东吴证券 SOOCHOW SECURITIES

# AI Agent深度（二）：2025 Agent元年，AI从L2向L3发展

证券分析师：张良卫
执业证书编号：S0600516070001
联系邮箱：zhanglw@dwzq.com.cn

证券分析师：周良玖
执业证书编号：S0600517110002
联系邮箱：zhoulj@dwzq.com.cn

研究助理：张文雨
执业证书编号：S0600123070071
联系邮箱：zhangwy@dwzq.com.cn

2025年5月4日

# 核心观点



1. **我们认为2025年是Agent元年：**AI正从L2（推理者）向L3（Agent/智能体）进化，标志着AI从“思考”走向“行动”。这一转变由四大驱动力促成：①**技术成熟度达到临界点：**强大的多模态基础模型（能理解视觉信息如GUI界面）和成熟的强化学习训练方法已准备就绪。②**标杆产品的出现：**行业领导者（如OpenAI, Google, Anthropic）推出了关键产品（如Operator, DeepResearch），基准测试（如RE-Bench）显示顶尖Agent在特定任务上的效率已可匹敌甚至超越人类专家。③**MCP协议的普及将促进Agent生态的互联互通**。④**市场需求驱动：**经历了大模型能力竞赛（2023年）和初步应用探索（2024年）后，市场（尤其是B端）迫切需要AI能够落地解决复杂业务问题、自动化多步骤流程，并带来显著的生产力提升，Agent的出现恰好满足了这一需求。

2. **为什么要关注Agent？我们认为其重要性在于：**①**深度自动化：**Agent具有深度自动化、指数级效率提升和成本优化潜力，将人类从重复性劳动中解放出来，聚焦更高价值的创造性工作。②**通往AGI：**Agent（L3）是通往通用人工智能（AGI）和具身智能的关键环节。③**重塑互联网入口：**Agent可能改变用户获取信息和完成任务的方式，挑战传统搜索引擎，并可能使操作系统、浏览器或“超级App”成为新的核心入口。我们预计入口级通用Agent的竞争将在2025年下半年开启。

3. **Agent的竞争格局是“巨头环伺，新锐突破”：**①**巨头环伺：**大型科技平台（OpenAI, Google, 微软；国内BAT、字节、华为等）凭借模型、数据、算力、生态优势主导通用Agent和平台生态的构建。②**垂直机会：**垂直领域凭借深度领域知识和工作流整合仍有创新机会，但长期面临通用Agent能力提升的威胁。初期AI应用价值高度依赖模型能力，但简单的“浅层套壳”产品（即Wrapper）缺乏壁垒，易被颠覆。真正的护城河在于复杂工作流的可靠编排、高质量工具集成能力和深度领域知识。

4. **投资建议：**①**重视Agent投资窗口：**2025年是布局Agent领域的重要窗口期，需密切跟踪基础模型（尤其多模态、推理、规划）、强化学习、工具调用可靠性、推理成本优化以及标准化协议（如MCP）的进展。②**长期配置平台巨头：**拥有强大基础大模型、算力、数据和生态系统的大型科技平台公司是Agent时代的核心受益者，最有可能主导通用Agent的发展，并能整合或取代单一功能应用，具备长期配置价值。例如海外的Google、微软，以及与OpenAI、Anthropic深度绑定的公司；国内的阿里、腾讯、字节（未上市）。③**关注垂直领域领跑者：**在通用Agent能力尚未完全成熟之前，那些在特定垂直赛道已经建立深厚领域知识壁垒、拥有清晰商业模式和客户基础的垂直Agent提供商具有较高的短期增长潜力。我们认为知识工作领域（如编程、研究、法律）将是最先落地的场景，其中，编程领域会是最快落地、最先实现PMF和商业化的领域，已有成功案例（如Cursor、Devin）。其他垂直应用也值得关注：我们总结了30家上市公司在垂类Agent方面的布局，其产品基本符合Agent定义且具有垂直领域的比较优势。例如出版校对（果麦文化）、电商外贸（焦点科技）、企业服务（创业黑马）、美学设计（美图公司）等。建议关注其利用AI Agent解决具体行业痛点的能力和商业化进展。

5. **风险提示：**技术成熟度风险，高成本风险，商业模式不确定性风险，竞争加剧风险。

# 目录


一、 为什么说 2025 年是 Agent 元年？
AI 从 L2 向 L3 进化
驱动力：技术成熟度达到临界点；行业领导者推动，标杆产品验证；市场需求驱动
定义：不是所有的 AI 模型/产品都叫 Agent；Agent 的四个必要构成（缺一不可）；Agent 的智能程度是有层次和梯度的

二、 Agent 为何重要？
深度自动化、指数级效率提升、解放人类生产力与创造力
通往 AGI 和具身智能的关键环节
重塑互联网流量入口格局
预计入口级 Agent 大战将于 25H2 开启

三、 竞争格局：模型即产品，通用 Agent 将由大厂主导
Agent 领域的竞争维度
模型即产品：爆款应用背后是模型能力更新、浅层套壳产品终将被颠覆
Big Giants：角逐 AGI、通用 Agent 和流量入口
Niche market：垂直 Agent 长期面临通用 Agent 的威胁、垂直 Agent 的价值在于深耕领域知识、谈谈 Cursor 的壁垒

四、 Agent 将最先落地于知识工作（尤其是代码）
Agent 最先落地的行业和场景预测
代码/软件开发领域的进展与观点
法律 AI Agent 对比

五、 投资建议

六、 风险提示

# 一、为什么说2025年是Agent元年？

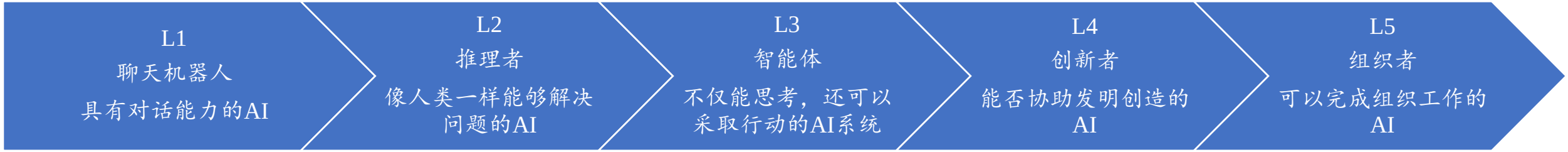
# 我们认为2025年将是Agent之年

**OpenAI将AI发展阶段分为L1到L5五个阶段。我们认为，**AI正从L2（推理者）向L3（Agent）进化，Agent代表了AI从“思考”走向“行动”的关键一步，是继大模型之后的下一个重要发展阶段和业界寻求的新突破口。驱动力来自：**技术、产品、需求。**

**L1 - 聊天机器人 (Chatbot)：**以ChatGPT（2022年底发布）为代表，具备自然语言交互能力。机器直接输出文字或回答。相较于机器学习时代，AI Chatbot 实现了“通用性”，不再局限于特定场景或单一问题，而是能处理广泛的语言任务。这是从基于规则、机器学习、神经网络、Transformer架构一路发展过来的通用大模型阶段。在这一阶段，**交互模式**是主要是输入-输出模式，用户提问，模型回答。

**L2 - 推理者 (Reasoner)：**具备更强的推理能力，能够处理更复杂的问题。用户能看见模型的推理过程。代表产品如OpenAI的o1系列、DeepSeek R1。相较于L1阶段，引入了强化学习和思维链（CoT）技术，模型在输出最终答案前会进行多步思考。

**L3 - 智能体 (Agent)：**能够自主规划和执行复杂任务的智能体。具备记忆、规划、工具使用和行为记忆四大核心能力。相较于L2阶段，AI从被动的“信息处理/推理”走向主动的“与外部世界交互和执行”。能调用工具（如浏览器、API）、操作软件界面，形成“指令->思考->交互->观察->再思考...”的闭环系统。




**ChatGPT的出现** **Deepseek R1标志着从L1到L2** **Operator标志着从L2到L3**

# Why Now？——技术成熟度达到临界点



**驱动力一：技术成熟度达到临界点。**支撑通用Agent发展的关键技术要素，特别是强大的多模态基础模型（能理解视觉信息如屏幕内容）和成熟的强化学习训练方法（能训练Agent与环境交互），已经发展到相对成熟的阶段。

- **从L0到L1：**标志GPT-3、ChatGPT（2022年底）为标志。背后的技术驱动力是Transformer架构的出现，使得训练更大、更通用的语言模型成为可能。在这一阶段，实现了“通用性”(Generality)，模型不再局限于特定场景，而是能够处理广泛的自然语言任务，像一个巨大的知识库。

**从L0到L1的技术演进路径**

| 基于规则 | 机器学习 | 神经网络 | transformer | GPT-3 | chatgpt |
|---|---|---|---|---|---|
| 依赖人工规则，灵活性差<br>智能处理的数据量较少 | 数据驱动，泛化能力增强<br>分类固定量数据 | 深度学习<br>复杂数据处理能力提升 | 自注意力机制<br>并行计算 | 预训练+微调<br>强泛化<br>自然语言生成能力 | 人对结果的反馈成为学习过程的一部分 |

# Why Now？ ——技术成熟度达到临界点

东吴证券 SOOCHOW SECURITIES

- **从L1到L2：**L1到L2的技术演进，核心在于大模型基础上的推理能力突破，涉及多步推理训练、检索增强、逻辑融合等关键技术，使AI从“会说”进化到“会想”，实现更高层次的智能。从L1到L2的跃迁，是AI从“语言表达”到“认知推理”的质变，这为AI在科学发现、复杂决策、自动规划等高价值场景的应用奠定了基础。技术突破包括：
  - 多步推理训练：通过链式思维（Chain-of-Thought, CoT）等方法，训练模型在给出答案前进行多轮、分步骤的推理。
  - 检索增强生成（RAG）：结合外部知识库，提升模型的事实一致性和推理深度，减少“幻觉”。
  - 更高质量的数据与反馈机制：采用专家数据、复杂问题集和强化学习等方式，持续优化模型的推理表现

| | L1 Chatbot | L2 Reasoner |
|---|---|---|
| 主要能力 | 自然语言生成 | 复杂推理与决策 |
| 技术核心 | 大规模transformer<br>预训练 | 思维链CoT<br>RAG<br>MoE等 |
| 代表模型 | GPT-3<br>ChatGPT | O1<br>Deepseek r1<br>Strawberry等 |

