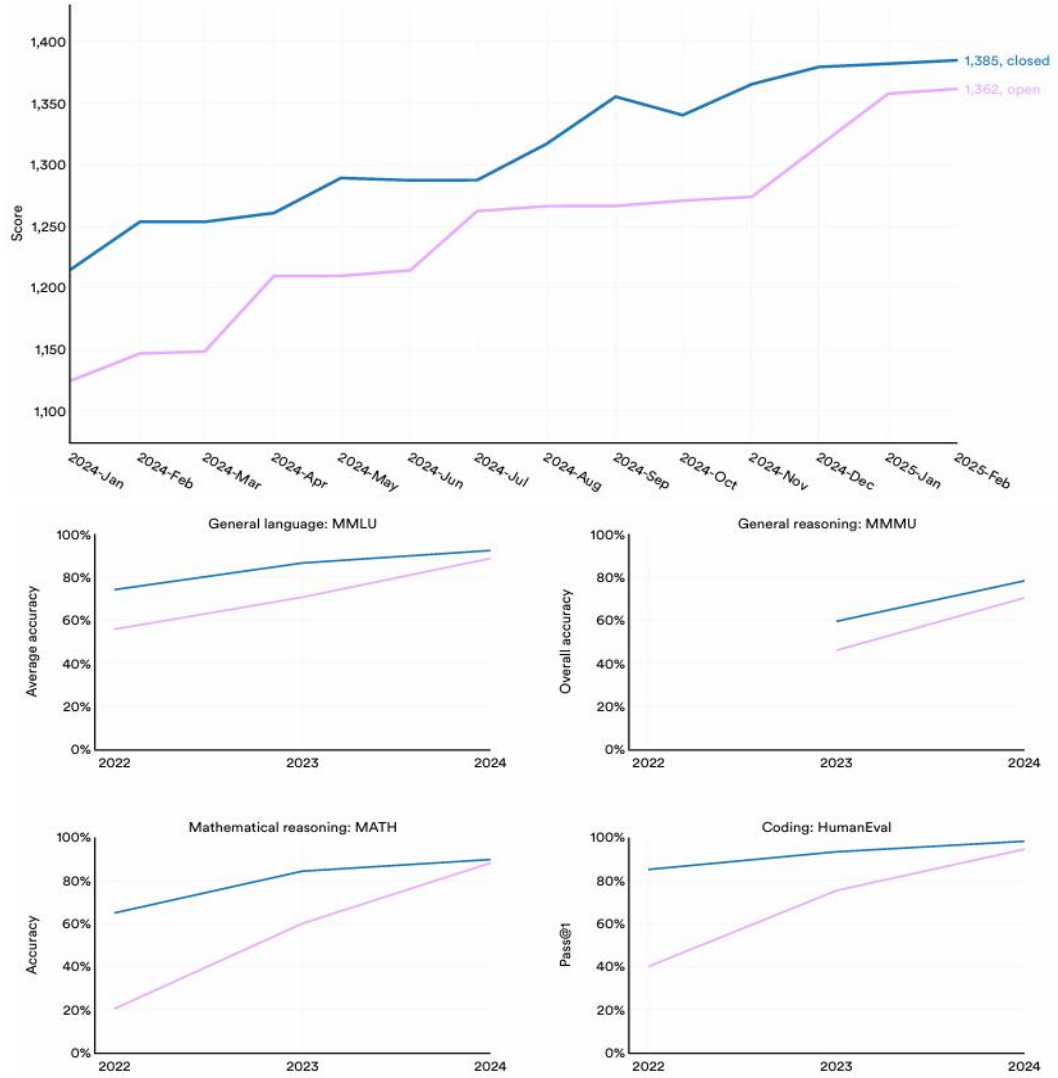
资料来源：Stanford University-《Artificial Intelligence Index Report 2025》-2025年-P64，国信证券经济研究所整理

开源与闭源模型差距缩小，推动AI应用落地

- **得益于Meta发布的Llama 3.1以及DeepSeek V3、R1等高性能开源模型的推出，开源与闭源之间模型差异快速缩小。**据斯坦福大学数据，2023年闭源与开源大模型之间存在明显的性能差距，在Chatbot Arena排行榜中，2024年1月初领先的闭源模型比顶级开源模型高出8.0%，而2025年2月差距缩小至1.7%，类似的趋势也出现在其他问答类基准测试中。2023年闭源模型几乎在所有主要基准测试上优于开源模型，但到2024年这种差距显著缩小，例如，2023年底闭源模型在MMLU基准上领先开源模型15.9个百分点，而到2024年底这一差异缩小至仅0.1个百分点。

- 开源模型允许开发者直接访问、修改和优化模型代码，降低了AI技术的使用门槛，用户可根据自身需求进行定制化开发，使模型更容易适配金融、医疗等垂直行业需求，加速大模型应用的普及。同时，用户无需支付闭源模型调用费用，使用大模型的成本显著降低，刺激AI应用在付费意愿较低的用户中渗透。随着开源模型与闭源模型之间的差距逐步缩小，下游企业可直接在企业中接入相关模型，并获得与顶尖闭源模型等同的应用表现，极大推动AI在各个垂类领域的应用。例如，通过微调Llama模型，AT&T在客户服务搜索响应上取得了近33%的提升；埃森哲基于Llama 3.1构建了用于ESG报告的定制大模型；北京中医药大学深圳医院部署DeepSeek赋能医院运营管理等。

图：开源与闭源模型之间差距快速缩小



资料来源：Stanford University-《Artificial Intelligence Index Report 2025》-2025年-P95，国信证券经济研究所整理

目录

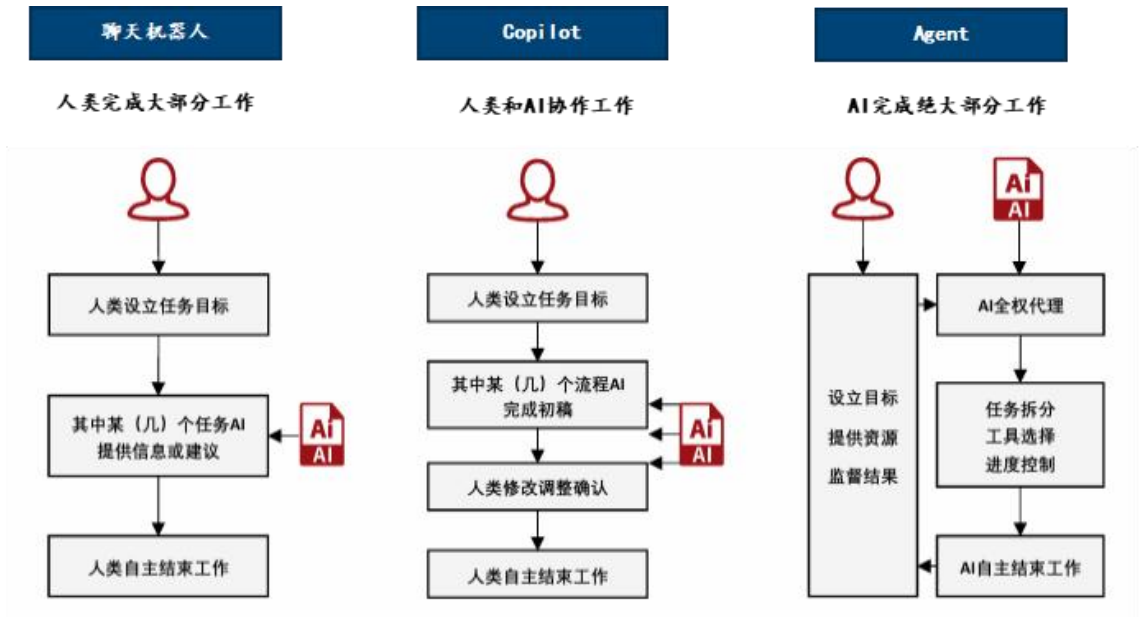
AI应用快速迭代，人机协同从Copilot转向Agent

国信证券 GUOSEN SECURITIES

- **AI Agent（人工智能体）是一种能够感知环境、进行决策和执行动作的智能实体。**不同于传统的人工智能，AI Agent具备通过独立思考、调用工具去逐步完成给定目标的能力。AI Agent和传统大模型的区别在于，大模型与人类之间的交互是基于prompt实现的，用户prompt是否清晰明确会影响大模型回答的效果，而AI Agent的工作仅需给定一个目标，它就能够针对目标独立思考并做出行动。

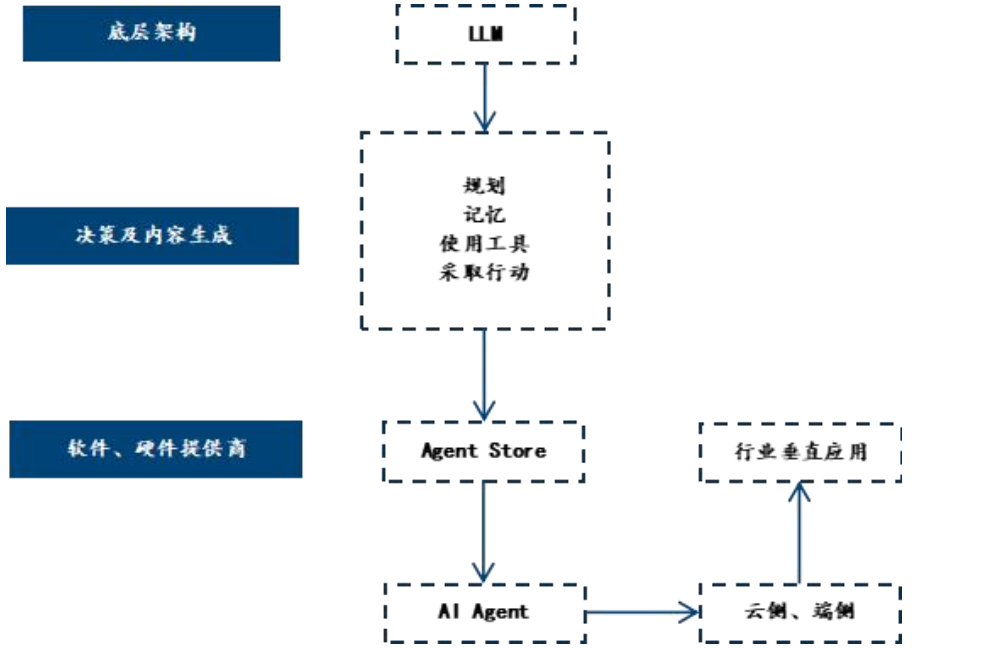
- **基于大模型的Agent不仅可以让每个人都有增强能力的专属智能助理，还将改变人类与AI协同的模式。**随着大模型的发展，人类与模型的协同方式从最初的聊天机器人转变为Copilot，并逐步向Agent探索。Agent的落地将给AI应用带来颠覆性变化，打开AI在垂直行业渗透的入口。随着自然语言处理、机器学习和生成式AI的进步，AI Agent的多功能性和部署量将急剧增长。

图：人类与AI交互方式转变



资料来源：头豹研究院，Frost & Sullivan，国信证券经济研究所整理

图：AI Agent打开垂直行业应用入口

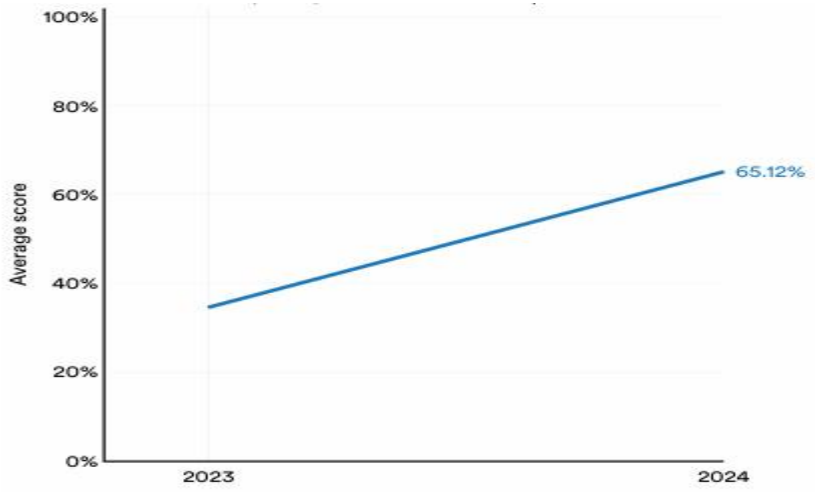


资料来源：头豹研究院，Frost & Sullivan，国信证券经济研究所整理

模型Agent能力快速提升，测试分数不断刷新

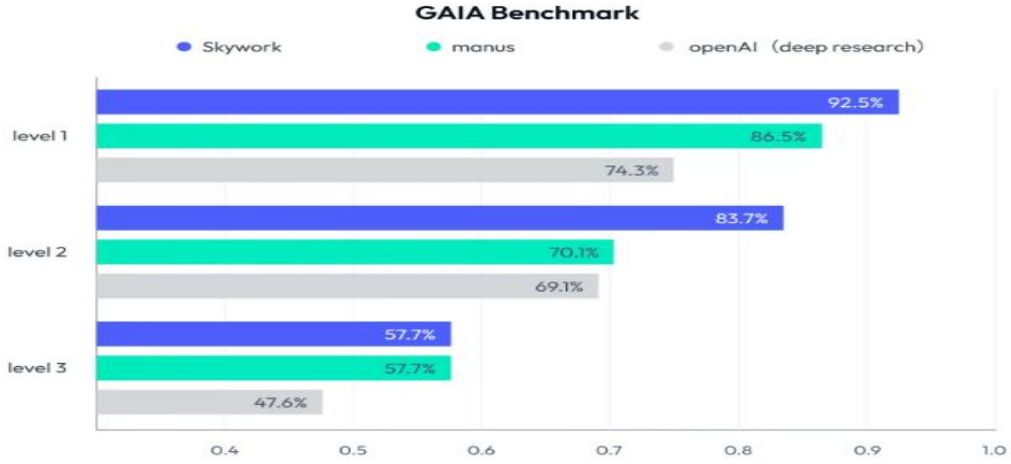
- GAIA是由Meta于2024年5月推出的一个面向通用AI助手的基准测试，包含466道问题，旨在评估AI系统执行广泛任务的能力，包括推理、多模态处理、网页浏览和工具使用等。与那些简单、类似考试风格的问题不同，GAIA使用复杂、多步骤的问题来挑战AI模型，这些问题可能需要搜索开放网络、解读多模态输入，并在复杂情境中进行推理。GAIA可以根据解决问题所需的步骤数量和所需的不同工具数量分为三个难度级别：
- 1）Level 1：问题通常不需要工具，或最多使用一个工具，不超过5步；
- 2）Level 2：问题通常涉及更多步骤，大约在5到10步之间，且需要结合不同的工具；
- 3）Level 3：问题是为接近完美的通用助手设计的，需要执行任意长度的操作序列，使用任意数量的工具，并访问一般世界。
- 当研究人员发布GAIA时，他们发现现有的大语言模型（LLM）在表现上远远落后于人类。例如，使用插件的GPT-4仅能正确回答15%的问题，而人类受访者的正确率则高达92%。当前模型在GAIA上的表现迅速提升，在2024年，表现最佳的系统得分达到了65.1%，相比2023年记录到的最高分提高了大约30个百分点。2025年5月，昆仑万维的天工智能体登顶GAIA，刷新SOTA得分，平均得分来到78.0%。