**AlphaGo Zero快速超越了Alpha Go**


资料来源：datahub，东吴证券研究所

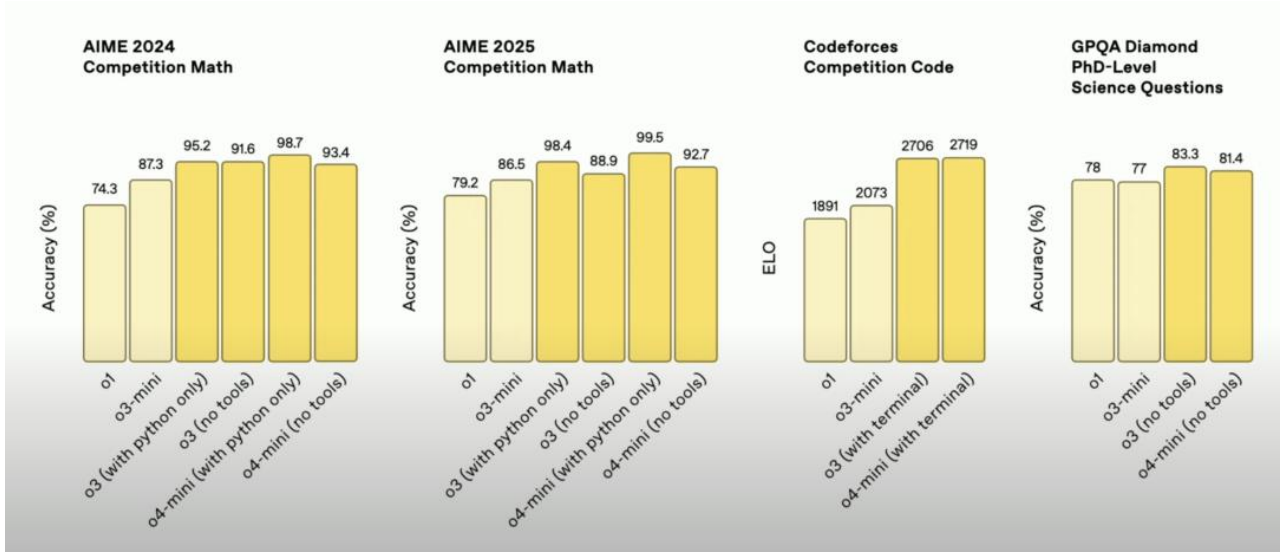
# Why Now？ ——技术成熟度达到临界点

东吴证券 SOOCHOW SECURITIES

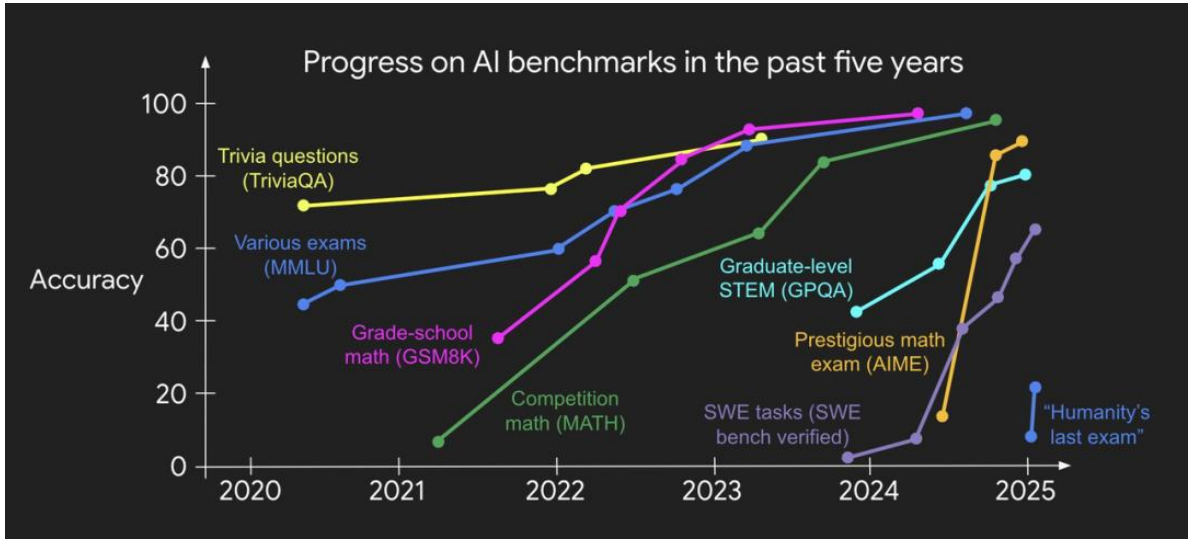
➢ **从L2到L3：关键的技术要素（强大的多模态基础模型和成熟的强化学习训练方法）已经趋于成熟，达到了可以支撑通用 Agent 发展的阶段。OpenAI在2025年1月发布Operator，更是印证和点燃了这一行业共识。具体来说，关键的成熟要素包括：**

- **强大的基础模型：**像Claude Sonnet 3.5这样强大的、原生的多模态基础模型已经出现。这些模型具备了足够好的视觉理解、语言理解和基础推理能力，能够“看懂”图形界面（如网页、操作系统界面），这是构建基于GUI（图形用户界面）的Agent的前提。而在过去（例如OpenAI在2016年尝试类似项目时），缺乏这样强大的基础模型是导致失败的关键原因。
- **成熟的强化学习技术与框架：**以强化学习为核心的 Post-training技术在2024年通过O1、O3等模型在纯文本领域被证明是极其有效的，能够显著激发和提升基础模型的深层推理和规划能力。行业将这种成功的范式应用到多模态领域，以训练出能够与环境交互、执行任务的Agent。

**o3 模型和 o4-mini 模型在数学和代码能力上表现出色**




**过去五年AI持续刷新各类排行榜**




资料来源：机器之心，澎湃新闻，东吴证券研究所

# Why Now？——行业领导者推动，标杆产品验证



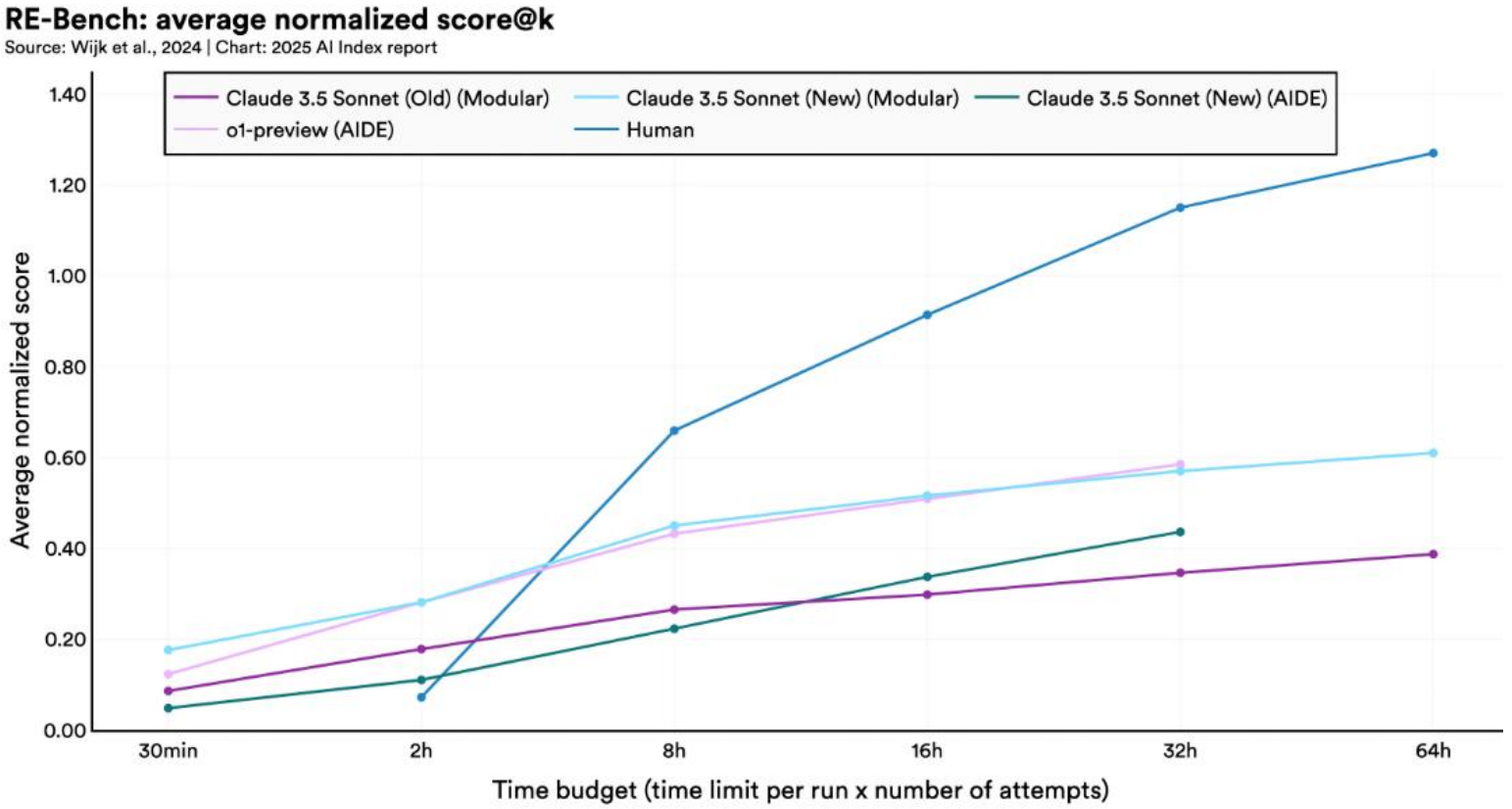
**驱动力二：行业领导者推动，标杆产品验证。** OpenAI、Anthropic、Google等头部公司发布关键产品（如Operator, DeepResearch）和技术协议（如MCP），并投入研发，起到了引领和示范作用。相对成型的Agent产品开始涌现（例如Manus、AutoGLM、Genspark等），验证了技术可行性，并点燃了行业共识，标志着Agent从**设想**走向**相对成熟**的产品阶段。

| 工具 | 底层模型 | 核心技术 | 自主性级别 | 多模态能力 |
|---|---|---|---|---|
| OpenAI Operator | 定制CUA模型 | 浏览器自动化、视觉理解 | 高（网页交互） | 强（视觉理解） |
| Manus | Claude Sonnet 3.7 | 多智能体架构、Linux沙盒 | 高（跨领域任务） | 强（文本、图像、代码） |
| Devin | 未公开 | 远程执行环境、规划系统 | 高（软件开发） | 中（主要文本和代码） |
| Cursor | 多个大模型 | 代码上下文理解、智能补全 | 中（辅助编码） | 弱（主要代码处理） |
| AutoGPT | 可定制LLM | 任务分解、互联网连接 | 高（自主执行） | 中（文本和图像） |
| Windsor.ai | 专有AI模型 | 数据归因、营销分析 | 中（数据处理） | 弱（主要结构化数据） |
| Deep Research | Gemini 1.5 Pro | 多步骤研究、网页浏览 | 中（研究执行） | 强（文本、图像、PDF） |
| ChatGPT Canvas | GPT-4 | 代码编辑、多文件管理 | 低（辅助编辑） | 弱（主要代码处理） |

资料来源：lamatic，智通财经，新智元，东吴证券研究所

# Why Now？——行业领导者推动，标杆产品验证

2024年的RE-Bench基准测试显示：在2小时短时限内，顶尖AI Agent得分是人类专家的4倍；但将时间放宽到32小时，人类表现则反超部分Agent。**这表明Agent在特定任务上已能匹敌人类专家，且更快、更经济，但人类在长时策略上仍有优势。**

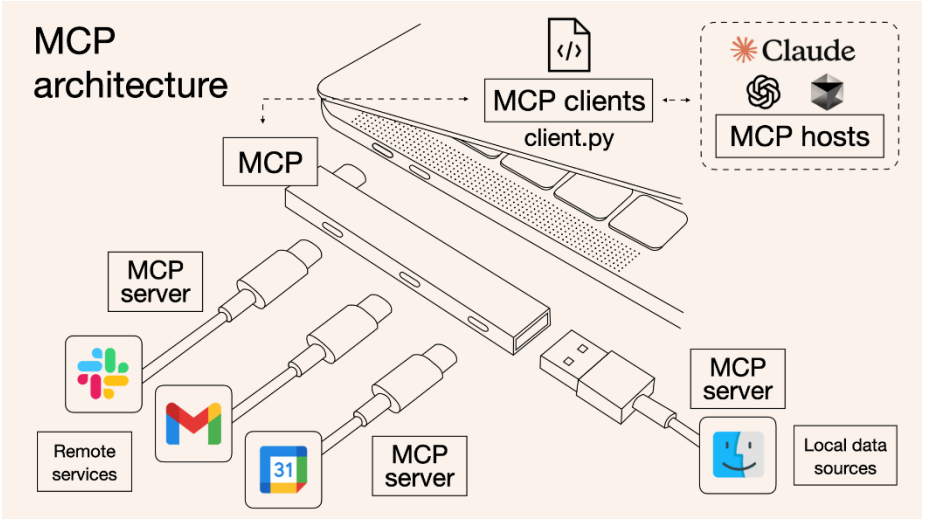


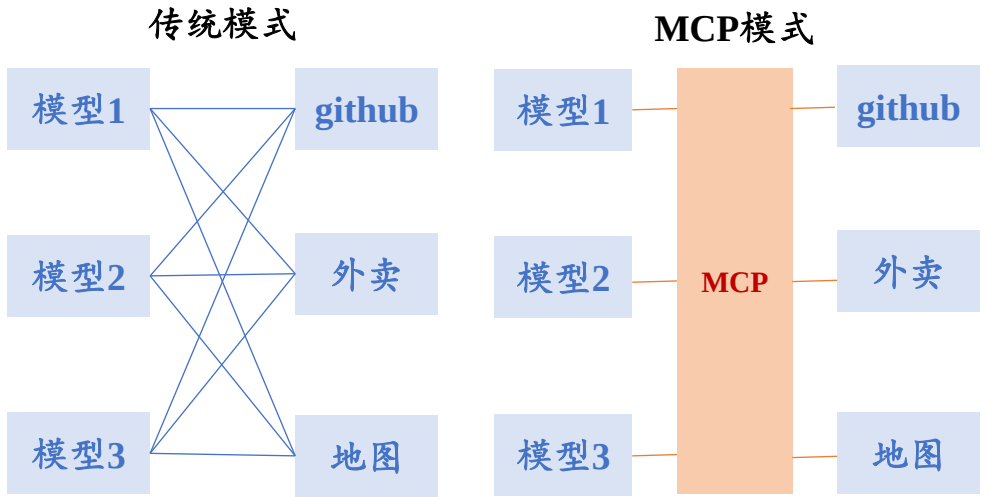
# Why Now？——MCP的普及助推Agent互联互通

在MCP出现之前，Agent 想利用外部工具或数据源（例如调用一个API、查询数据库、读取Slack消息、操作某个软件），面临着巨大挑战：接口各异、定制开发成本高、生态割裂。

**MCP的普及有助于推动Agent行业互联互通。**MCP（Model Context Protocol，模型上下文协议），是由Anthropic提出的一个开放协议，旨在统一大型语言模型（LLM）/Agent与外部工具、数据源之间的通信方式。MCP通过提供一个开放、统一的通信标准，可以解决Agent与外部世界交互的碎片化和高成本问题。它的普及将极大地降低集成门槛，增强不同模型、Agent和工具间的互操作性，催生出一个更加繁荣、开放和互联互通的Agent生态系统，最终赋能更强大、更通用的AI Agent应用。

| 在MCP出现之前agent的挑战 | MCP的普及有助于促进互联互通 |
| --- | --- |
| ①接口各异：每个工具、每个数据源都有自己独特的API接口或交互方式。<br>②定制开发成本高：Agent开发者需要为每一个想要连接的工具编写特定的适配代码，以理解该工具的输入输出格式和调用逻辑。同样，工具开发者如果想让自己的服务被不同的Agent调用，也可能需要适配多种不同的Agent框架。<br>③生态割裂：这种点对点的、定制化的连接方式，导致整个生态系统是割裂的。Agent A可能只能使用它专门适配过的工具集X，而Agent B只能使用工具集Y，它们之间难以共享或调用对方生态中的工具，形成了“数据孤岛”和“能力孤岛”。 | ①建立“通用语言”：MCP提供了一套标准化的规则和格式，定义了Agent（通过MCP Client）如何向工具（MCP Server）发出请求、传递参数，以及工具如何返回结果。这就像为AI Agent和外部工具之间建立了一种通用的“交流语言”。<br>②降低开发与集成复杂度：Agent开发者不再需要为每个工具编写定制化的适配器。只需要让Agent支持MCP协议，理论上就能与任何同样支持MCP的工具进行交互。工具/数据源提供者只需将自己的服务通过一个MCP Server暴露出来，就能被所有支持MCP的Agent发现和调用，降低了接入AI生态的门槛。就像USB-C统一了各种设备的物理连接和数据传输标准一样，MCP旨在统一Agent与工具的“数字连接”。<br>③促进互操作性，催化生态系统繁荣：当Agent和工具都遵循同一标准时，它们之间的互操作性大大增强。这意味着用户或开发者可以更自由地组合来自不同提供商的模型、Agent框架和工具，构建出更强大、更灵活的解决方案，打破了原有的供应商锁定或生态壁垒。标准化是生态繁荣的基础。MCP的普及将鼓励更多开发者参与Agents生态，形成一个更加开放、组件化、可互相协作的Agent生态系统。 |



# Why Now？——MCP的普及助推Agent互联互通



**和其他工具调用方式（Function Calling, A2A, Browser Use）相比，MCP的优势是什么？——通用性、互操作性、低门槛**

**1、Function Calling** 是 OpenAI 的早期尝试，开发者在调用 LLM API 时可以定义一组可用的函数（工具）。当用户需要执行某个功能时，模型不会直接执行，而是会输出一个包含函数名和所需参数的 JSON 对象。开发者接收到这个对象后，自己编写代码去执行相应的函数，并将结果返回给模型，让模型继续生成回复。

**Function calling 的缺点是**，没有定义一个通用的、跨平台、跨模型的标准，每个开发者都需要根据 OpenAI 的规范来实现。

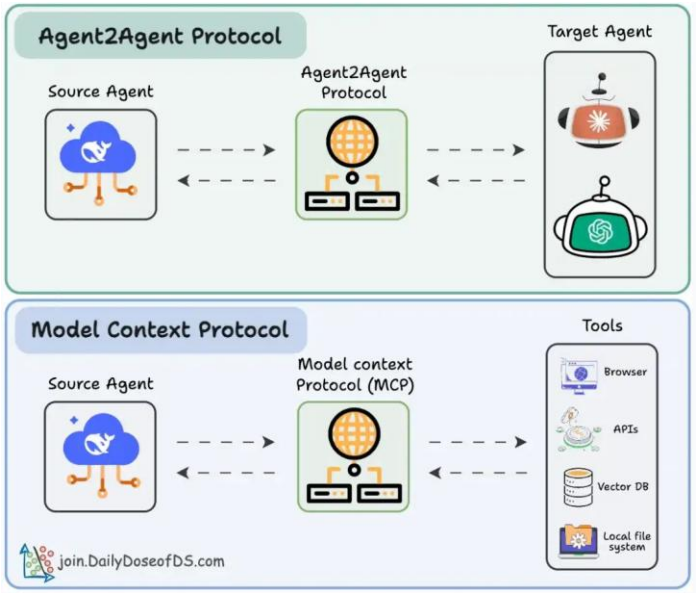
**2、MCP** 旨在建立一套通用的协议或规范，来定义 Agent 如何发现、理解和调用各种工具，以及工具如何返回结果。

相较于 Function Calling，**MCP的优势在于：①统一度量衡：** MCP 就像是为工具调用设定了国际标准（如米、千克），取代了之前各种自定义、不兼容的“度量方法”（类似 Function Calling 的非标准化状态）。**②互操作性 & 降低门槛：** 有了统一标准，开发者开发的 Agent 可以更容易地调用任何遵循 MCP 规范的工具，反之亦然。工具开发者只需支持 MCP，就能被众多 Agent 使用。这大大降低了工具集成和生态构建的门槛。