图：GAIA得分快速提高



资料来源：Stanford University-《Artificial Intelligence Index Report 2025》-2025年-P147，国信证券经济研究所整理

图：天工智能体刷新GAIA成绩单



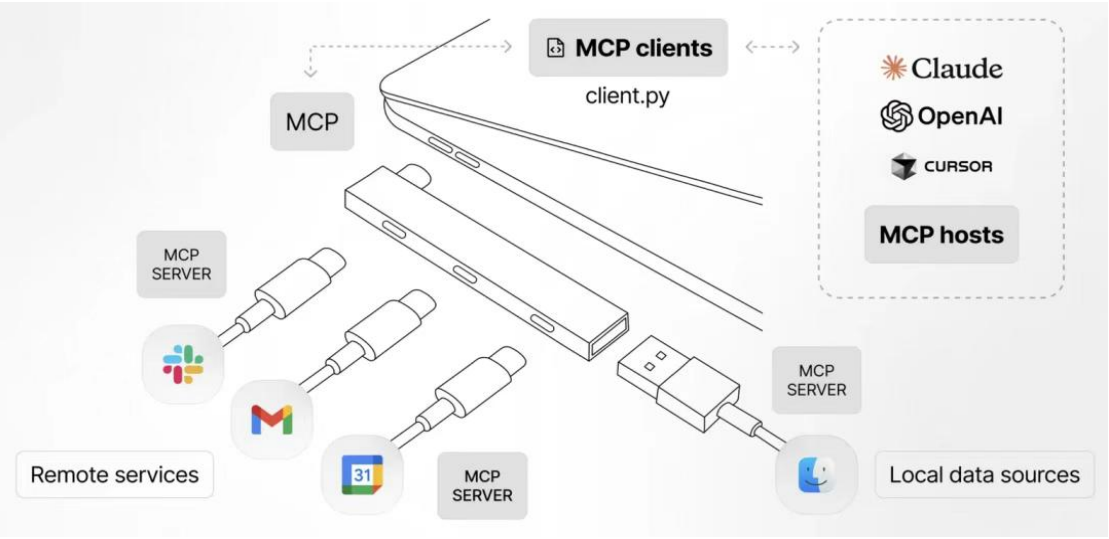
资料来源：公司官网，国信证券经济研究所整理

MCP扩展AI能力边界，推动Agent加速落地

- **通过支持添加MCP，AI Agent能力边界被进一步扩展。**MCP（Model Context Protocol）是由Anthropic提出的开放标准，旨在为AI模型与外部工具之间建立安全、双向的连接。在MCP出现之前，AI要集成工具需要针对每个工具进行定制开发，缺乏统一标准，集成效率低。而MCP协议提供了可插拔、可扩展的框架，允许AI无缝对接数据源、文件系统、开发工具、Web浏览器等外部系统。通过集成MCP扩展，Agent可以访问和利用各种外部工具和服务，丰富了Agent的功能范围，使其能够执行更复杂的任务。同时，MCP提供了标准化的接口，AI可以根据具体需求快速接入新的工具或数据源，对于Agent的可用性以及生态构建均有显著推动作用。

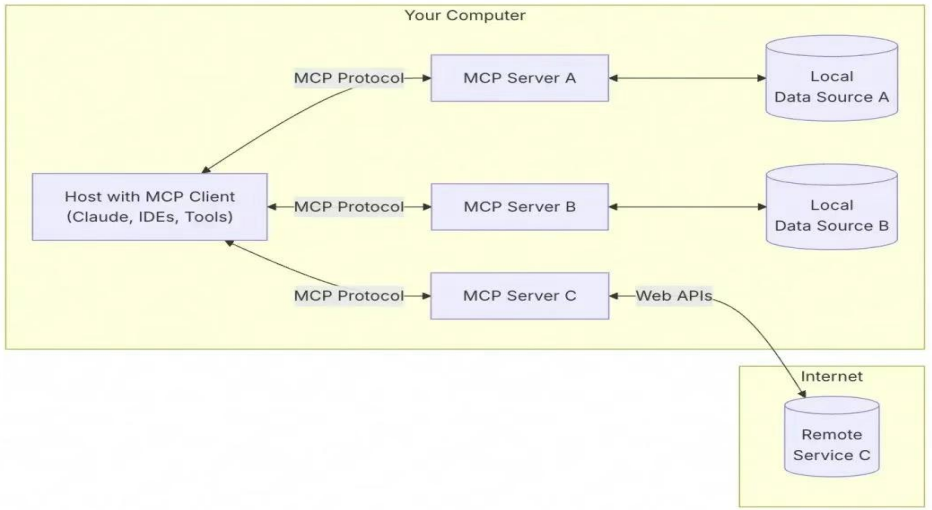
- **海内外内大厂纷纷布局MCP，相关生态迅速丰富。**海外方面，微软宣布在Copilot Studio、GitHub Copilot等产品中支持MCP，并将在Windows 11中集成MCP；谷歌支持Gemini AI模型使用MCP协议；AWS上线MCP Servers，每个服务器专注于特定领域，协同提供全面解决方案。国内方面，百度智能云率先宣布千帆大模型平台接入集成MCP，支持通过千帆AppBuilder SDK开发的组件无缝转化为MCP Server模式；阿里云百炼平台推出全生命周期MCP服务，覆盖高德地图、GitHub自动化等15类场景，支持一键开通并集成至智能体；腾讯云大模型知识引擎已接入MCP，用户可调用平台精选或自定义MCP插件搭建应用；火山引擎发布大模型生态广场MCP Servers，实现工具调用、模型推理到应用部署的全链路开发闭环。MCP有望成为AI时代的HTTP协议，可大幅提效AI应用开发、加速生态扩展。

图：MCP提供了统一连接标准



资料来源：阿里云官网，国信证券经济研究所整理

图：MCP遵循客户端-服务器架构



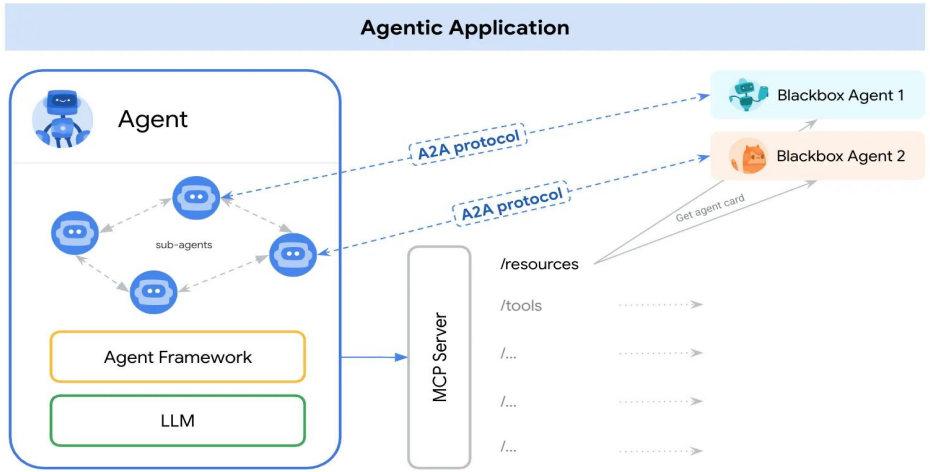
资料来源：阿里云官网，国信证券经济研究所整理

谷歌发布A2A协议，打通AI落地复杂应用场景

- **A2A与MCP互补，加速Agent生态完善。**2025年4月，Google正式发布Agent2Agent Protocol（简称A2A），为用于链接不同封闭Agent，并实现其相互操作的开放协议，该协议为不同类型的智能体之间搭建了高效沟通与协作的桥梁，无论是独立Agent与独立Agent、独立Agent与企业Agent，亦或是企业Agent与企业Agent，都能借助该协议实现通信交互和事务协作。A2A协议与MCP互补，A2A负责解决Agent间的通信问题，MCP解决Agent与工具间的通信问题，有望提升Agent在下游领域的应用效果，推动Agent生态系统的完善与发展。

- **A2A获得多个科技巨头支持，推动AI应用向复杂工作流落地。**随着Agent应用的逐步落地，单一Agent难以独立完成多领域任务（如同时处理数据分析、文档生成等），需依赖团队协作，而不同厂商的Agent因技术栈差异无法直接协作，形成信息孤岛，从而阻碍Agent应用落地。A2A协议可通过任务自动分配与结果同步，减少人工干预，同时整合不同领域Agent的优势，完成跨系统复杂任务。A2A协议构建在HTTP、Server-Sent Events(SSE)、JSON-RPC等常用标准上，企业无需大规模改造自身IT技术栈，就能平滑接入多代理环境。在用户发起任务后，客户端智能体通过Agent Card定位目标智能体，通过代理间相互发送消息，包括上下文信息、用户指令、执行结果等形成协同网络，依次或并行地处理不同环节。A2A协议当前已得到了50多家谷歌技术合作伙伴的支持和贡献，包括Atlassian、Salesforce、SAP、ServiceNow等，为依赖多源数据和需要嵌入复杂工作流的AI应用提供了走向大规模落地的生态支撑。

图：A2A与MCP协议互为补充



资料来源：阿里云官网，国信证券经济研究所整理

图：A2A已获得50多家科技公司支持

资料来源：谷歌官网，国信证券经济研究所整理

海内外智能体快速发展，测评成绩不断刷新

- **全球通用领域Agent快速发展，应用效果快速提高。**海外方面，2025年1月，OpenAI上线了其首个AI Agent Operator，能够与电脑交互，完成浏览网页、填写表格、预定餐厅等相关任务。2月，OpenAI发布Deep Research，由o3模型提供支持，能够帮助用户进行信息查询与分析，输出综合报告。5月，Anthropic发布Claude 4，可自主编程数小时并在推理过程中使用工具。同时发布编程Agent Claude Code，通过GitHub Actions支持后台任务，与VS Code和JetBrains进行了原生集成，可直接在文件中显示编辑内容，实现无缝结对编程。国内方面，3月，Monica正式对外发布通用型AI Agent产品Manus，提供多种处理现实世界任务的案例，包括旅行规划、股票分析等，测评成绩超越Deep Research。4月，MainFunc发布Genspark，采用整合多AI模型的混合代理（MoA）系统，包含80多个工具集和10多个高级数据集，可协调多个AI工具高效执行各项任务。字节跳动发布扣子空间，除通用Agent外还提供华泰A股助手等专家Agent，支持飞书多维表格、高德地图等MCP。5月，昆仑万维发布天工智能体，采用Deep Research技术，能够生成文档、PPT、播客和音视频多模态内容，提供5个专家级Agents和1个通用Agent，接入数十个MCP，刷新GAIA测评新高。

图：Agent领域新品密集发布

时间	产品/技术	参与者	意义
2024年10月	微软Dynamics 365集成自主AI Agent	微软	增强企业管理软件的智能化水平，实现多业务领域的自动化，提高企业运营效率
2024年10月	中国移动灵犀消息智能体	中国移动	推动AI在通信领域的应用，提升用户通话等场景中的体验，拓展智能体在通信场景中的应用边界
2024年12月	谷歌Gemini 2.0驱动的Project Astra（谷歌AI助手）	谷歌	拓展AI Agent在多领域的应用，推动多模态AI Agent的发展，为用户提供更全面、智能的服务
2024年12月	中国电信星辰智能体	中国电信	通过自主规划和工作流两大模式解决大模型落地痛点，快速搭建“会展助手”智能体，提升工作效率
2024年12月	字节跳动MarsCode	字节跳动	为国内开发者提供一款功能强大的AI代码编程工具，提高编程效率，推动国产AI编程工具的发展
2025年1月	Claude 3.5 Sonnet升级版	Anthropic	提升AI在代码开发和计算机操作模拟方面的能力，为开发者提供更智能的辅助工具
2025年1月	AI Agent Operator	OpenAI	拓展AI Agent的功能边界，为用户提供更强大的自动化任务执行和知识生成能力，推动AI Agent在多领域的应用，代表OpenAI拉开L3级Agent时代序幕
2025年1月	阿里云推出通义千问Qwen2.5-Max模型	阿里巴巴	提升国内大模型在多模态交互和复杂任务处理方面的能力，为AI Agent的开发提供更强大的基础模型支持
2025年1月	拓尔思拓天大模型AI Agent工具链	拓尔思	降低AI Agent的创建门槛，推动AI在多个领域的应用落地，具备任务规划、流程编辑与自动执行功能
2025年2月	GitHub Copilot Agent模式	GitHub	提升AI在代码开发中的自主性和智能性，推动软件开发模式的变革，提高代码开发的效率和质量
2025年2月	DeepSeek-R1	幻方量化	降低AI Agent开发门槛，推动开源生态与行业应用的结合，为国内AI Agent的发展提供新的技术支撑和开源资源
2025年2月	Deep Research	OpenAI	帮助用户进行深入、复杂的信息查询与分析，以研究分析师的水平创建综合报告
2025年3月	Manus	Monica.im	工具链整合能力的规模化跃升迎来中国AI Agent重大突破时刻，推动AI Agent从对话智能升级为生产力操作系统
2025年3月	AutoGLM 沉思	智谱	能够模拟人类的思维过程，完成从数据检索、分析到生成报告，核心链路的技术与模型于4月全面开源，进一步推动生态发展
2025年4月	Genspark	MainFunc	整合多AI模型的混合代理（MoA）系统，包含了80多个工具集和10多个高级数据集，在GAIA Benchmark中表现超越Manus、OpenAI Deep Research等产品
2025年4月	扣子空间	字节跳动	从回答问题，到解决问题全线打通，拥有专家Agent生态并首创探索/规划双模式，MCP扩展集成，拓展Agent能力边界
2025年4月	阿里云发布通义千问Qwen3模型	阿里巴巴	总参数量235B，专家模型数量128个，性能测评全面超越R1、OpenAI-o1等顶尖模型，部署成本仅为满血版R1的25-35%，在专门评估模型Agent能力的BFCL评测中刷新了榜单记录，同时原生支持MCP协议
2025年5月	Claude 4及Claude Code	Anthropic	Claude 4擅长处理复杂的编程问题，可以自主编程数小时，并可在推理过程中使用工具。编程Agent Claude Code通过GitHub Actions支持后台任务，直接在文件中显示编辑内容，实现无缝结对编程
2025年5月	天工超级智能体	昆仑万维	全球首个开源的deep research Agent框架，拥有5个专家AI Agent和1个通用AI Agent，支持一站式生成docs、slides、sheets等多种模态内容，所有信息都可以溯源
2025年5月	DeepSeek-R1模型升级	幻方量化	在数学、编程与通用逻辑等多个测试中超越Qwen3，整体表现接近OpenAI-o3，复杂推理任务表现显著提升，蒸馏思维链后训练的DeepSeek-R1-0528-Qwen3-8B在数学测试中与Qwen3-235B相当
2025年6月	AI编程智能体Gemini CLI开源	谷歌	提供包括代码编写、问题调试、项目管理、文档查询以及代码解释等功能，同时具备通用Agent能力，可利用Veo 3模型生成视频、连接MCP访问外部数据库等
2025年7月	Kimi K2模型开源	月之暗面	具备更强代码能力、擅长通用Agent任务的MoE架构模型，总参数1T，激活参数32B，可接入owl、Cline、RooCode等Agent/Coding框架，完成复杂任务或自动化编码
2025年7月	ChatGPT agent	OpenAI	融合Operator网站交互能力、Deep Research整合信息技巧以及ChatGPT智能对话优势，能够浏览网页、筛选结果，在需要时提醒安全登录、运行代码、进行分析，还能直出PPT和Excel汇总发现结果