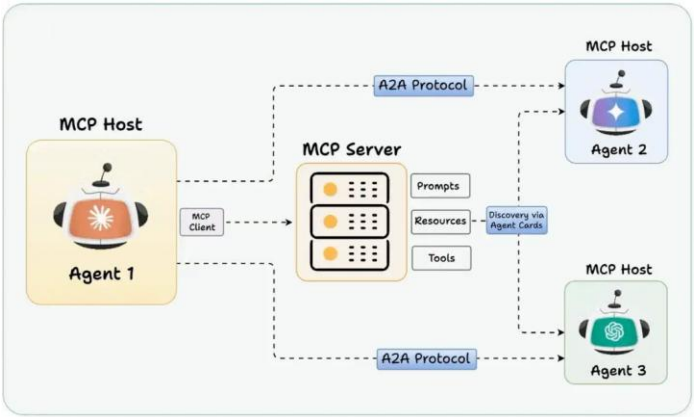
**3、A2A (Agent-to-Agent)** 是Google 提出的概念，声称不仅能让 Agent 调用工具（Tool），还能实现 Agent 与 Agent 之间的直接交互。

**但我们认为这其实是概念冗余：**从工程角度看，一个 Agent 本身也可以被封装成一个符合 MCP 规范的 Tool。因此，通过 MCP 协议，已经可以间接实现 Agent 调用另一个 Agent（作为工具）。A2A 并没有带来根本性技术突破，更像是一种“KPI 工程”或争夺标准化话语权的战略行为，而非必要的技术创新。

**A2A的原理和MCP类似**




**A2A可以融入到MCP框架中**



# Why Now？——MCP的普及助推Agent互联互通



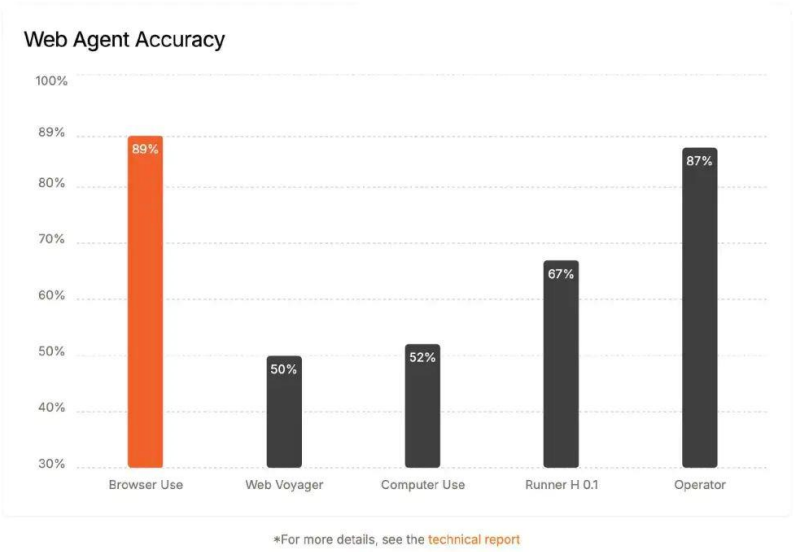
**和其他工具调用方式（Function Calling, A2A, Browser Use ）相比，MCP的优势是什么？——通用性、互操作性、低门槛**

**4、Browser Use** 让 Agent 能够像人一样操作浏览器，浏览网页、提取信息、填写表单、点击按钮等。但需要明确的是，Browser Use和MCP并不互斥。Browser use的驱动方式分为两种：MCP驱动和GUI操作。

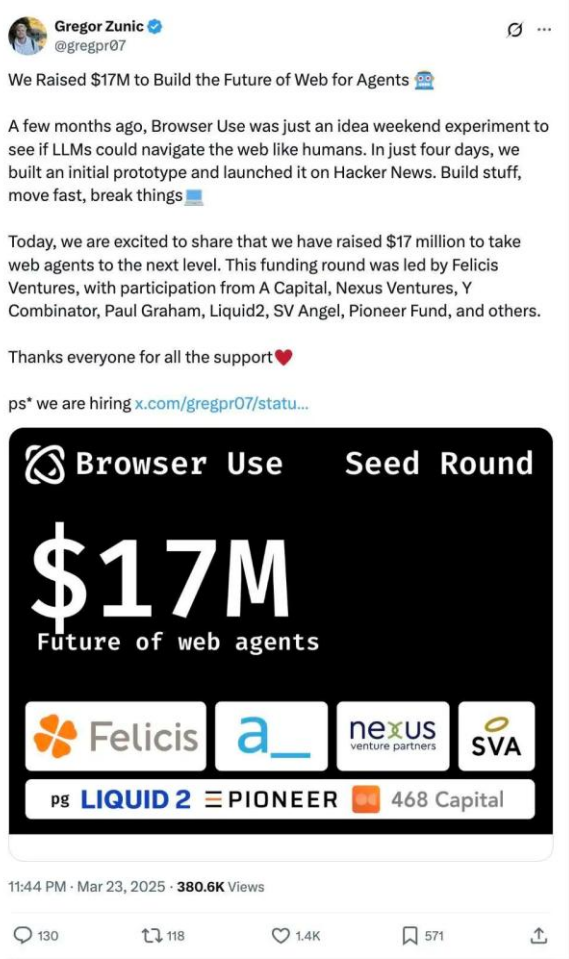
- **MCP驱动（更成熟、常用）：**这并不是让 AI 真的“看”屏幕去点。而是通过调用浏览器提供的 API（例如，获取网页 DOM 结构、执行 JavaScript、模拟网络请求等），或者将这些浏览器操作封装成符合 MCP 标准的工具，然后让 Agent 通过代码调用这些工具来间接“操作”浏览器。现在很多所谓的 Browser Use 演示，其背后很可能就是这种基于代码/API/MCP 的方式。代表产品有Browser User和Manus。

- **GUI操作（尚不成熟）：**这是真正意义上的“看屏幕、点鼠标”。Agent 接收浏览器窗口的截图，通过视觉模型识别界面元素（按钮、输入框等），计算出坐标，然后通过模拟鼠标点击和键盘输入来进行操作。这种方式目前面临准确性和稳定性的瓶颈，因为视觉模型在精确识别和定位界面元素（尤其是动态或复杂的网页）时容易出错，导致点击错误位置或无法完成操作。

**Browser Use的准确率较高**




**初创产品Browser Use，由两名学生在4天时间内开发完成，可以理解网页内容。该初创公司2025年3月融资1700万美元。**



# Why Now？——市场需求驱动



**驱动力三：市场需求驱动。**

回顾AI发展历程，如果说2023年是“模型竞赛年”（以LLM本身性能竞赛为标志），那么2024年则是“应用探索年”。在2024年，涌现了大量基于LLM的应用，例如各种聊天机器人、写作助手、简单的Copilot等。企业投入资源进行尝试，希望将AI能力融入业务流程。

然而，2024年的应用探索也暴露出一些局限性。许多应用可能只是**“薄封装”**，未能深入解决核心业务痛点；或者其自动化能力仅限于相对简单的单点任务，难以应对跨系统、多步骤的复杂工作流；带来的生产力提升往往是局部的、渐进式的，未能完全达到市场最初的高期望，也使得AI投入的ROI不够清晰。

**进入2025年，市场心态发生了转变，特别是对于需要为AI投资寻求明确商业价值的To B而言：**

①从“尝试”到“落地”：企业不再满足于概念验证（PoC）或小范围试点。他们需要能够**真正部署到生产环境中、稳定可靠、能与现有系统集成、并产生可衡量业务成果**的AI解决方案。市场渴望看到AI技术从“玩具”或“助手”变成真正能干活、能解决问题的“员工”或“自动化引擎”。

②渴望自动化“更复杂任务”：简单的问答、基础的内容生成等“低垂果实”已被初步采摘。企业现在关注的是那些**更耗时、更繁琐、涉及多个步骤、需要调用不同工具或信息源的复杂流程**。例如，自动完成一份包含数据搜集、分析、图表生成和报告撰写的市场研究报告；或者自动化处理一个需要查询订单系统、物流系统、与客户沟通并执行退款操作的客服请求；亦或是完成整个软件开发周期中的部分环节。这些是传统自动化或简单AI应用难以触及的领域。

③期待“更显著”的生产力提升：市场不再满足于10%或20%的效率提升。他们期待的是**数量级（例如数倍甚至更高）的生产力飞跃**，能够真正重塑工作方式、显著降低成本、或者将人力解放出来从事更高价值的创造性或战略性工作。

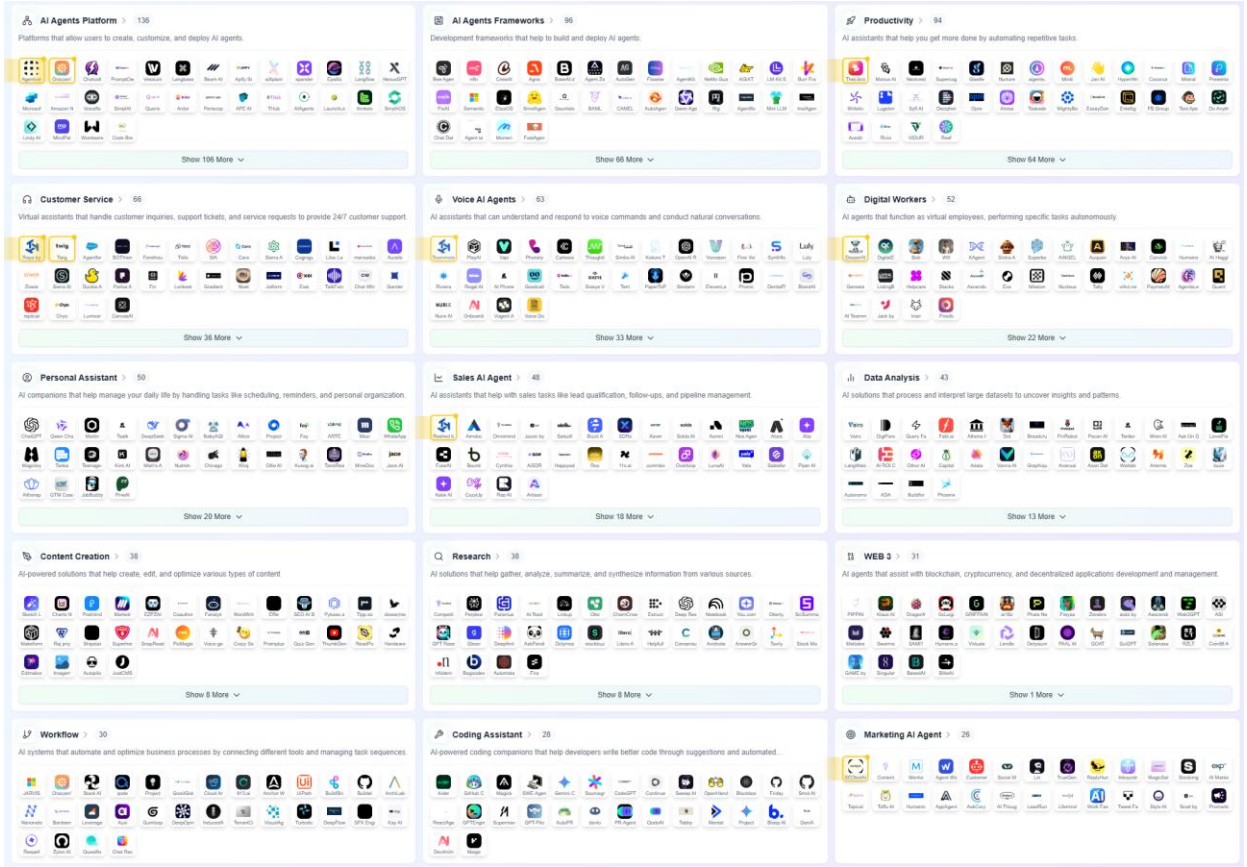
**而AI Agent（智能体）的出现，恰好精准地契合了市场的这种新期待**：①为复杂任务而生：Agent的核心能力（如自主规划、记忆、工具使用）使其天然适合处理**多步骤、需要与外部环境（如网页、软件、API）交互的复杂任务**，这正是市场所需要的。②强调“执行”与“行动”：不同于L1/L2主要停留在“对话”或“推理”，L3 Agent的设计目标就是**完成任务、采取行动**，这与企业追求“落地”和实际效果的需求高度一致。③潜力巨大：通过自动化更复杂、更耗时的工作流，Agent有望带来**指数级的效率提升和生产力解放**，满足市场对“显著”价值回报的期待。

# 定义：不是所有的AI模型/产品都叫Agent

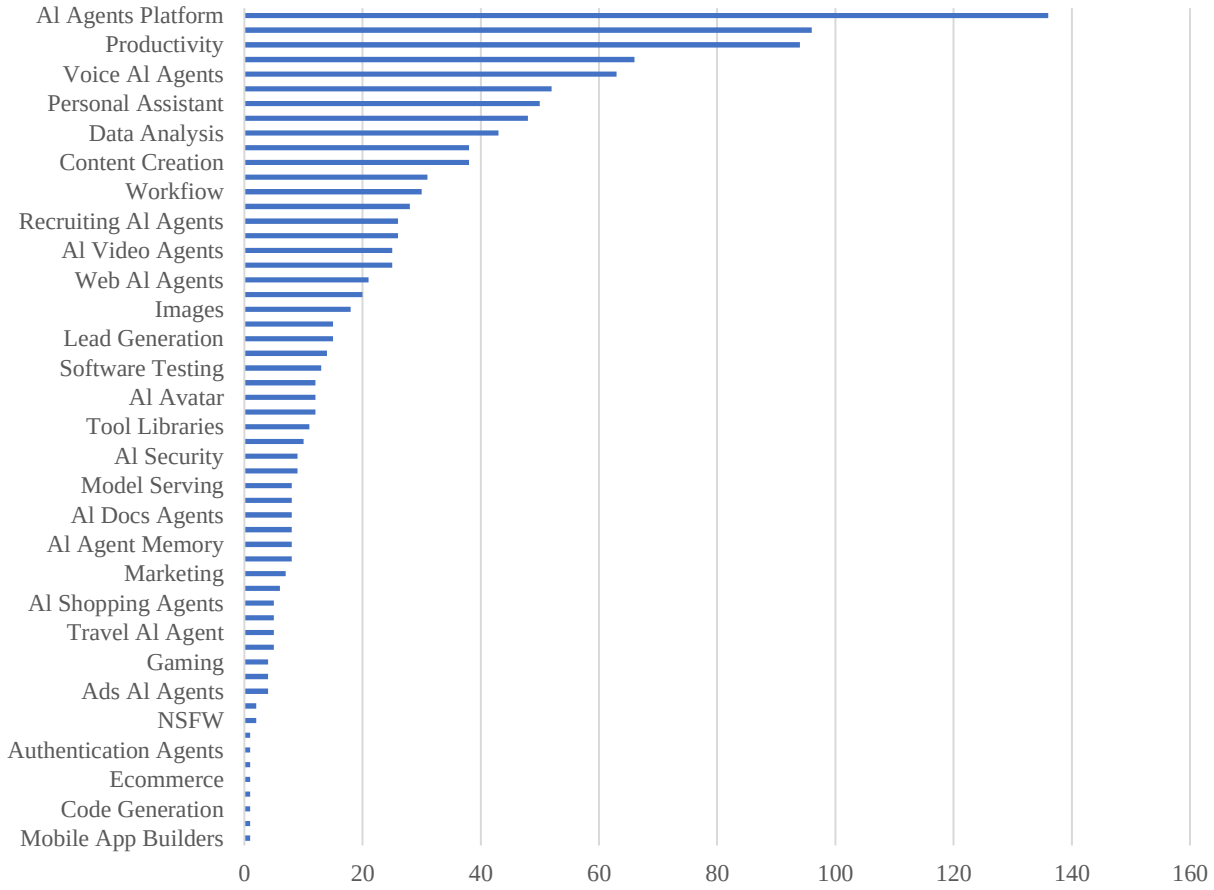
东吴证券 SOOCHOW SECURITIES

**当前存在着大量的垂类Agent。根据AI Agents Directory统计，截至2025/4/7，全球共有1211个AI Agent，覆盖57个不同类别，其中数量较多的有Agent平台（136个）、生产力Agent（94个），客户服务Agent（66个），个人助手Agent（50个）等。虽然部分应用可能并不属于严格意义上的Agent（需要有调用工具的能力和规划执行的能力等），但也能直观上反映当前应用生态的复杂多样。然而，这些都能被称之为Agent吗？**

**AI Agent landscape**




**按照行业分类的AI Agent数量（截止25/4/7）**




资料来源：AI Agents Directory，东吴证券研究所

# 定义：不是所有的AI模型/产品都叫Agent

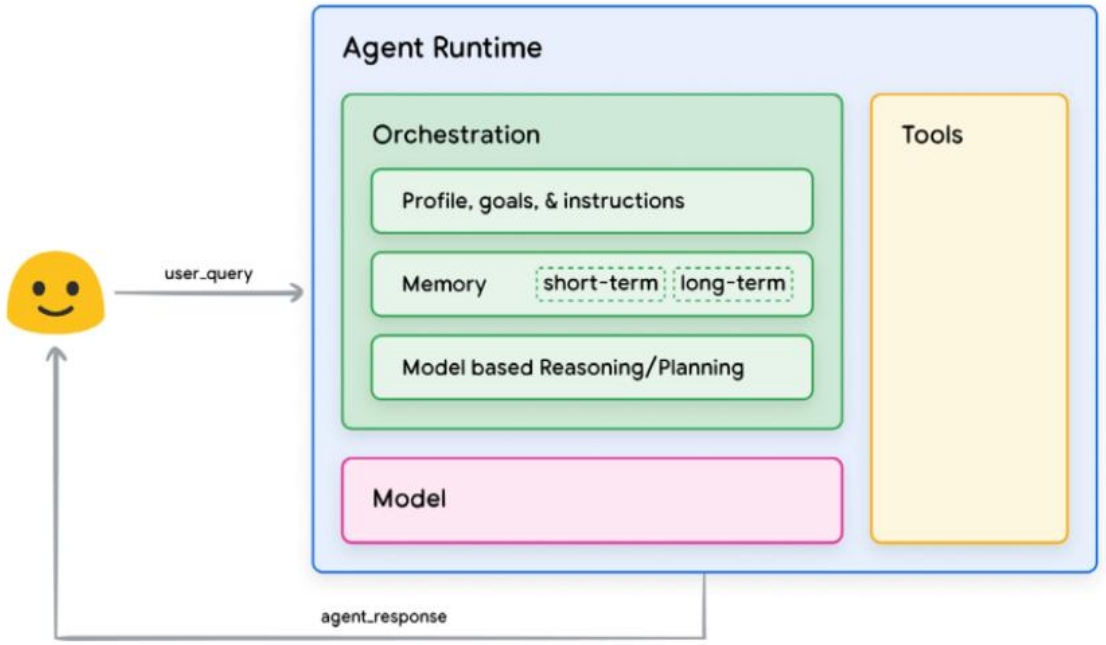
关于Agent的讨论往往存在定义混乱的问题。以至于一千个人眼中有一千个Agent。

**我们认为，只有同时具备了这四项能力（对话能力、推理能力、长记忆、工具调用），才能被称之为Agent。**其中，工具调用是最核心的区分要素。只有对话能力的是 Chatbot；只有对话和推理能力的是 Reasoner。

而工具调用又是建立在前三个基础之上的。Agent必须理解用户的指令，记住过去对话的内容，记住其任务目标、分解动作和已执行的步骤、遇到的问题，才能顺利地完成多步骤任务。

| | Chatbot | Copilot & Assistant | 半自动Agent | 全自动Agent |
|---|---|---|---|---|
| 对话能力 | √ | √ | √ | √ |
| 推理能力 | √ | √ | √ | √ |
| 长记忆能力 | | √ | √ | √ |
| 调用工具的能力 | | | √ | √ |
| 规划的能力 | | | | √ |