资料来源：公司官网，国信证券经济研究所整理

目录

中国走出自身发展路径，本土AI形成生态闭环

- **中美作为全球AI领域的领导者，目前已走出不同的发展路径：不同于美国算力至上的路径，中国AI发展更侧重效率，以显著少于预期的硬件配置实现了先进的AI能力。**例如，DeepSeek-V3仅用557.6万美元的训练成本便取得与GPT-4o等顶尖闭源模型相当的成绩；仅需4张H20即可本地部署235B的Qwen3旗舰版MoE模型，部署成本仅为DeepSeek-R1的35%。

- **当前中国已构成了政府推动+本土产业链+人才储备+数据红利共同构成AI生态闭环，成为中国AI持续发展的核心动能。**政府方面，早在2017年我国便发布了《新一代人工智能发展规划》，并在后续发布了《人工智能人才培养行动计划（2024-2026年）》等文件，为AI产业的基础理论研究、产业应用落地等方面提供了政策支撑。同时，我国亦高度重视对AI公司的财政支持，据财政部数据，2025年中央预算将拨款3981.2亿元用于科学技术，同比+10%，将重点推进半导体、人工智能等领域发展。人才储备方面，当前中国已成为全球最大的AI人才聚集地。据MacroPolo数据，2022年顶级AI研究人员中有28%在中国工作，较2019年的11%大幅提升。据斯坦福大学数据，截至2023年，中国AI领域论文发表数量位居全球第一，占比达23.2%，相关论文引用量占全球所有AI论文引用的22.6%。专利方面，据R&D World数据，全球AI授权专利数量自2010年以来已增长超30倍，中国在全球AI专利申请中占据主导地位，截至2024年总专利数达12945。

图：中国AI领域论文发表数量占全球23.2%

资料来源：Stanford University-《Artificial Intelligence Index Report 2025》-2025年-P34，国信证券经济研究所整理

图：中国AI专利数量在全球占主导地位

排名	国家	总AI专利数（2024年）	公司数量	每家公司专利数	主要司法管辖区
1	中国	12,945	523	24.8	CN, US, EP
2	美国	8,609	257	33.5	US, EP, CN, JP
3	韩国	1,537	43	35.7	KR, US, CN, EP
4	日本	1,537	41	37.5	JP, US, CN, EP
5	德国	784	18	43.6	EP, US, CN
6	英国	369	11	33.5	GB, EP, US
7	荷兰	249	7	35.6	EP, US, CN
8	瑞典	243	4	60.8	EP, US
9	芬兰	180	1	180	EP, US
10	台湾	156	11	14.2	TW, US, CN

资料来源：Morgan Stanley-《China - AI: The Sleeping Giant Awakens》-2025年-P24，国信证券经济研究所整理

芯片产业本土化加速，数据底座推动AI发展

- **美国不断加码AI芯片的出口管制，中国形成国产芯片的完整链条。**当前国内已形成覆盖从材料、设备到设计、制造及终端应用的完整链条，材料领域沪硅产业已实现12英寸硅片量产、先进制程领域中芯国际实现14nm制程芯片量产、芯片设计领域昇腾910系列芯片算力达256TFLOPS（FP16），光刻机、封测等领域国产替代稳步推进，本土芯片占比逐步提升。据IDC数据，2024年中国AI芯片出货量约270万颗，其中中国本土AI芯片品牌出货量超82万颗，占比达30%，据TrendForce预期2025年国产芯片占比将提升至40%。同时，国产芯片厂商积极围绕DeepSeek、Qwen等模型打造适配本土芯片的软件栈、工具链等生态组件，生态格局进一步完善。

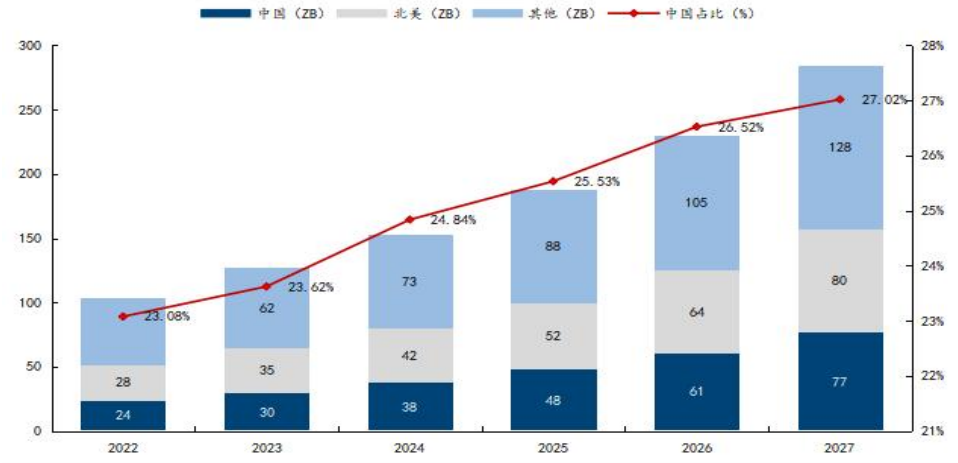
- **随着模型参数数量的增加，数据成为AI发展的核心资源。**2020年，GPT-3在约3740亿个Tokens上进行训练，到2024年Llama 3.3则在约15万亿个Tokens上进行训练，数据集规模呈指数增长。据Epoch AI的统计，大模型训练数据集的规模约每八个月就会翻倍增长，数据成为制约AI发展的关键。我国已成为全球数据量增长最快的国家，据工信部统计，截至2024年我国移动互联网用户达15.7亿户。我国已建成全球规模最大的5G网络，5G用户数占全球的52%，在5G的助力下，智慧城市、远程医疗等新型应用场景得以快速落地，推动数据爆发式增长。我国高度重视数据资产，在2023年10月揭牌国家数据局，负责协调数据相关政策、推动数据基础设施建设，为AI等产业发展提供基础。据IDC预期，中国每年产生的数据量将从2022年的24ZB增长至2027年的77ZB，CAGR达26%。同时，中国拥有字节跳动和快手两大视频平台巨头，每天约生成8000万条新视频，总视频长度已超过YouTube，为多模态模型迭代提供大量数据支撑。

图：中国自研芯片快速追赶

公司	介绍	产品	制程	算力	应用场景
海光信息	成立于2014年，最初通过与AMD成立合资企业获取架构授权。受美国出口管制影响，AMD后续IP许可被叫停，公司已研发出可广泛应用于服务器、工作站的高端处理器产品	深算三号	7nm	392 TFLOPS（INT8）	大数据处理、AI、商业计算等
寒武纪	2016年自中国科学院计算技术研究所孵化，已形成全面覆盖云端、边缘端和终端场景的系列化智能芯片产品布局	MLU220	16nm	128 TOPS(INT8)	图形计算，智能电网、智能制造等边缘计算场景
		MLU370	7nm	256 TOPS（INT8）	AI训练、推理或混合型人工智能计算加速
华为	发布昇腾310/910后，已形成基于达芬奇架构的Atlas板卡、服务器与集群产品族，覆盖端、边、云全场景	昇腾910	7nm N+2	640 TOPS（INT8）	AI训练、数据中心、云计算等
阿里平头哥	成立于2018年，为阿里巴巴的芯片设计主体，业务横跨数据中心、AI推理与RISC- V IP授权，依托玄铁RISC- V系列处理器在IoT生态保持领先	含光800	12nm	825 TOPS（INT8）	AI训练、推理、数据中心等
百度昆仑芯	前身为百度智能芯片及架构部，2021年完成拆分独立，依托芯片+集群+生态策略，已迭代至第三代产品并在智算中心实现规模化部署	昆仑芯2代	7nm	256 TFLOPS	数据中心、AI推理、训练等
壁仞科技	成立于2019年，定位高性能通用GPU供应商，目前壁砺GPU及软件平台已在多地智算中心落地	BR100	7nm	1024 TOPS（INT8）	AI训练、推理等
摩尔线程	成立于2020年，由前NVIDIA、AMD技术团队创办，产品线覆盖数据中心AI加速、云游戏到桌面显卡，并提供物理引擎、数字人等完整软件栈	S80	7nm	14.4 TFLOPS	图形计算等
沐曦	成立于2020年，致力于为异构计算提供全栈GPU芯片及解决方案	MXN100	7nm	160 TOPS（INT8）	AI推理、智算中心等
天数智芯	成立于2015年，专注于开发用于人工智能领域的通用GPU	智铠100	7nm	384 TOPS（INT8）	AI训练、推理，金融、医疗、教育智算中心等

资料来源：公司官网，国信证券经济研究所整理

图：国产AI芯片性能逐步追赶



资料来源：IDC，国信证券经济研究所整理

中美模型差距缩小，依靠开源模型走出自身生态

国信证券 GUOSEN SECURITIES

- **美国长期在AI研究和模型开发方面占据主导地位，最新数据显示中国与美国顶尖模型之间的差距正在迅速缩小。**据斯坦福大学数据，2023年美国模型在性能上显著超越中国模型，在LMSYS聊天机器人竞技场中，2024年1月表现最好的美国模型比最佳的中国模型高出9.26%。到了2025年2月，差距缩小到仅1.70%。2023年底，双方在MMLU、MMMU、MATH和HumanEval等基准测试上的性能差距分别为17.5、13.5、24.3和31.6个百分点，到2024年末差异显著缩小至仅0.3、8.1、1.6和3.7个百分点。

- **中国在开源AI模型方面已领先全球，Qwen3登顶榜首。**据斯坦福大学数据，在SOTA模型中，中国的推理模型有1/2是开源的，非推理模型中有1/3是开源的，而美国的推理模型仅有1/8、非推理模型仅有1/7为开源。在HuggingFace的开源AI模型排行榜上，中国占据了前十名中的40%席位，其中Qwen2.5位列全球第一。在LMArena模型竞技场当中，Kimi-K2成为评分最高的开源模型。

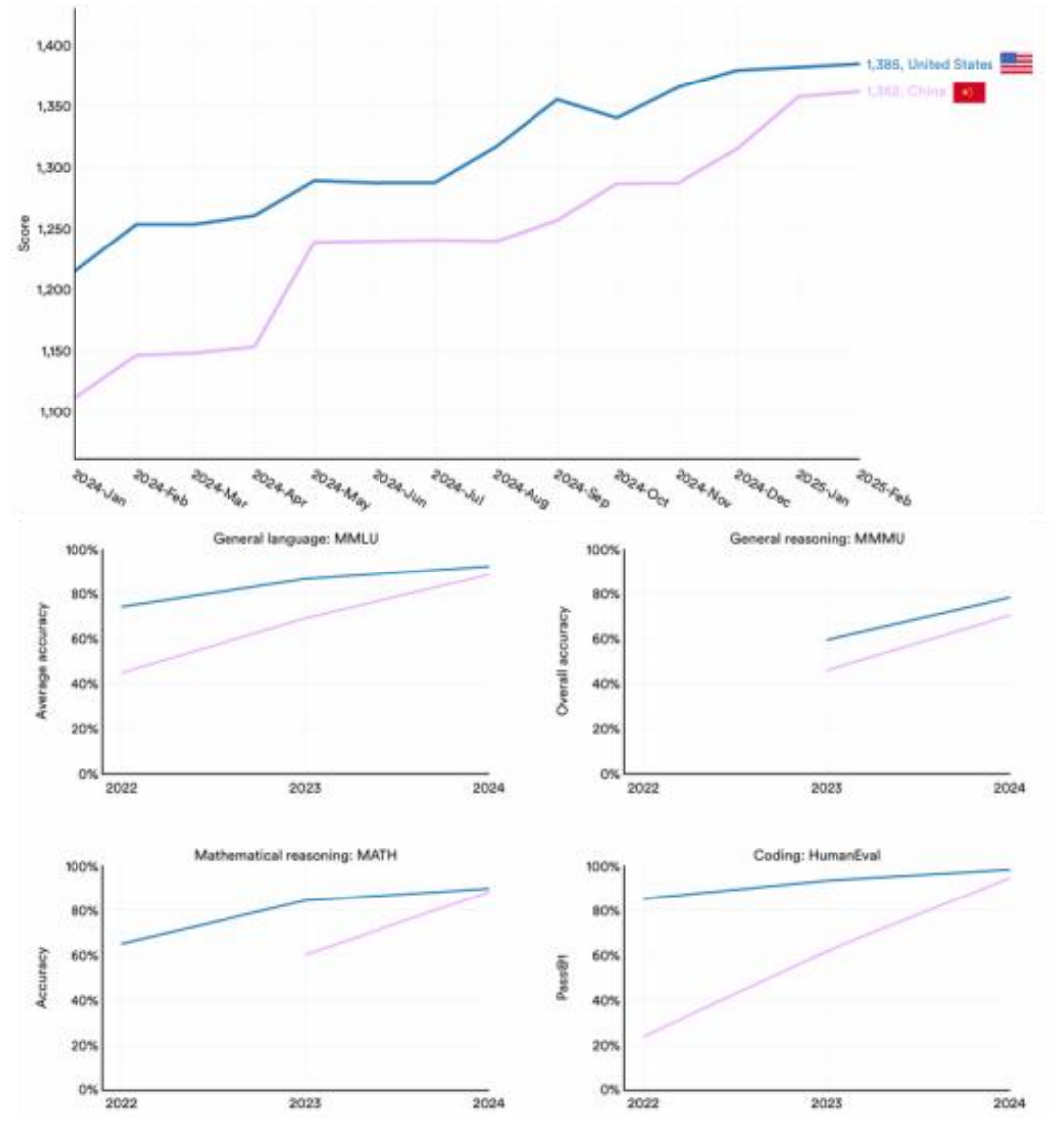
图：中国开源模型登顶多项榜单

Model	217 / 217	Overall	Hard Prompts	Coding	Math	Creative Writing	Instruction Following	Longer Query	Multi-Turn
gemini-2.5-pro		1	1	1	1	1	1	1	1
o3-2025-04-16		2	2	1	1	4	2	6	4
chatgpt-4o-latest-2…		3	2	1	7	2	2	1	1
grok-4-0709		3	2	2	1	2	2	2	4
gpt-4.5-preview-202…		3	4	2	6	1	2	2	1
kimi-k2-0711-preview		5	2	2	7	7	7	5	1

Rank	Type	Model	Average	IFEval	BBH	MATH	GPQA	MUSR	MMLU…
6		Qwen/Qwen2.5-72B-Instruct	47.98 %	86.38 %	61.87 %	59.82 %	16.67 %	11.74 %	51.40 %
22		Qwen/Qwen2.5-32B-Instruct	46.60 %	83.46 %	56.49 %	62.54 %	11.74 %	13.50 %	51.85 %
23		mistralai/Mistral-Large-Instruct-2411	46.52 %	84.01 %	52.74 %	49.55 %	24.94 %	17.22 %	50.69 %
40		meta-llama/Llama-3.3-70B-Instruct	44.85 %	89.98 %	56.56 %	48.34 %	10.51 %	15.57 %	48.13 %

资料来源：HuggingFace，LMArena，国信证券经济研究所整理

图：中国模型能力正在逐步追赶美国

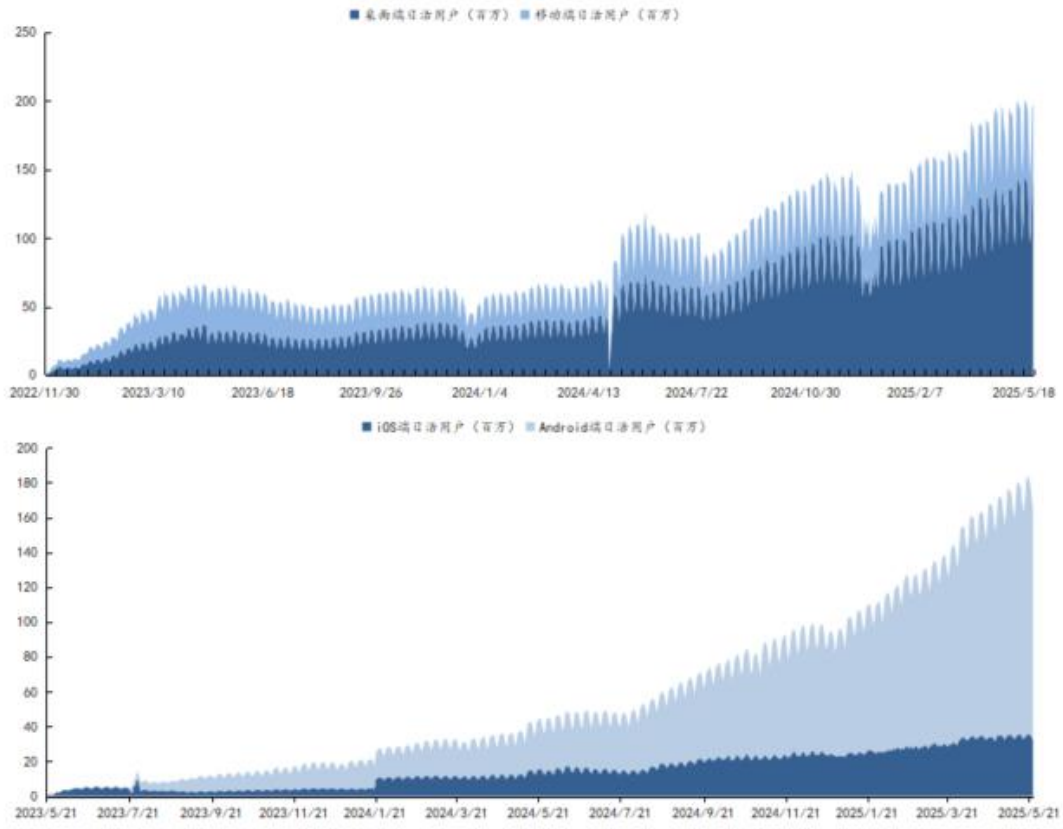


资料来源：Stanford University，国信证券经济研究所整理

AI日活快速增长，为商业化提供基础

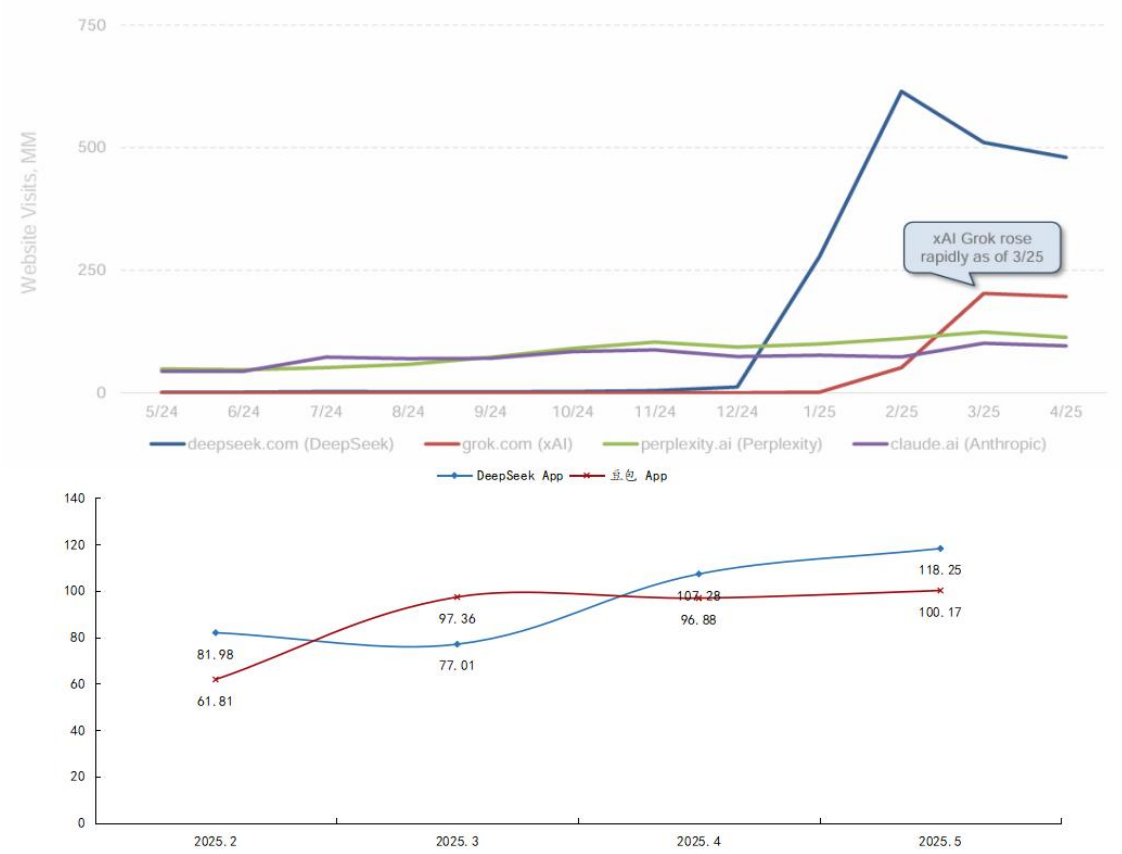
- **全球AI模型流量均持续上涨，为应用侧发展提供基础。**海外侧，ChatGPT日活流量呈上升趋势，2025年5月19日网页端当日日活达到2.02亿，创下历史新高，同比增长135.7%，当前网页端日活用户维持在2亿左右；移动端，5月20日单日日活达到1.84亿，同比增长328.6%。2025年1-3月，Grok日活人数增长迅速，Perplexity、Claude日活呈现稳定增长态势。国内方面，AI日活流量也呈现上升趋，2024年12月起，DeepSeek日活快速增长，访问量在4月份达到约5亿次。移动端，日活自1月份持续增长，据AI产品榜数据，DeepSeek五月份MAU达1亿，环比增长3.39%。豆包日活持续提升，五月份MAU达1.18亿，环比增长10.23%。

图：ChatGPT活跃用户统计



资料来源：SimilarWeb，国信证券经济研究所整理

图：其他AI日活数据



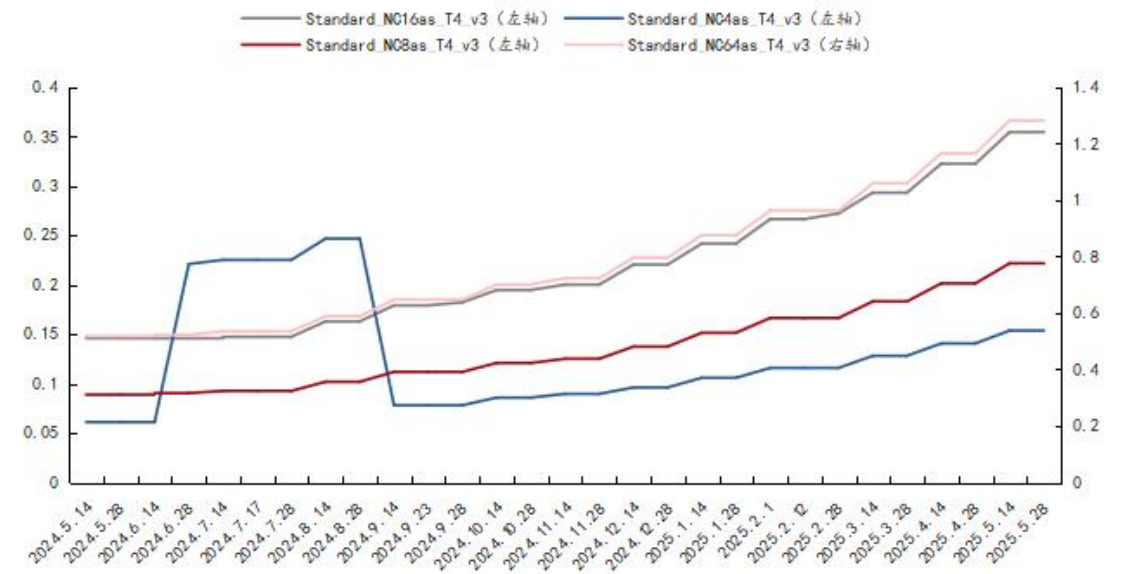
资料来源：BOND-《Trends-Artificial Intelligence》-2025年-317，AI产品榜，国信证券经济研究所整理

推理需求逐步提升，侧芯片价格上升

- **当前全球AI应用快速发展，各行业Tokens调用量迅速提升。**根据微软FY25Q3财报电话会议数据，公司平台季度处理了超100万亿个Tokens，同比大幅增长5倍，3月处理创新高的50万亿个Tokens，已有超过10000家组织使用了新的智能体服务来构建、部署和扩展智能体。据英伟达FY26Q1财报数据，推理需求正在迅速增长，谷歌等客户的日均Tokens调用量已经达数万亿级别，远超传统Chatbot时代的负载水平。多模态、Agentic AI等新形态模型的快速涌现，正在拉动实时推理、大规模低延迟计算的新增长。

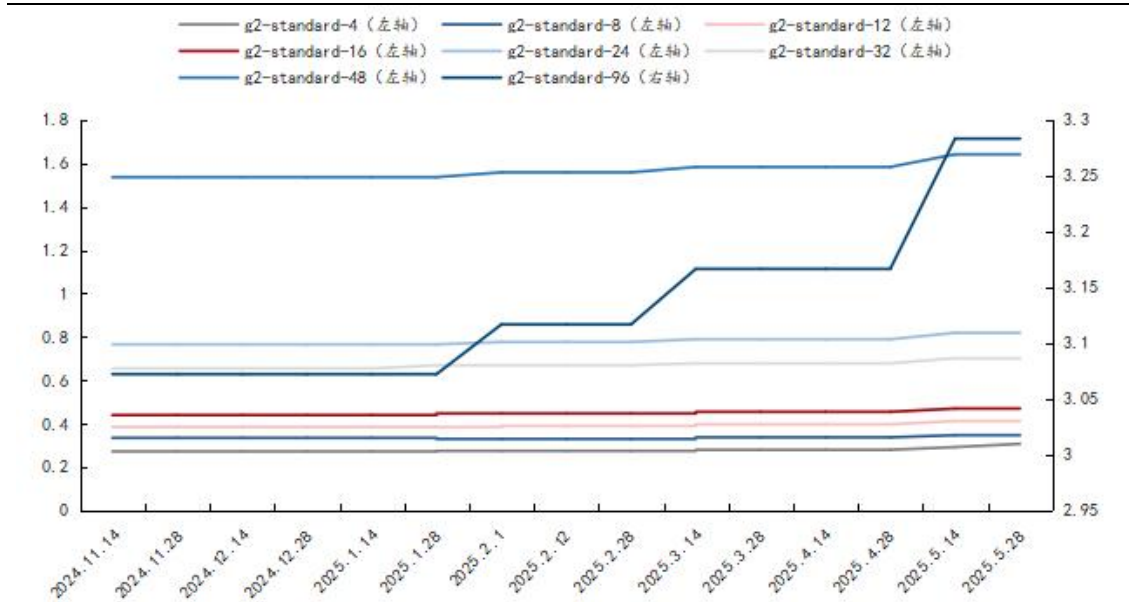
- **数据显示，各家云厂商推理芯片租赁价格均有不同程度上涨。**Azure的T4租赁实例Standard_NC64as_T4_v3平均价格由去年5月份0.52美元/小时上涨到1.28美元/小时，AWS的T4租赁实例g5g.metal平均价格由去年5月份0.62美元/小时上涨到0.74美元/小时。谷歌L4租赁实例g2-standard-96平均价格由去年11月份3.07美元/小时上涨到3.28美元/小时。各家云厂商普遍上调推理侧芯片租赁价格，显示出推理侧Token调用高增长，今年AI应用正加速落地。