**Agent的构成**



# 定义：Agent的智能程度是有层次和梯度的



**尽管如此，我们仍然无法准确定义AI Agent，例如：**

- 一个 AI 系统仅仅能响应指令、生成内容就够了吗？还是要看它是否能为了达成某个特定目标而主动采取一系列行动？
- Agent 的“行动”是否必须对外部世界（数字或物理）产生状态改变？生成信息、报告或建议算不算定义中的“行动”？
- 这个系统是只在内部进行计算和推理，还是需要感知外部环境的状态，并能对环境施加影响（无论是数字环境还是物理环境）？
- 在执行任务的过程中，系统是严格按照预设步骤执行，还是能够根据当前情况自主进行决策、选择策略或调整计划？
- 需要多大程度的独立决策和执行能力才能称之为 Agent？需要人类确认或干预到什么程度就不再是（完全自主的）Agent？
- 系统完成任务是仅靠自身内置的知识和能力，还是需要识别并调用外部的资源或工具（如 API、数据库、其他软件）来辅助完成？
- Agent 交互的“环境”必须是动态的、不可预测的吗？与一个静态数据库交互算不算环境交互？
- 系统处理任务是一次性的“问答”或“生成”，还是能够在持续一段时间内保持对目标和上下文的认知（记忆），以完成需要多个步骤或较长时间才能完成的任务？
- Agent 的“大脑”是什么？它与底层的LLM是什么关系？Agent 是 LLM 本身，还是一个围绕 LLM 构建了规划、决策、执行框架的系统？

**这些问题其实是同一个问题，即，Agent需要智能到什么程度，才可以被称之为Agent？**

**再比如：**

- AI 搜索（如Perplexity、DeepResearch、New Bing）是Agent吗？一个能理解复杂问题、自主上网搜索、阅读并整合信息，最终生成一份摘要报告或直接答案的 AI 搜索系统，是 Agent 吗？它“使用”了浏览器或搜索引擎作为工具，并“行动”生成了报告，这是否足够？如果这个 AI 搜索系统只是呈现整合后的信息，而没有根据这些信息去执行下一步的、改变外部状态的动作（比如基于搜索结果去预订、购买或发送邮件），它与一个高级的L2 Reasoner的核心区别是什么？Deep Research 这类工具，其“Agent”属性体现在哪里？仅仅是研究过程的自动化吗？
- AI 编程（如Github Copilot、Cursor、Devin）是Agent吗？GitHub Copilot根据代码上下文提供建议，开发者选择采纳。它有环境感知（代码上下文），也有行动（生成代码建议），但自主性较低，它算 Agent 吗？还是更像一个“智能感知代码的 L1 模型”？
- AI推荐系统是Agent吗？一个能分析你的历史行为、理解你的偏好，并主动推送（行动）相关内容或商品的推荐引擎，它具备目标（提升用户参与度/转化率）、环境感知（用户行为数据）和行动（推送），它算 Agent 吗？它的“自主性”和“规划”体现在哪里？

# 二、Agent为何重要？解放生产力、走向AGI和具身智能、挑战入口格局

# 2.1 Agent将带来革命性的变化——深度自动化

AI Agent作为下一代AI应用形态，将带来革命性的变化，远超简单的信息检索或内容生成。Agent是能够主动执行任务、解决问题的数字化劳动力或超级助理，其核心价值体现在：

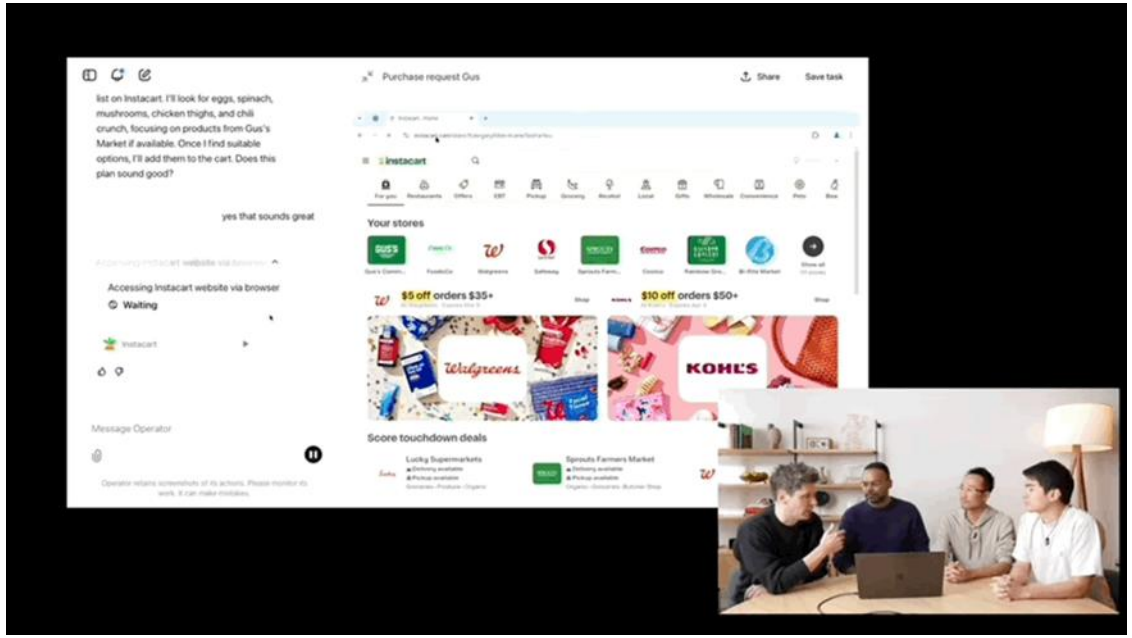
①深度自动化：

超越简单重复：不同于RPA或传统脚本主要处理固定流程的重复性任务，Agent能够理解模糊指令，自主规划并执行复杂的、多步骤的、甚至需要适应变化的认知型任务。例如，Operator能模拟人类操作任意GUI界面完成预订或购物，Devin能自主完成软件开发中的编码、调试、测试等系列环节。

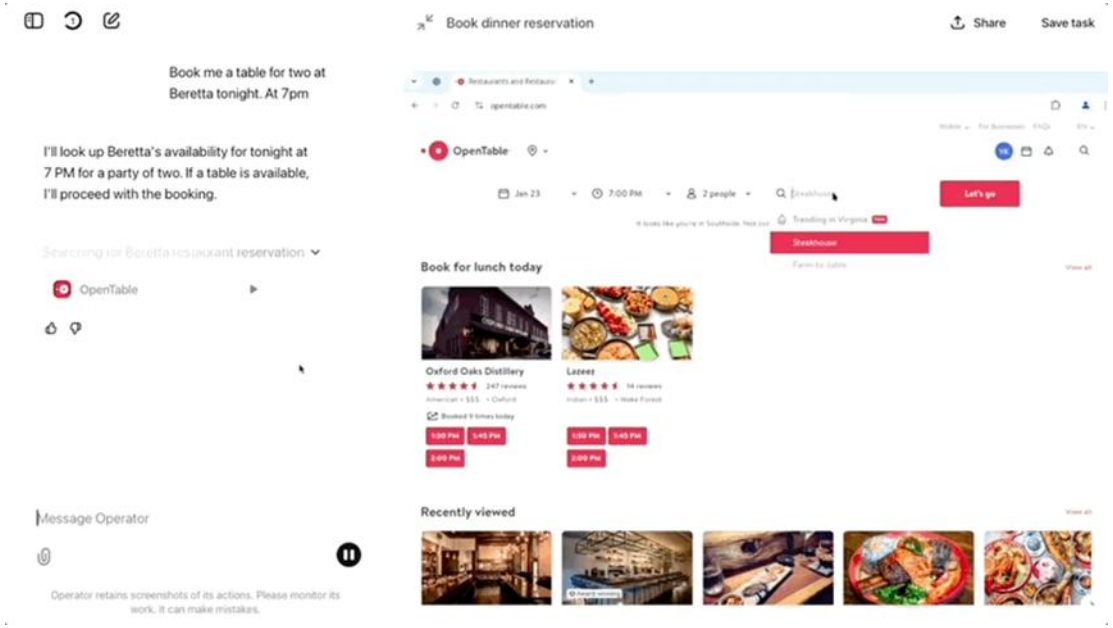
端到端流程：Agent有潜力打通原本需要多个人类角色、多个软件系统协作才能完成的端到端工作流，实现更高层次的自动化。

认知自动化：其核心是自动化需要思考、判断、与数字世界交互的“知识工作”，而不仅仅是体力或简单的点击操作。

**用operator采购商品**




**用operator自动预定餐厅**




资料来源：openai，东吴证券研究所

# 2.1 Agent将带来革命性的变化——指数级效率提升



②指数级效率提升：

- 速度与规模：对于特定任务，Agent的处理速度可以远超人类（如RE-Bench短时限测试表现）。更重要的是，Agent可以7x24小时不间断工作，并且可以轻松扩展（理论上增加算力即可增加Agent数量），实现人力无法比拟的规模化效率。
- 成本优化潜力：虽然当前推理成本较高，但通过自动化高价值、高成本的人类劳动（尤其是专业知识工作，如软件开发、法律咨询），长期来看具有巨大的成本节约潜力。一个高效的Agent理论上可以替代或增强多个人类员工的生产力。
- 减少错误与提升一致性：对于定义清晰的任务，Agent有望减少人为错误，提高执行的一致性和标准化水平（尽管当前可靠性仍是挑战）。

③解放人类生产力与创造力：

人机协作新范式：Agent不仅仅是替代，更是强大的增强工具和协作伙伴。它们可以承担复杂流程中繁琐、耗时的部分，让人类专家（开发者、研究员、律师等）从重复性劳动中解放出来。