图：Azure T4实例租赁价格



资料来源：Azure官网，国信证券经济研究所整理

图：谷歌云 L4实例租赁价格



资料来源：谷歌云官网，国信证券经济研究所整理

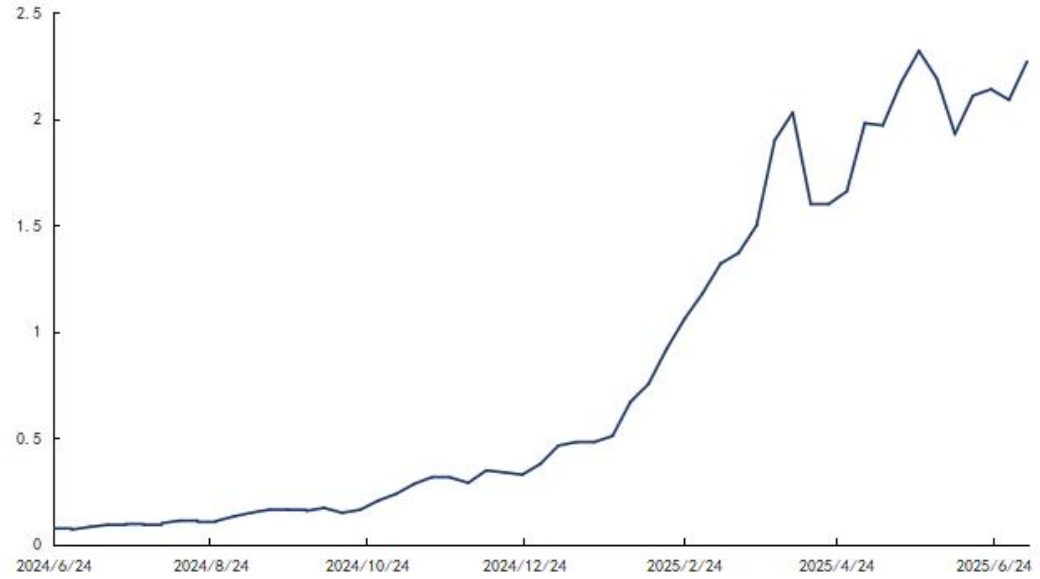
API调用量迅速提升，国产模型表现亮眼

国信证券 GUOSEN SECURITIES

- **OpenRouter是一个通用API平台，为开发者提供来自OpenAI、Anthropic、Google以及60多家实验室的400多个大语言模型的访问权限，所有模型均可通过统一的端点调用。**开发者无需为每个AI提供商集成不同的SDK，只需将应用指向OpenRouter的API，即可使用相同的OpenAI兼容请求格式访问任意模型。平台作为一个智能路由层，会依据价格、延迟、可用性等实时指标，把请求智能路由到最优端点，并在主线路超时或限流时自动故障转移。

- 据OpenRouter数据，进入2025年后平台API调用量呈现快速增长趋势，显示出下游需求的持续放量，截至2025年7月7日，平台周API调用量达2.27T，同比增长268.52%。从模型层面看，国产模型表现亮眼，DeepSeek V3周调用量达239B，环比增长63%。Kimi K2周调用量达33.9B，超Grok 4以及Claude 3.5 Haiku等海外顶尖模型。

图：API调用量快速增长



资料来源：OpenRouter官网，国信证券经济研究所整理

图：OpenRouter调用量数据（截至2025年7月18日周调用数据）

排名	模型名称	公司	使用量（tokens）	变化趋势	排名	模型名称	公司	使用量（tokens）	变化趋势
1	Claude Sonnet 4	anthropic	329B	↑ 20%	11	Mistral Nemo	mistralai	40.6B	↑ 7%
2	Gemini 2.0 Flash	google	277B	↑ 5%	12	GPT-4.1 Mini	openai	39.8B	↓ 2%
3	DeepSeek V3 0324 (free)	deepseek	239B	↑ 63%	13	GPT-4.1	openai	34B	↑ 5%
4	Gemini 2.5 Flash	google	172B	↑ 25%	14	Kimi K2	moonshotai	33.9B	new
5	DeepSeek V3 0324	deepseek	157B	↑ 14%	15	R1 (free)	deepseek	33.5B	↑ 14%
6	Gemini 2.5 Pro	google	149B	↑ 7%	16	Grok 4	x-ai	31.1B	↑ 1,285%
7	Gemini 2.5 Flash Preview 05-20	google	118B	↓ 43%	17	GPT-4o-mini	openai	27.1B	↓ 15%
8	R1 0528 (free)	deepseek	63B	↑ 59%	18	Gemini 2.5 Flash Preview	google	26B	↓ 33%
9	Gemini 2.5 Flash Lite Preview 06-17	google	61.8B	↑ 112%	19	Claude 3.5 Haiku	anthropic	23.1B	↑ 111%
10	Claude 3.7 Sonnet	anthropic	53.2B	↓ 4%	20	Gemini 2.0 Flash Lite	google	22.5B	↑ 26%

资料来源：OpenRouter官网，国信证券经济研究所整理

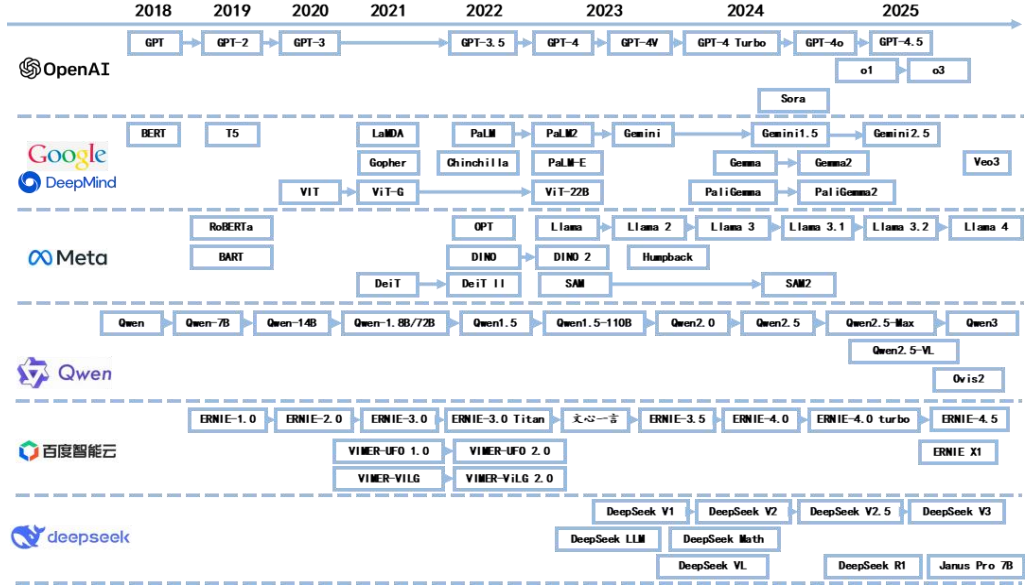
目录

AI应用有望重塑C端流量入口，互联网巨头具备先发优势

- **AI应用有望重塑流量入口，各个厂商积极卡位。**回顾互联网发展历史，超级应用大都由用户基数庞大的基础需求出发，成为庞大的用户流量入口，再逐步覆盖更多的需求层级，聚合和链接背后庞大的应用生态，掌握用户流量的分配权。进入AI时代后，随着大模型C端应用围绕个性化、强交互等方向实现价值增量，相关应用的流量迅速提高，有望推动新的入口级应用出现。当前各类厂商纷纷布局和卡位AI应用，结合自身资源和技术迭代趋势布局最具价值的领域，主要分为以下策略：1）投入聊天机器人，布局通用人工智能；2）切入生产力场景，以内容创作、垂直专业、工作效率类AI应用抢抓用户流量；3）布局垂类领域AI+解决方案，掌握垂直行业C端流量入口。

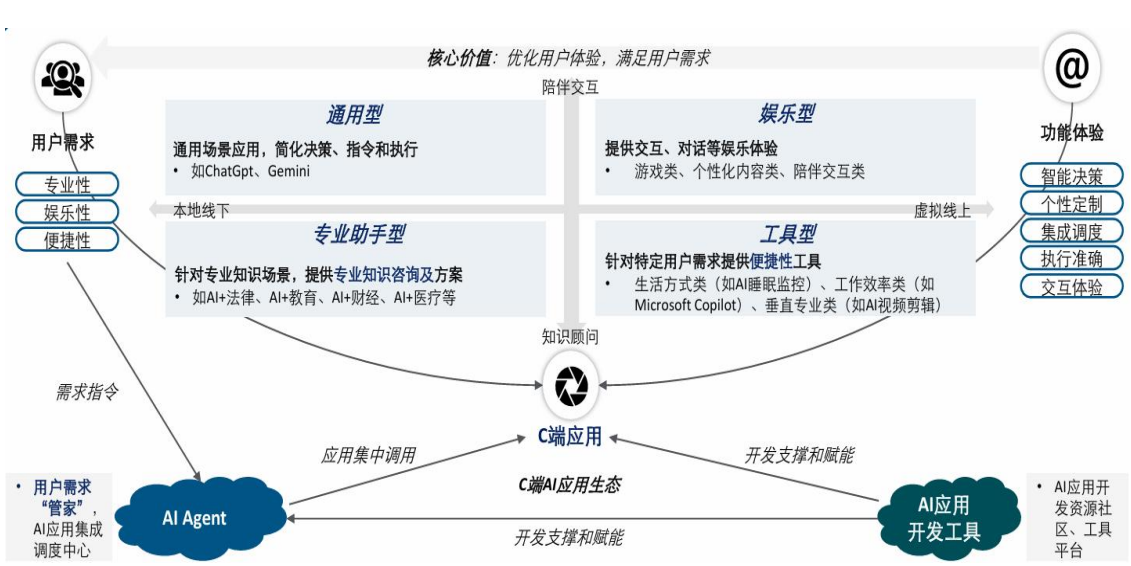
- **传统互联网巨头在AI领域具备先发优势，可利用专有数据和用户参与度将AI功能集成到现有的应用当中，在AI应用渗透领域具备先发优势。**互联网生态围绕超级应用（如微信、淘宝、谷歌等）发展，这些应用在平台上提供全面服务，从而形成了庞大的用户基础和高度的用户参与。通过这些应用，主要互联网公司可以访问具有行为、社交和商业特征的专有用户数据，对提供定制化的AI服务至关重要。得益于庞大的用户基础和高用户参与度，将AI功能整合到现有应用能够促进市场的采用，在相关领域具备先发优势。

图：各大厂布局通用AI进程



资料来源：公司官网，国信证券经济研究所整理

图：C端AI应用生态构成



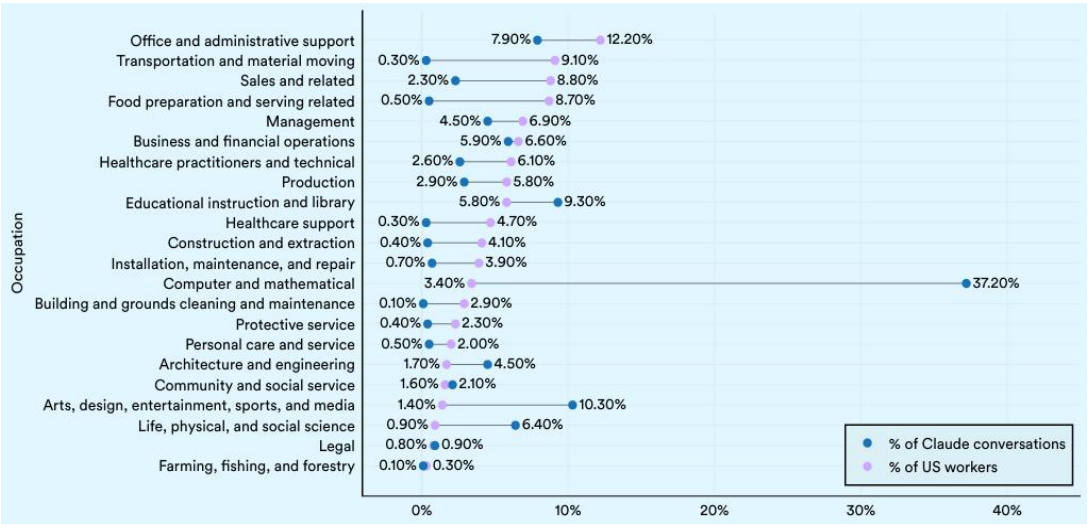
资料来源：德勤，国信证券经济研究所整理

AI被用于辅助工作，人机深度协同存在较大空间

- **编程成为人机协同的主要领域，办公类任务AI占比较低。**在TO C领域，当前AI正逐步融入用户工作当中，据Anthropic数据，在与Claude的对话中软件工程相关任务占据了数据集中最大的比例，达37.2%的对话涉及代码调试、网络故障排查等内容，第二大类任务是写作与编辑，占比为10.3%，这两类职业在美国经济中分别仅占3.4%和1.4%，远低于办公室行政和销售类岗位的比例，但成为AI的高频使用场景。科学和教育领域的职业也显示出占比例更高的AI使用率。

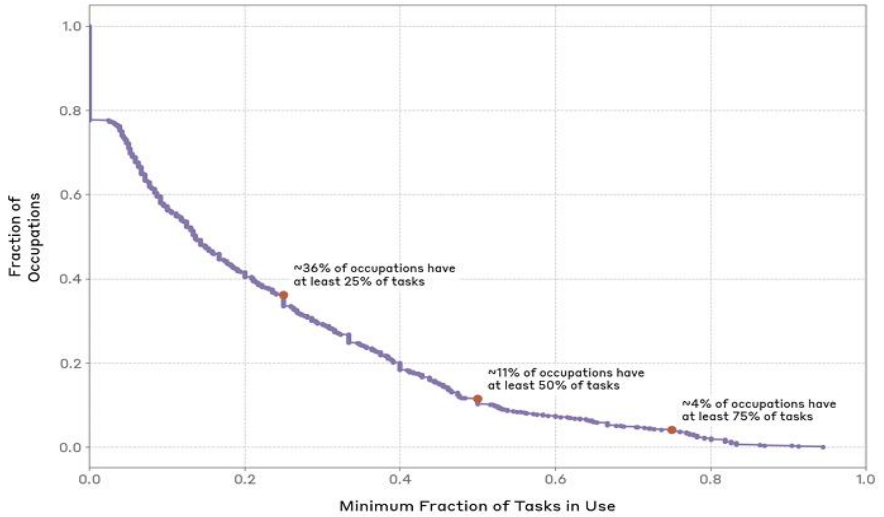
- **AI逐步融入工作，全面自动化仍存在提升空间。**Anthropic的研究发现，约36%的职业在其相关任务中，至少有25%的用户使用了AI，这表明AI的应用已经显著渗透到技术领域之外。然而，只有约4%的职业在其75%或以上的任务中使用了AI。此外，在57%的情况下，AI被用于辅助用户工作，而在43%的情况下，则是自动化完成任务，即AI直接替人执行任务。这表明对整个职业类别的全面自动化尚未发生，人机深度协同具备较大的提升空间。同时，AI的采用率在处于最高工资四分位的职业中达到峰值，但在工资的两个极端（即最低和最高工资）均显著下降。对于需要较多前期基础的职业（通常要求具备学士学位），其AI使用率比其在整体劳动力中的占比高出50%，而需要极少和极高基础的职位，AI采用率则相对较低。

图：AI主要应用领域



资料来源：Kunal Handa等-《Which Economic Tasks are Performed with AI? Evidence from Millions of Claude Conversations》-Anthropic-2025年-P6，国信证券经济研究所整理

图：国产AI芯片性能逐步追赶



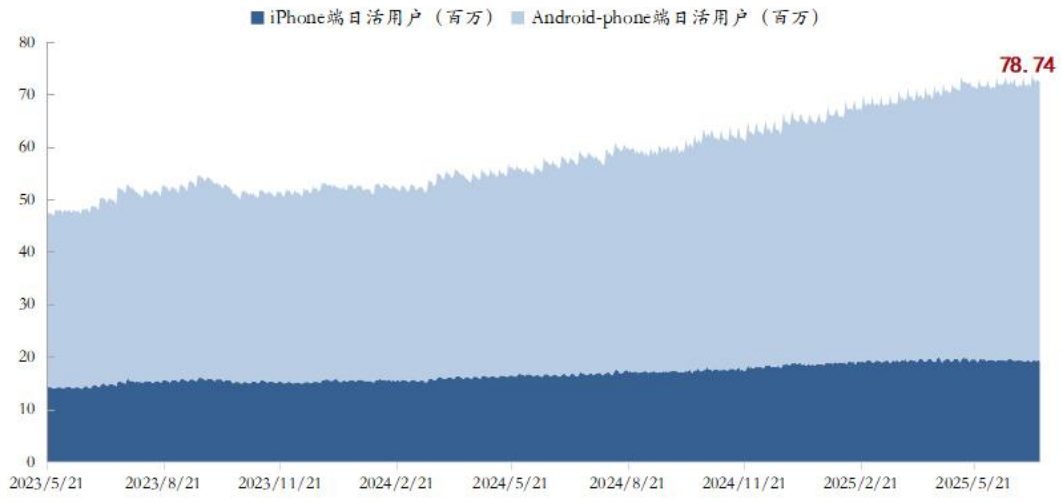
资料来源：Kunal Handa等-《Which Economic Tasks are Performed with AI? Evidence from Millions of Claude Conversations》-Anthropic-2025年-P7，国信证券经济研究所整理

AI+社区平台：Reddit

- **公司是全球性的社区平台，平台以兴趣为中心将用户组织在subreddit（子社区）中，社区内容由用户创建、筛选与维护。**当前拥有超10万个活跃的subreddit，其中有500多个社区订阅人数超100万。通过覆盖不同类型的话题，平台积累了庞大的数据资源，截至2024年底，平台拥有超20亿条帖子和200亿条评论，提供了真实的用户视角。平台中每个人都有投票权，帖子在社区里的展示优先级完全取决于用户投票结果。每个社区可以制定独特的社区规则，由版主根据社区背景与文化执行，从而创造和谐的社区环境。同时，平台运用AI根据用户搜索、话题、互动行为等因素推荐帖子与社区，推动2024年阅读帖子超30秒的访问量增长了30%。截至2024年，约53%的用户来自美国以外地区，但超90%的内容仍以英文发布，用户平均每天搜索查询超过4000万次。公司主要收入来源是通过在移动应用程序和网站上销售广告，其他收入包括内容授权（主要买家为OpenAI以及Google）、Reddit Premium订阅服务以及Reddit Gold和可收藏头像。

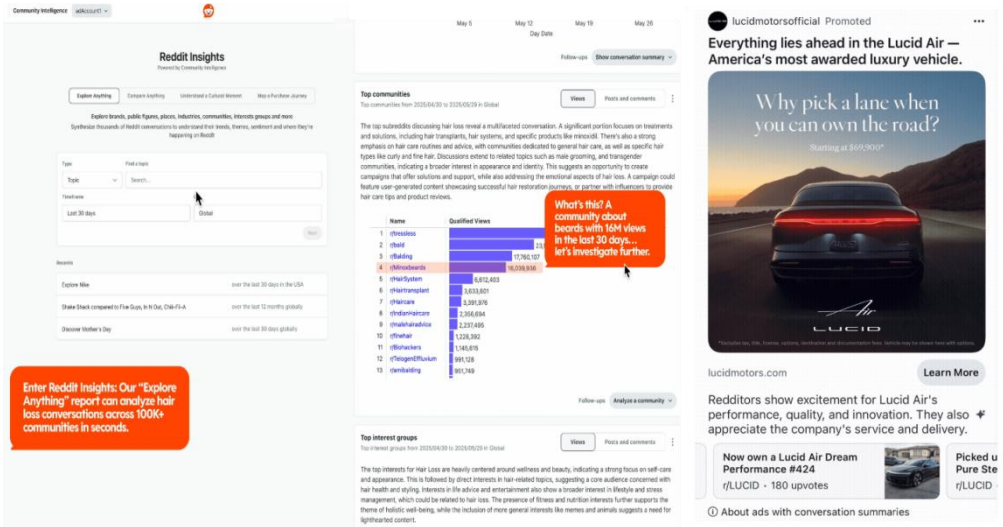
- 2025年6月，公司推出AI驱动的新功能：1）Reddit Insights：AI驱动的分析工具，使广告商掌握最新热门话题并据此规划营销策略；2）Conversation Summary Add-ons：AI驱动的对话摘要功能，自动筛选用户对品牌的正面评价和推荐，并展示在广告旁边。Jackbox Games和Lucid等品牌已参与早期测试，相比标准图片广告，新工具使得广告点击率提升了19%。

图：公司App端日活快速提升



资料来源：SimilarWeb，国信证券经济研究所整理

图：Reddit全新AI产品



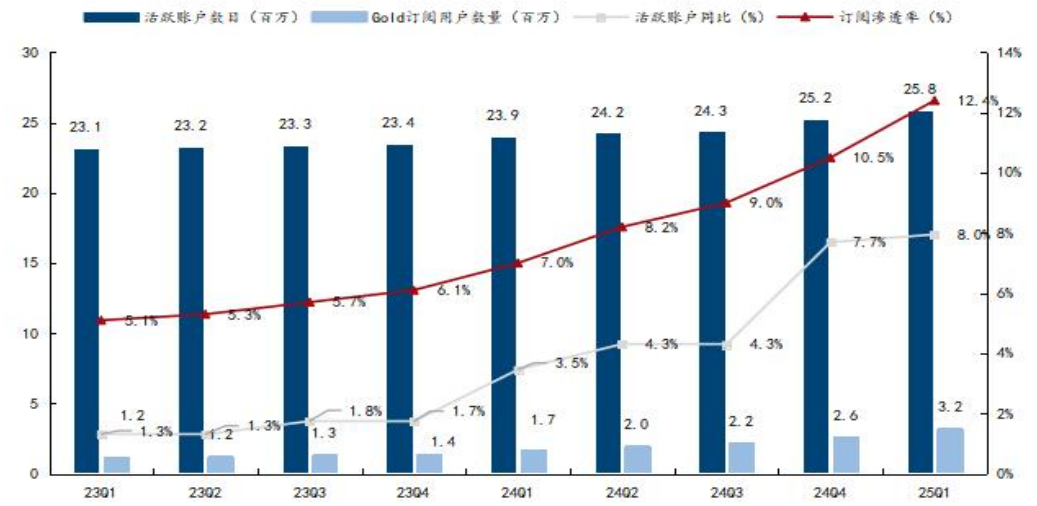
资料来源：公司官网，国信证券经济研究所整理

AI+金融：Robinhood

- **公司是美国首家提供免佣金股票交易且没有账户最低限额的经纪人，目前已发展出交易、理财、支付三大功能。**交易为公司最大收入来源，主要通过Robinhood APP、桌面交易平台Robinhood Legend为用户提供股票及衍生品、加密货币等交易服务，当前已支持包括比特币、狗狗币在内的22种加密货币交易。同时，公司为用户提供融资融券、高利息储蓄（Cash Sweep）、订阅用户服务（5美元/月，为用户提供研究报告、Level II市场数据等权益）、信用卡（Robinhood Gold）等其他服务。

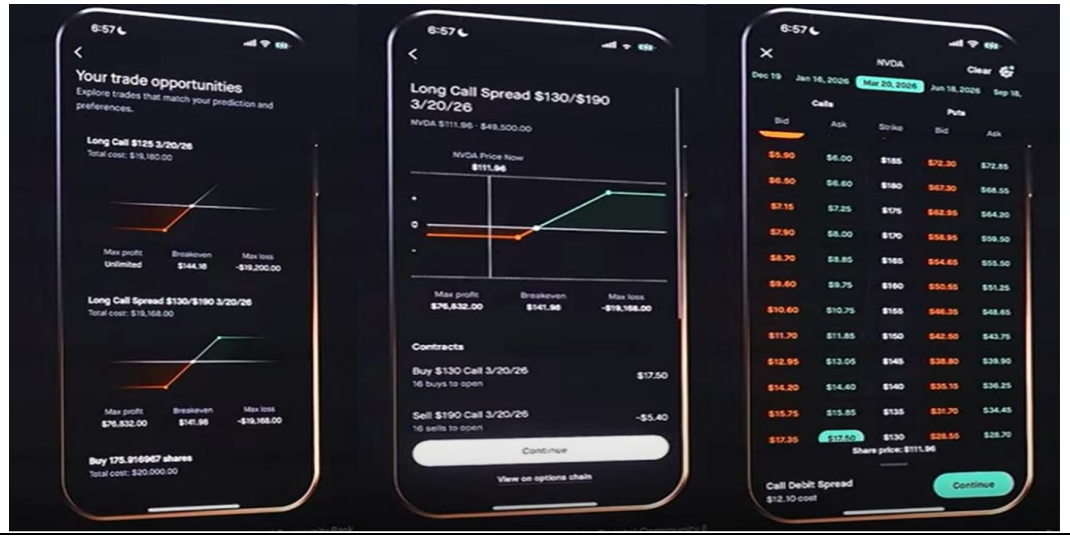
- **收购Pluto Capital，推出AI投资助手。**2024年7月，公司宣布收购AI投资研究平台Pluto Capital，计划将基于大模型的市场数据分析、定制投资策略、实时投资组合优化等功能集成到平台当中。2025年3月，公司在The Lost City of Gold活动中推出基于AI的投资助手Robinhood Cortex，旨在帮助用户更好地把握市场动向、识别投资机会，相关功能将面向Gold订阅用户推出。Robinhood Cortex具备以下功能：1）个股摘要：帮助投资者解答股票涨跌原因，只需进入股票的详情页，AI会快速生成影响该股票价格的分析，支持展示数据来源以及音频播放功能；2）交易构建器：简化交易流程、为投资者提供新策略，通过分析价格信号、技术面等信息，根据用户的要求在市场上筛选出可能的交易机会，例如用户只需选定目标价格区间，AI将自动生成多个期权交易策略，用户可根据需求对方案进行调整。

图：公司订阅用户渗透率快速提升



资料来源：Bloomberg，国信证券经济研究所整理

图：Robinhood Cortex相关功能



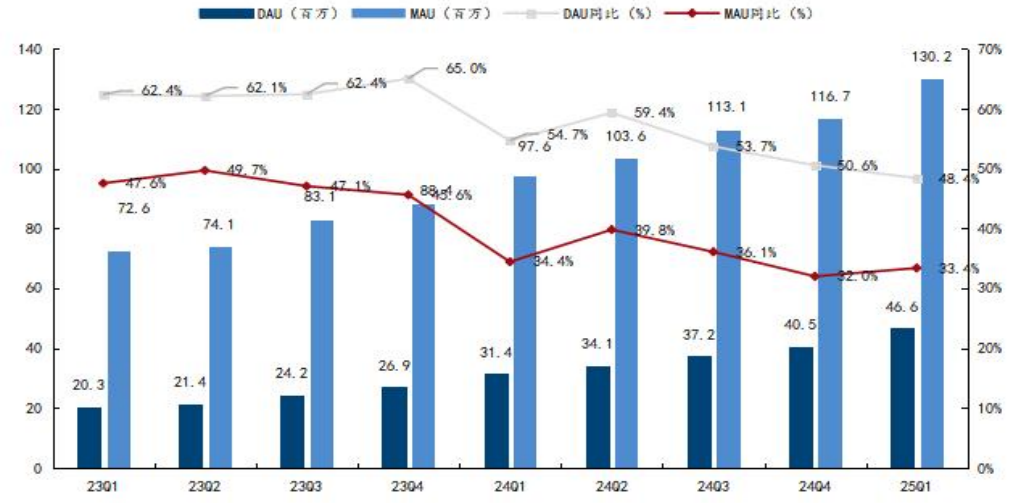
资料来源：公司官网，国信证券经济研究所整理

AI+教育：多邻国

- **公司是全球月活人数最多的语言学习平台，主要产品为Duolingo APP，面向全球各年龄段用户提供40多种语言的学习课程，覆盖语法、词汇、听说读写等多个方面。**公司核心商业模式为收取订阅会员费，80%以上收入来自订阅收入，包括以下类型：1）Super Duolingo：普通会员订阅服务，12.99美元/月或83.99美元/年，可享受无广告学习、拥有无限测试机会、跟踪个人学习进度数据、解锁额外语言内容等功能；2）Super Duolingo家庭计划：约119.99美元/年，最多允许6个账户共享Duolingo Super会员福利；3）MAX：2023年4月发布的基于GPT-4的AI高级订阅，30美元/月或168美元/年，功能包括答案释义（个性化错题分析）、角色扮演（与不同的AI角色进行互动对话）、视频通话（可以与角色Lily自由对话）。除订阅外，公司还提供Duolingo ABC、Duolingo English Test等其他业务。

- **公司的AI应用通过提升用户体验以及内部降本来创造价值：**1）区别于通用聊天机器人，Lily具有鲜明个性，可为用户提供情绪价值，从而提升用户粘性。2024年9月新增视频通话功能后，Max订阅在四季度占到总订阅的5%，25Q1提升至占总订阅用户的7%；2）AI协助课程、播客DuoRadio内容制作，由AI生成新的148种语言课程、耗时约1年，此前100多门课程内容花了12年，DuoRadio从2门课程增加到25门以上并将总集数从300集增加到超15000，只用了两个季度，曾经需要5年以上，同时节省了99%的成本。