聚焦高价值活动：人类可以将时间和精力投入到更需要创造力、战略思考、复杂决策、情感沟通和人际协作等AI尚不擅长的高阶任务上。

赋能创新：通过自动化原本难以完成或成本过高的复杂分析与操作，Agent可能催生新的科学发现、商业模式或艺术创作，拓展人类能力的边界。

| | 自动化重复性、流程化的数字/知识工作 | 软件开发与编程辅助 | 垂直领域的专业Agent |
|---|---|---|---|
| 代表产品 | OpenAI Deep Research, Perplexity | GitHub Copilot, Cursor, Devin | 如营销/人力资源等行业的Agent |
| 具体能力 | • 信息研究与报告生成：自动搜集、整理、分析信息并生成报告，辅助研究人员、分析师等知识工作者。<br>• 操作软件和网页：自动执行需要与软件界面或网页交互的任务，如填写表单、预订差旅、处理邮件、管理日程、关闭广告、计算退款等。<br>• 数据处理与分析：自动执行数据提取、清洗、初步分析等任务。 | • 代码生成、补全、调试：提升开发者效率。<br>• 复杂开发任务执行：能够理解需求、规划步骤、编写代码、配置环境、测试、修复 Bug 等更完整的开发流程。<br>• API 调用与集成：Agent 利用编码能力与其他系统或服务交互。 | • 客户服务：处理标准化的客户请求，如查询订单、处理退款等。<br>• 销售/市场营销：自动化部分销售流程，如潜在客户筛选、邮件营销等。<br>• 人力资源：辅助处理简历筛选、安排面试等流程化任务。<br>• 特定行业：如法律文书辅助、医疗信息查询与初步分析等 |

# 2.2 Agent（L3）是通往AGI的关键环节

AI 的发展遵循一个从简单到复杂的层级结构，通常参考 OpenAI 提出的 L1 到 L5 框架。**Agent (L3) 是承上启下的关键阶段**。它不仅需要 L2 的推理和规划能力，更核心的是增加了与外部世界（数字世界或物理世界）交互的能力，形成闭环系统。这与 L1/L2 主要停留在与人交互或纯粹内部思考不同。

**虽然 L1 到 L3 的路径相对清晰，但从 L3 (Agent) 到 L4 (创新者) 存在一个巨大的鸿沟。区别在于：**

L1-L3 本质是**遵循指令 (instruction following/execution)：** AI 的主要任务是理解并完成人类给定的指令或目标。评价标准相对明确（任务是否完成，结果是否正确）。而 **L4 (Innovator) 要求创造力与原创性**: 它需要能够**超越指令 (beyond instructions)**，产生新的想法、方法、知识，甚至设定新的目标。评价标准变得模糊，不再是简单的“对不对”，而是“好不好”、“新不新”。

虽然终极目标是 AGI，但**短期内** Agent 通往 AGI 的路径体现在其自动化复杂任务的能力上。无论是自动化研究信息收集（只读 Agent），还是自动化软件操作、任务执行（读写 Agent），核心都是**解放人类的注意力**，提升生产力。通过不断提升自动化水平和处理任务的复杂度，Agent 的能力逐步逼近甚至超越人类特定领域的水平，为**最终**实现AGI奠定基础。

# 2.2 Agent（L3）是走向具身智能的关键环节



当前以“只读”型 Agent（如 AI 研究助手）为代表的应用已初步展现出明确的产品市场契合点（PMF），主要服务于知识工作者。

下一步的关键是从“只读”进化到“读写”型 Agent，即赋予 AI 执行操作、调用工具（如浏览器、邮件客户端、API）、与外部世界交互的能力（如 OpenAI Operator、Monica 的探索）。

虽然“读写” Agent 潜力巨大（能自主完成订票、发邮件、甚至发布悬赏任务等复杂操作），但其发展会更谨慎，因为涉及安全、权限和潜在风险，需要配合监控、对齐和防滥用措施。

随着记忆（Memory）和在线学习（Online Learning）这两大关键技术的突破，Agent 的能力将进一步飞跃，可能实现 Agent 指挥 Agent、更个性化、能实时学习适应新情况。

未来可能出现为 AI 设计的专用工具，进一步提升其效率，超越人类工具的限制。

Agent 的普及将极大解放人类注意力，从重复性工作中解脱，可能带来生产力的指数级增长，改变工作和生活方式。

**目前，绝大多数 AI Agent 主要活跃在数字世界中。这是因为数字世界**

- 环境结构化: 数字环境（如网页、软件界面、API）通常具有相对清晰的结构、明确的输入输出规则和可预测性。
- 信息易获取: 数据以文本、代码、图像等形式存在，相对容易被模型理解和处理。
- 行动成本低/可逆: 数字操作（如点击、输入、调用 API）通常成本低廉，且很多操作是可撤销或影响有限的。

然而，这种数字世界的局限性也很明显：AI 的能力被束缚在屏幕和网络之内，无法直接感知和影响我们生活的物理现实。真正的通用智能必然要求能够理解并作用于物理世界。

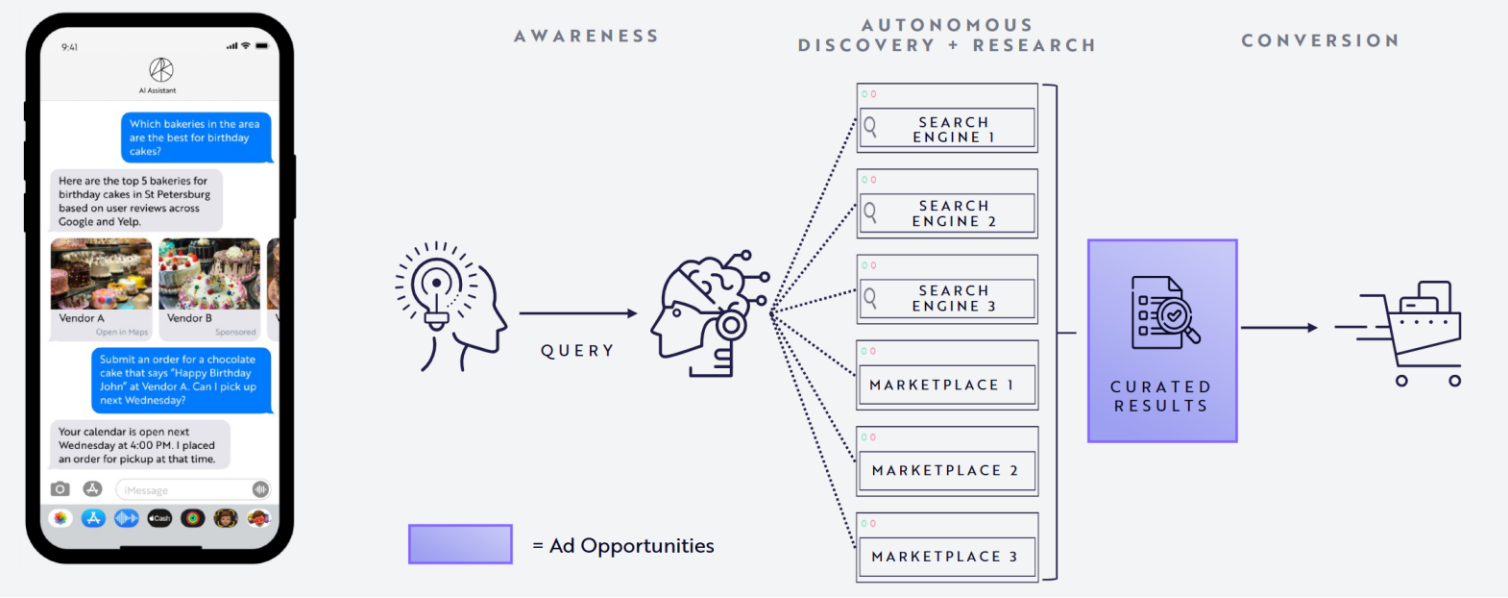
# 2.3 Agent将重塑互联网流量入口格局

**我们认为，AI Agent的发展可能对现有的互联网入口格局产生深远甚至颠覆性的影响。入口可能更加集中，价值链可能重构：**可能出现少数几个主导性的通用Agent。传统依赖流量分发的入口（如搜索引擎、应用商店）面临挑战，能直接完成任务或提供核心能力的Agent平台和服务商可能获得更大价值。

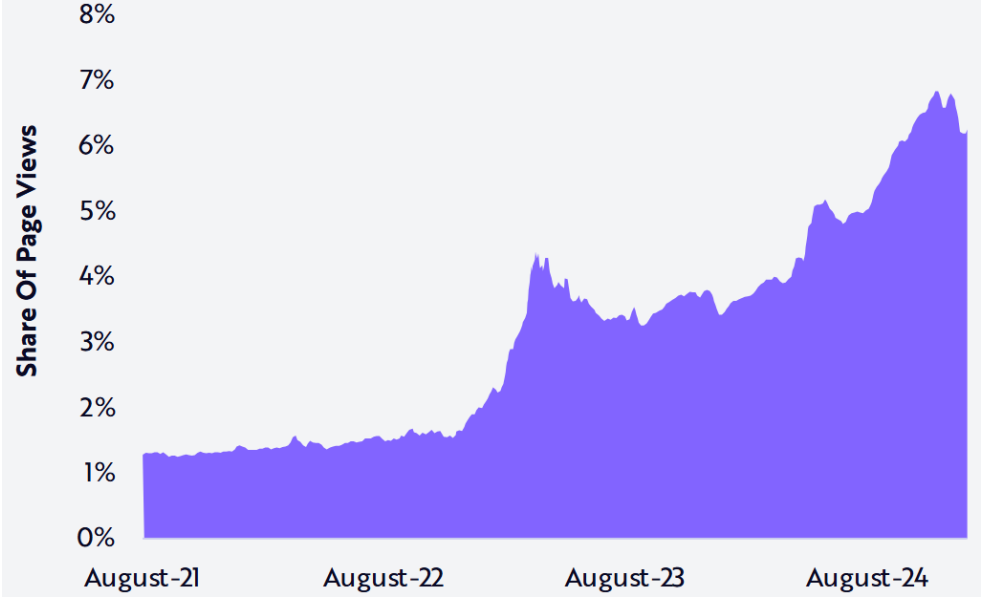
**1. 对传统搜索引擎的挑战与重塑：**

- **信息获取方式改变：**用户可能不再需要通过关键词搜索，然后浏览一堆链接来寻找答案或服务。Agent可以直接理解用户的复杂意图（例如“帮我规划一个周末去杭州的旅行，包含交通、住宿和景点，预算2000元”），然后整合信息、调用工具（订票、订酒店API）、进行规划，并直接给出完整方案甚至完成预订。这大大削弱了传统搜索引擎作为信息“门户”的角色。
- **搜索即执行：**Agent将搜索从“查找信息”升级为“完成任务”。入口的价值不再仅仅是分发流量到其他网站，而是直接满足用户的最终需求。像Perplexity、Google的AI Overviews以及Deep Search/Research，都体现了这种趋势——搜索结果本身就是答案或解决方案的一部分。