图：公司活跃用户数量快速增长



资料来源：Bloomberg，国信证券经济研究所整理

图：与Lily视频通话功能

资料来源：公司官网，国信证券经济研究所整理

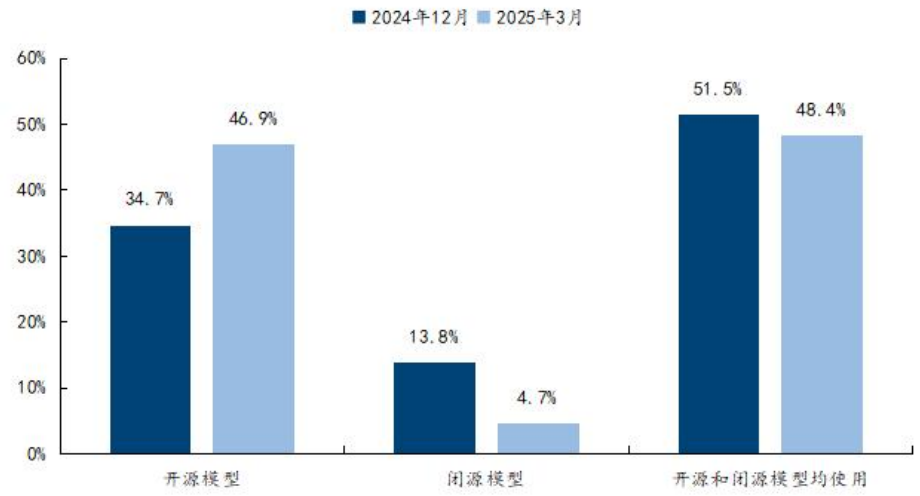
目录

开源模型提升企业投入意愿，刺激国内上云需求

- **开发工具和生态的繁荣大幅降低行业应用门槛，加速产业智能化落地进程。**多种开发平台汇聚了多样化开发框架、工具组件、算法资源、数据集等，开发者可快速调用、微调模型，从而验证想法并构建AI应用。开源模型具备可控性强、可定制性强以及社区支持丰富等优势，推动更多企业采用AI作为技术解决方案。据阿里云数据，开源模型的采纳比例持续提升，2024年12月企业使用开源模型的比例为34.7%，到2025年3月使用开源模型的比例增长至46.9%。

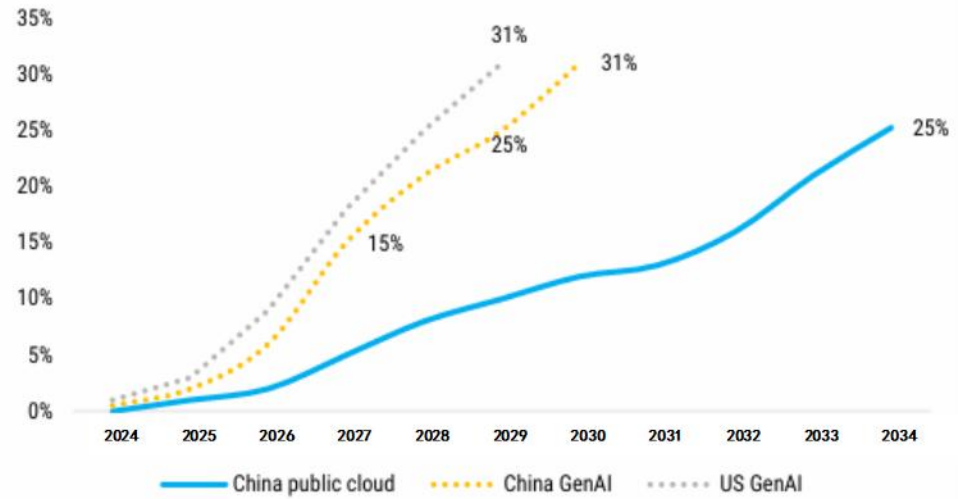
- **AI技术和解决方案已深入到传媒、医疗、机器人、制造等多个行业，通过创新产品和服务、优化生产流程来推动行业的智能化转型。**例如，凯撒医疗利用生成式AI门诊记录工具Abridge完成超1000万份就诊摘要，自2024年10月到2025年3月期间Abridge年经常性收入从5000万提高到1.17亿美元；AI代码编辑器Cursor已成为数百万程序员的首选编辑器，每日自动生成或修改超10亿字符代码，自2023年3月到2025年4月期间年经常性收入从100万提升至3亿美元。据Morgan Stanley数据，到2030年中国企业的生成式AI工作负载渗透率将达到31%，美国的生成式AI工作负载渗透率将在2029年达31%，中国企业AI的采用进程将比美国落后约12个月，并在后续达到类似水平。同时，AI应用将提升企业上云意愿，预计未来将有超25%的工作负载被迁移至公有云中。

图：企业采用AI比率快速提升



资料来源：阿里云-《中国人工智能发展报告（2025）》-2025年-P6，国信证券经济研究所整理

图：AI将逐渐在企业端代替人工



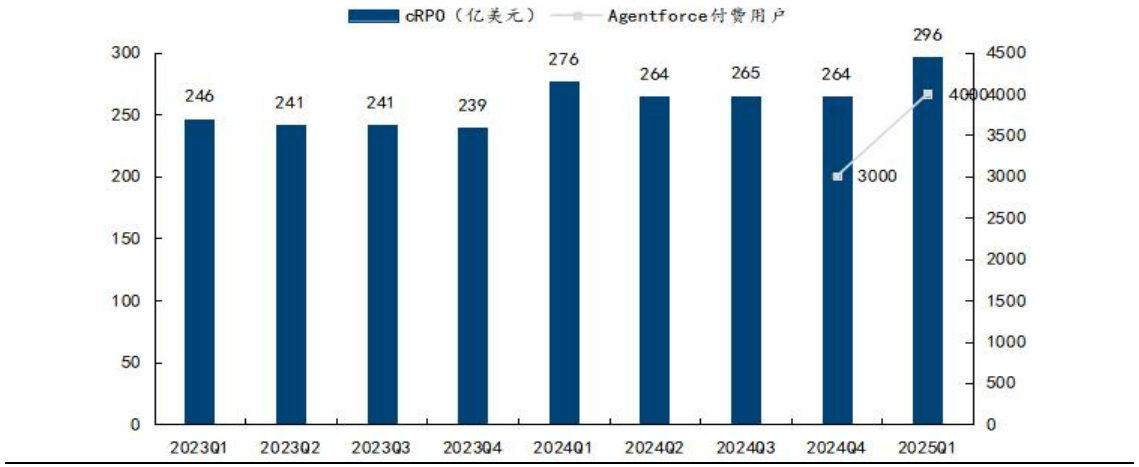
资料来源：Morgan Stanley，国信证券经济研究所整理

AI+CRM：赛富时

- **赛富时主要提供客户关系管理技术，帮助企业通过数据、AI和自动化与客户建立联系。**公司以AI驱动的Salesforce平台为核心，整合销售、服务、营销、AI、分析等多项功能，通过连接客户数据帮助企业打造完整的客户视图。公司通过直接销售以及间接合作伙伴向全球企业销售服务，通常采用订阅模式，同时支持第三方利用公司平台和开发工具创建可扩展功能和新应用，这些应用可以与公司产品捆绑销售，也可以单独销售。

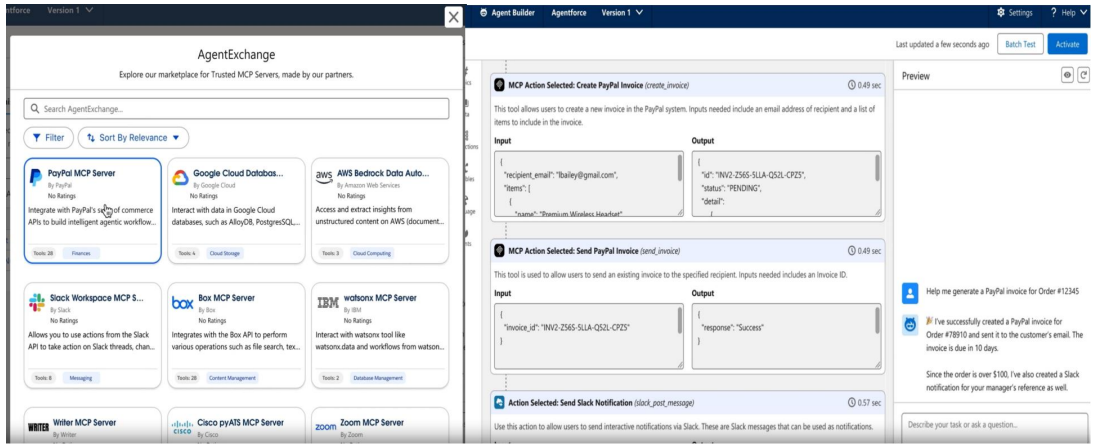
- 2025年6月，公司正式发布Agentforce 3.0，通过Data Cloud对Atlas推理引擎进行改进：1）响应速度快50%；2）除OpenAI外还支持Claude，未来将支持Gemini，提供更广泛的模型选择，在某个模型运行放缓或出现故障时会在后台自动切换至其他模型；3）现已在加拿大、英国、印度、日本和巴西上线，并新增6种语言支持；4）**已获得FedRAMP High认证，通过Government Cloud Plus服务成为美国公共部门组织使用的工具。**上线以下全新功能：1）指挥中心：为管理者提供统一视图，以监控AI智能体的运行状态、衡量绩效并优化业务结果；2）MCP以及A2A：内置MCP客户端，可用MuleSoft把API转成MCP服务接入Agent，用Heroku部署自定义MCP服务器适配企业内部系统，在Slack里让Agent读取消息、发文件、做决策等；3）AgentExchange：AppExchange的AI替代方案，构建AI智能体应用商店生态，不仅使客户能够快速配置可信的第三方Agent模版，还可以使用包括AWS、Box、Cisco、Google Cloud、IBM等超30家合作伙伴的MCP服务器，通过安全的AI智能体网关无缝接进系统。

图：公司cRPO及Agentforce付费用户快速提升



资料来源：公司官网，国信证券经济研究所整理

图：Agentforce 3.0融合MCP生态



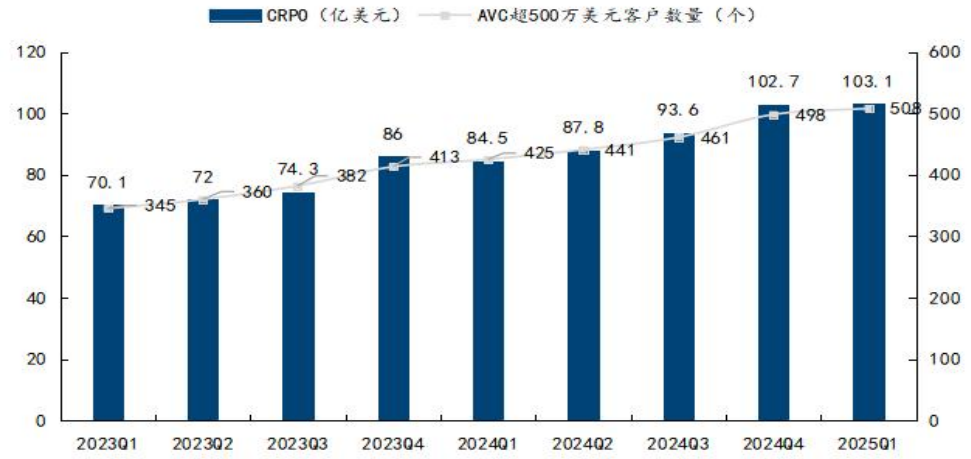
资料来源：公司官网，国信证券经济研究所整理

AI+工作流管理：ServiceNow

- **公司致力于构建数字化企业的端到端智能工作流自动化平台，产品基于其智能平台Now Platform交付。**Now Platform为基于云的低代码开发平台，内嵌AI功能，能够连接不同部门、系统内的数据，使组织能够快速构建自定义应用程序、自动化流程和工作流。基于Now Platform，公司将业务纵向延伸CRM、HR等领域。Now Assist是公司提供的基于生成式AI的解决方案，用于提升工作效率和员工体验，当前Xanadu版本的Now Platform已将Now Assist集成到Creator（代码）、ITSM（IT管理）、HRSD（人力）以及CSM（客户关系管理）等56种流程当中。公司于年初推出的Agent调度平台AI Agent Orchestrator，可自动调用不同部门不同功能的Agent完成任务。公司同时推出AI Agent Studio，为基于自然语言构建AI智能体的可视化平台，用户只需描述业务目标、代理角色与任务逻辑即可通过自然语言生成智能体，由AI Agent Orchestrator管理、调度、测试与部署。

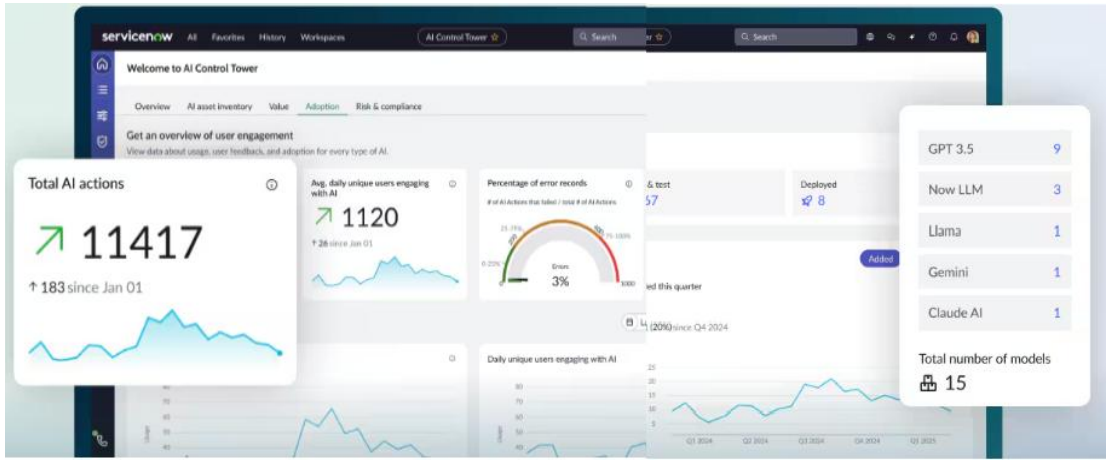
- 2025年5月公司发布更新后的ServiceNow AI平台，推出以下全新功能：1）AI控制塔：可视化AI数据中台，用于管理和展示不同Agent的调度情况；2）智能体网络：支持智能体跨工具、跨系统协作，已与微软、英伟达、Google、Oracle等实现对接，Agent可共享上下文、协调任务等；3）升级CRM系统：从响应式支持到主动式客户交互，包括报价配置、订单履行、客户服务与续约等全过程；4）Apriel Nemotron：与英伟达合作开发的推理大模型。2025年7月，公司通过OEM集成Gemini、Azure OpenAI、Claude等第三方模型。

图：公司cRPO及大客户数量快速增长



资料来源：公司官网，国信证券经济研究所整理

图：平台可视化所有智能体动态



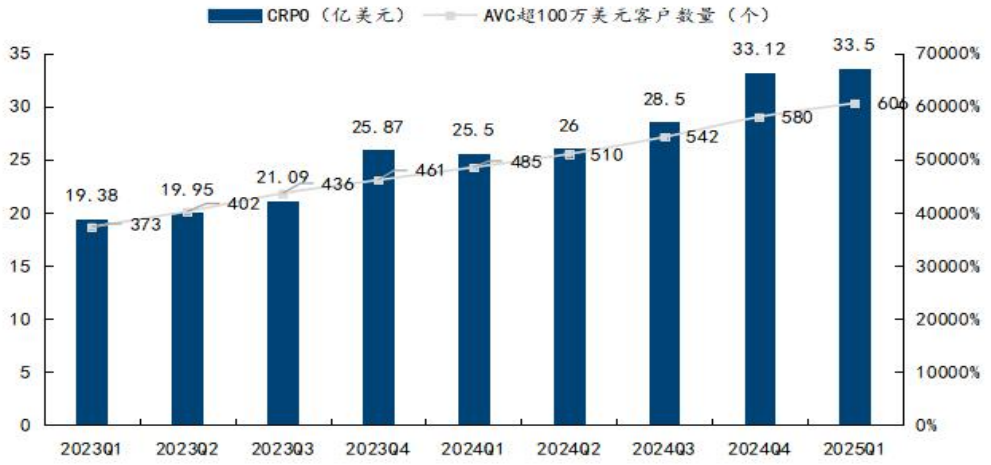
资料来源：公司官网，国信证券经济研究所整理

AI+数据库：Snowflake

- **公司专注于数据云服务，旨在通过创新平台帮助组织实现数据的统一管理和高效利用。**公司推出了AI数据云网络生态，部署于全球三大主流公有云上，覆盖全球47个地区，使客户、合作伙伴、开发者、数据提供方和数据消费者打破孤岛，通过将数据整合使AI解决业务问题、构建数据应用。公司采用按量计费商业模式，客户只需为实际使用的资源付费。同时，作为即服务交付的产品，公司的平台几乎不需要维护，使客户可以专注于业务而非基础设施管理。

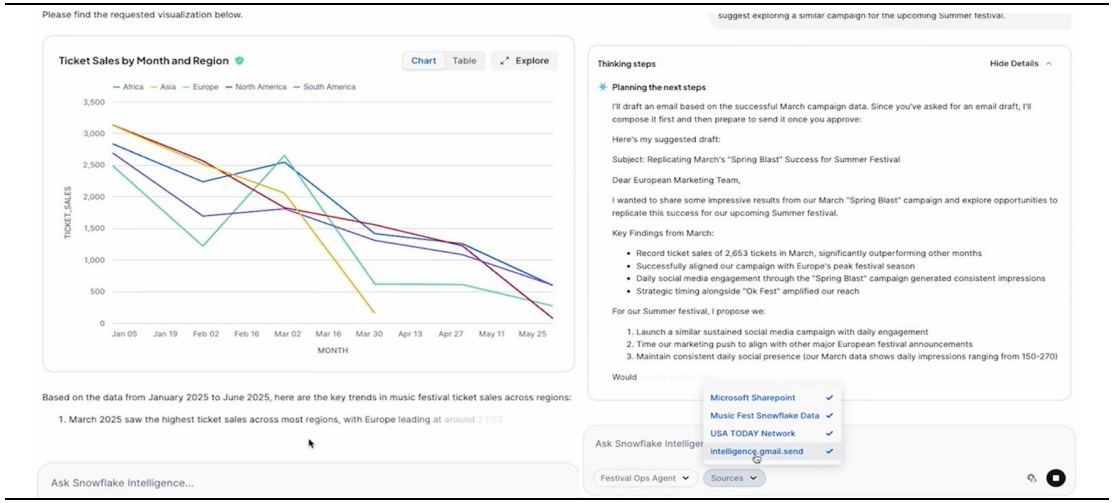
- 2025年6月，公司推出更新后的Snowflake Intelligence，允许用户使用自然语言发掘数据，支持提出复杂问题并自动操作，结合结构化和非结构化数据，并继承了Snowflake内建的数据治理和隐私保护机制。底层由Anthropic和OpenAI的大模型提供支持，并结合Cortex Agents技术。同时，公司发布多种全新AI产品：1）Data Science Agent：为数据科学打造的智能助手，通过自然语言自动执行机器学习全流程，包括数据准备、模型训练等；2）Cortex AISQL：将AI引入SQL语法中，支持多模态数据处理。团队可以使用SQL分析文档、图像等非结构化数据，包括增强版Document AI（可从复杂PDF中提取表格结构）和Cortex Search的增强检索功能；3）AI observability：对生成式AI应用的无代码监控，提供与多个主流大模型的集成，包括OpenAI、Anthropic、Meta、Mistral等，并确保模型运行在Snowflake的安全范围内。

图：公司cRPO及大客户数量快速增长



资料来源：公司官网，国信证券经济研究所整理

图：Snowflake Intelligence发掘数据并自动发送邮件



资料来源：公司官网，国信证券经济研究所整理

海外AI应用厂商汇总

图：海外AI应用厂商汇总

领域	公司	主营业务	AI应用
AI+企业服务	ServiceNow	企业自动化工作流平台	Now Assist：内置于平台工作流中，通过生成式AI提升企业在IT、客服、HR和开发等关键领域的生产力
	Salesforce	客户关系管理平台	Agentforce：可以通过简单的配置打造定制化的AI助手，同时面向客服提供AI服务助手
	SAP	ERP系统及其他解决方案	Joule：内置于SAP应用中，能够整合、分析企业数据，打破信息孤岛，提升业务流程效率
	Asana	工作管理和协作平台	AI Studio：无代码构建器，允许用户设计和部署关键工作流中的人工智能代理
	GitLab	DevOps平台	Duo：AI驱动的开发者工具，提供自动化测试、智能代码建议等功能
AI+教育	Duolingo	语言学习平台	Max：提供由GPT-4支持的AI角色互动对话等功能，该功能允许学习者与AI角色Lily进行互动对话
AI+金融	Intuit	财税管理和金融服务	Intuit Assist：AI财务助手，可以帮助用户完成创建发票、发送提醒以及优化营销活动等任务
	Robinhood	投资助手	Robinhood Cortex：帮助投资者解答股票涨跌原因，自动生成多个期权交易策略，用户可根据需求对方案进行调整
AI+电商	Shopify	电子商务平台	Magic：提供撰写产品描述、编辑图片、撰写博客内容和营销邮件，为客户提供AI客服支持等功能
AI+医疗	Tempus	综合基因组分析及临床诊断	olivia：利用生成式AI为患者及护理人员提供个性化健康管理支持，包括汇总档案、健康监测等功能
AI+广告	Applovin	广告投放平台	AXON 2.0：高效地在需求方和供应方之间进行大规模广告拍卖，极大提升广告投放的精准度和效率
AI+游戏	Unity	游戏开发引擎	Muse：AI辅助创作平台，加速视频游戏、数字孪生等实时3D应用程序及体验的创建过程
AI+设计	Adobe	数字媒体解决方案	Firefly：提供多种图像、视频编辑的功能，应用于图像生成、合成、修复及风格迁移等任务
	Autodesk	3D设计解决方案	Autodesk AI：集成在公司产品当中，通过自动化、分析和增强功能来改善设计与制造的工作流程
AI+数据服务	Palantir	数据分析平台	AIP：运用大语言模型，赋能传统数据分析平台，并允许客户构建定制化AI应用
	Snowflake	数据仓库及分析	Snowflake Intelligence：基于Cortex AI和Cortex Search，为企业提供基于数据基础构建的数据代理
AI+合同管理	DocuSign	合同管理平台	IAM：使组织能够集中存储、管理和分析来自任何来源的协议，并有效地将非结构化协议转换为结构化数据
AI+项目管理	Atlassian	项目管理工具	Atlassian Intelligence：帮助团队实现生成与转换内容、汇总关键信息、加速搜索等功能
AI+搜索	Elastic	数据搜索引擎	Elastic AI Assistant：帮助工程师实现解读日志消息和错误、优化代码、撰写报告、增强安全等功能

资料来源：公司官网，公司财报，国信证券经济研究所整理

中国AI应用厂商汇总

图：中国AI应用厂商汇总

领域	相关公司	Agent相关布局
通用企业Agent	用友网络	智友借助自然语言处理和智能调度系统，连接企业内部财务、人力、营销、供应链等多个企业级AI智能体，用户用自然语言发出指令，智友就能理解分析，自动分解任务，协调数百个专业分工智能体合作，无需人工参与即可完成任务并交付成果，用户只需在关键决策环节确认审批
	金蝶国际	发布了苍穹AI管理助手及其移动形态，为财务、人力、采购、开发等多个业务场景提供了个性化智能解决方案。苍穹Agent平台提供多模型混合应用能力，能够根据不同的业务需求，灵活调用DeepSeek、金蝶苍穹大模型等业界领先的AI大模型进行业务处理
	泛微网络	发布泛微·数智大脑Xiaoe.AI，为用户提供7*24工作的智能助理，帮助用户以自然语言对话实现找功能、找资料和语音办公事务等日常工作，可准确识别用户的工作意图，帮用户完成事项办理
	新国都	发布AI数字员工产品，基于企业私有化部署的GPT模型，通过企业数据的导入和训练，可以根据企业管理各模块的需求，生成不同功能的数字员工，涵盖人力、客服等多个领域
	新大陆	发布星驿付与慧徕店AI营销助手，具备智能语义识别、智能问答、降本提效等功能。公司积极推进商户运营行业智能体场景开发平台的孵化，构建起覆盖AI商户审核智能体验、AI客服智能体、AI风控反洗钱智能体在内的垂直领域智能体矩阵
金融	京北方	全面构建起面向未来的AI Agent，AI大模型服务平台搭载智能运维解决方案，能够实现更智能的资源调度、自动化运维和精准的故障预测，从而降低运营成本并提高系统稳定性。该平台结合阿里通义千问大模型的自然语言处理能力，探索智能化数据查询方案，利用语义理解技术识别运维人员的查询意图，并自动调用相关数据接口，优化查询流程，提高数据获取的便捷性和响应速度
	宇信科技	近期推出的AI-SCRM私域智慧运营平台4.0版本，集成DeepSeek等诸多金融大模型，本次升级创新推出的零代码Agent构建平台，将复杂的AI模型训练转化为直观的拖拽操作。运营人员通过图形化界面即可完成智能客服工作流编排、精准营销策略树搭建和自动化质检流程配置，重塑银行私域运营范式
	中科金财	AI Agent开发运行平台提供Agent创建、多基座模型调用、工作流定义等功能，能够根据行业场景需求自动路由调度最适合的大模型并完成Agent创建，已形成生成式业务流程Agent、智能客服Agent、智能信贷Agent、智能投研Agent、账户管理Agent、智能座舱Agent等产品，以打造多任务、复杂任务的智能体为目标，在部分产品中使用Multiple Agent架构
	顶点软件	全面融合DeepSeek等AI大模型，C6客户运营依托Agent构建客户画像，并据此生成高度适配的个性化营销服务方案，提高营销效率及准度；P6产品供给通过大模型打造智能投研能力，实现产品研究、产品运营和产品风控的提升；W6资产配置为客户提供精准的账户管理建议；E6效能管理运用多维度指标数据对绩效管理、AUM效能进行分析，精准识别异常数据
	天阳科技	推出DeepSeek版包含产融分析和拓客智能体的产融大模型产品，基于大数据+大模型+机器学习的分析能力，通过50+智能Agent协同矩阵，在数分钟内生成专业级产融报告，覆盖企业竞争力评估、营销策略、融资方案设计等客群经营全流程
	信雅达	发布FinDOC多模态文档智能体，支持文档解析、文档智能问答、多维度内容审核及自动化文档生成等核心功能，同时生态兼容设计支持无缝对接其他Agent应用开发平台
政务	新致软件	以新致新知人工智能平台、新知大模型、金融行业数字资产为核心基座，结合行业实际情况，推出各行业应用Agent，在金融领域专注于营销展业、产品解读、产品核验、智能核保、理赔助手和对练培训等应用。企业服务领域专注于政务办公场景，包括类案检索、卷宗生成以及汽车领域的营销智能工牌等应用
	博思软件	公司在智慧财政财务领域的智能探索、智能问答、智能协办、智能报告均有相关应用，基于财政一体化、运行监测知识和数据预训练，结合国产化通用大模型、向量库检索增强、知识图谱等技术，进行多应用场景微调，致力打造财政垂直领域AI智能中台和多场景AI Agent。同时，公司在政府采购等公共采购领域开展相关预研工作。
	久其软件	公司基于女娲GPT已开发了多个领域与行业化Agent，助力政企客户快速接入大模型、连接业务、调优、快速应用，降低大模型应用门槛，并解决业务系统融合等应用难题。已通过Agent智能体实现智能分析、智能统计等，帮助企业更高效地处理数据和进行决策
法律	华宇软件	发布法律行业垂类大模型华宇万象，构建了以大模型+为核心的应用生态。发布万象+Agent开发平台，在公安、政法委等多个行业客户单位部署上线，发布基于此平台搭建警情分析等智能体应用，助力客户新价值创造
	金桥信息	金桥与阿里合作研发多元解纷平台，AI技术不断赋能多元解纷业务，利用Agent技术提升司法和政务效率
医疗	嘉和美康	推出新一代智能电子病历平台（V7），深度融合AI前沿技术与临床实践，为临床工作人员提供AI助手和虚拟病房等智能数据交互功能，为诊断支持、辅助诊疗、病情预警、疾病风险预测提供支持
	国脉科技	发布居家养老场景AI智能体，功能包括前期接触与适应支持：帮助公司与家属进行更好沟通；日常生活与社交娱乐：为长者日常生活、兴趣培养和社交拓展提供全面支持；健康与安全保障：从日常健康监测到应急处理，为长者的身体和心理健康提供持续支持与保障
	迪安诊断	与华为云联合打造迪安医检大模型智能体——AI健管专家迪晓智，为面向B端、C端的AI健康管理平台，提供常规检验、功能医学检测和体检报告解读、基于检测结果生成健康促进书,提供疾病风险评估、就医科室推荐及生活方式干预建议、通过对话获取深度健康咨询等功能
工业	赛意信息	赛意·谷神平台将AI能力嵌入善为·LCDP低代码流程引擎，打造流程AI Agent，可根据企业业务需求和历史数据，自动推荐最佳的流程设计方案，帮助企业节省时间和资源，提高流程设计的准确性和效率。通过Agent+赛意ITSM运维平台构建自动化工单处理体系，整合智能客服与企业知识库。通过GPT客服接入实现IM对话、智能查询、互动问答等功能
	鼎捷数智	发布鼎捷Indeth AI智能体平台，可以降低AI开发门槛，快速搭建AI应用，应用已覆盖企业“研发设计、生产制造、质量管控、经营管理、服务售后”五大领域，具备采购参谋、财务参谋、文生设计、设备助手等功能
虚拟机	深信服	提供端点安全Agent、VDI Agent、云主机Agent等，保障企业网络安全和设备管理

资料来源：公司官网，公司财报，国信证券经济研究所整理

目录

风险提示

- AI应用落地不及预期、市场需求不及预期、行业竞争加剧、宏观经济波动、新技术研发不及预等。

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