**AI聚合电商信息**



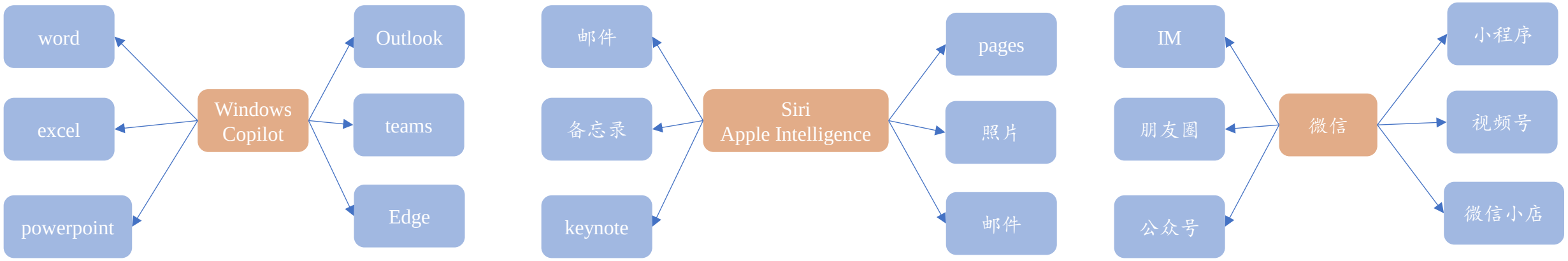

**ChatGPT, Claude, Perplexity和Bing的搜索份额**



# 2.3 Agent将重塑互联网流量入口格局

2. **Agent本身成为入口：**最具颠覆性的可能是，一个或多个强大的通用Agent成为用户上网和处理数字任务的首选入口。用户首先想到的是“问Agent”，而不是“打开某个App或网站”。这样的通用Agent有可能是APP、浏览器或者OS操作系统。
   - **OS级Agent：**操作系统（如Windows Copilot、未来更强大的Siri/Google Assistant）可以深度集成Agent能力，协调控制设备上的各种应用和数据。用户可能直接通过OS层的Agent下达指令，Agent负责调用合适的App或服务来完成，使得OS本身成为一个更核心、更主动的交互入口。
   - **浏览器集成Agent：**浏览器作为访问Web的主要工具，集成Agent可以辅助用户浏览、总结网页、写作、甚至自动化某些网页操作。这让浏览器从一个被动的页面加载器，变成一个主动的智能助手，增强了其入口地位，例如Edge Copilot、夸克、豆包（也有浏览器功能）。
   - **“超级App”入口地位巩固：**对于像微信这样的超级App，如果能成功集成强大的Agent能力，并打通其内部丰富的小程序、服务和社交关系，它可能成为一个极其强大的、覆盖生活方方面面的Agent入口。用户在一个App内就能完成大量任务。

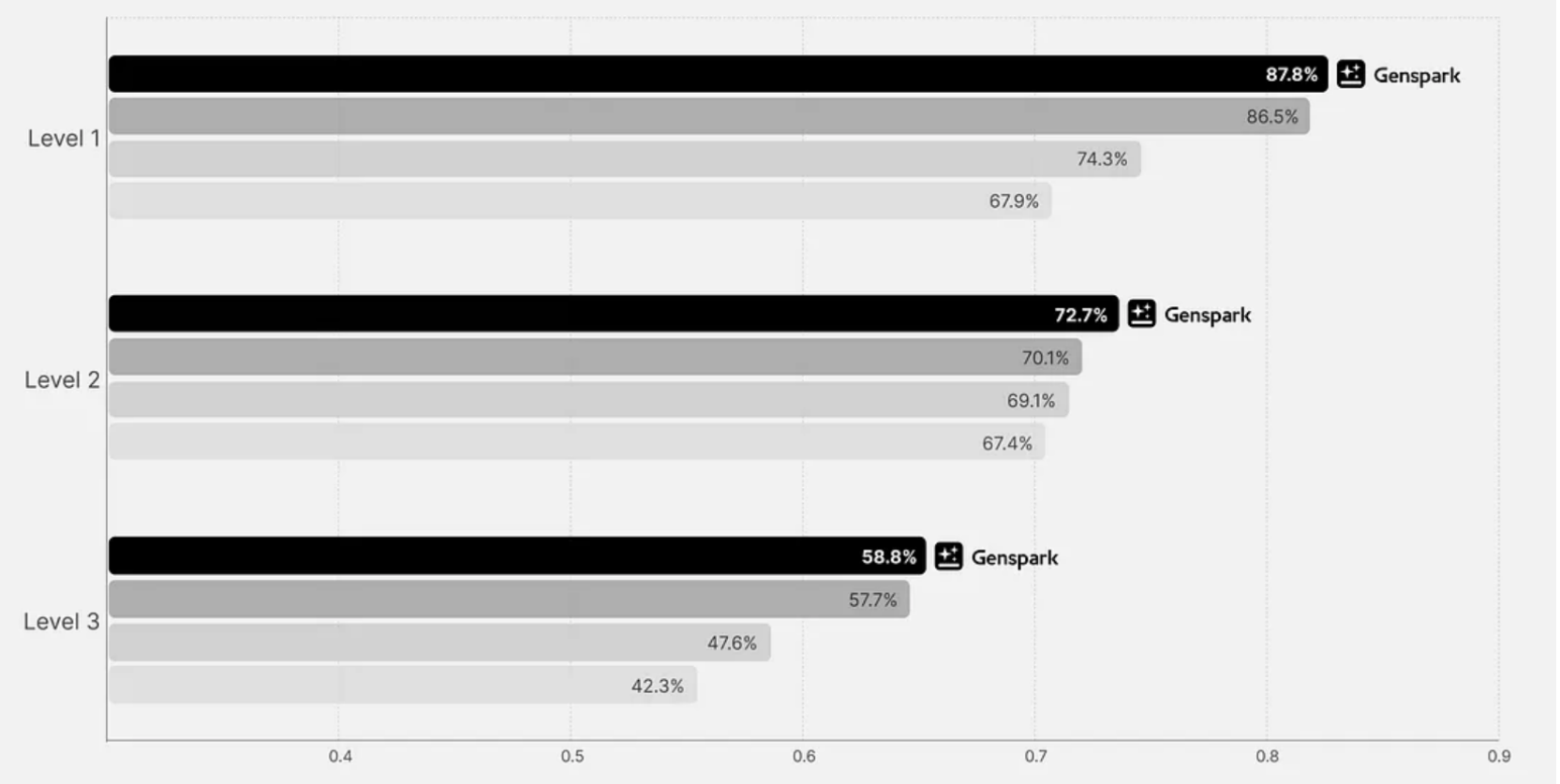
3. **部分APP被“管道化”：**一些功能单一的服务型App（如订票、打车、天气），用户可能不再需要直接打开它们，而是通过Agent来调用其背后的API或功能。这些App可能沦为Agent的“工具箱”，其自身的入口价值下降。**而复杂或体验型App，仍是入口：**对于需要沉浸式体验或复杂交互的应用（如游戏、专业创作工具、内容社区），用户可能仍然会直接打开App。但Agent也可能在这些App内部提供辅助功能。



# 2.3我们预计入口级Agent大战将于25H2开启

**我们预计，围绕“通用入口级Agent”的大战将在2025年下半年开始逐渐拉开帷幕**。为什么？因为L3级别Agent能力（能够系统性完成复杂任务）可能在一年内取得关键突破，一旦技术相对成熟，能够支撑起一个足够“通用”的Agent体验，各大有实力的玩家就会下场争夺市场主导权。

**Agent在L3的评分有望进一步突破**




资料来源：Genspark，东吴证券研究所

# 三、竞争格局：模型即产品，通用Agent将由大厂主导

# AI Agent的竞争格局是典型的“巨头环伺，新锐突围”



**AI Agent的竞争格局是典型的“巨头环伺，新锐突围”。**大型科技平台凭借模型、数据、资金和生态优势占据主导地位，并致力于构建平台和生态系统。然而，在基础设施、工具链以及需要深度领域知识的垂直应用方面，仍然存在创业公司和专业厂商的机会。成本、效率、交互体验和商业模式的创新将是未来竞争的关键。后续续密切关注技术演进、平台战略以及细分市场的动态。

**我们预计Agent领域的竞争将围绕六个维度展开**

| ■ 平台 vs. 应用 | ■ 通用 vs. 垂直 | ■ 成本与效率 |
| --- | --- | --- |
| 价值将主要沉淀在拥有核心模型和平台的巨头，还是能够创造独特价值的Agent应用开发商？<br>目前看平台方优势明显，平台公司倾向于将其Agent能力与其现有云服务、操作系统、办公套件等深度绑定，构建封闭或半封闭生态，增加用户迁移成本。<br>但应用层创新仍有机会。 | 通用Agent（如Operator）能力不断提升，是否会挤压垂直Agent的生存空间？<br>短期内，垂直Agent凭借领域知识仍有优势；<br>但长期看，通用Agent的泛化能力是巨大威胁。 | 推理成本是Agent大规模商业化的关键瓶颈。<br>模型效率、训练/推理优化、以及芯片成本将是重要的竞争维度。 |
| **■ 交互范式之争** | **■ 数据与护城河** | **■ 人才竞争** |
| Operator代表的直接GUI操作与Manus代表的“可见性”过程展示，以及未来可能出现的更优化的Agent专用接口，都反映了对最佳人机（或Agent-机）交互方式的探索。 | 高质量的训练数据（尤其是人类示范数据和特定领域数据）以及持续的用户反馈数据（尽管其提升智能的效率存疑，但对产品优化有用）是重要的竞争壁垒，但并非传统意义上的“数据飞轮”。 | 顶尖的AI研究员和工程师是各家争夺的核心资源。 |

# 模型即产品：爆款应用背后是模型能力更新



- 在LLM经历突破性发展的初期阶段，模型本身展现出强大能力（如对话、写作、编码、推理），以至于模型本身几乎就等同于产品。模型的“magic moment”往往直接定义了新的产品可能性。例如：
  - GPT-3.5解锁了Chatbot形态；
  - Claude Sonnet解锁了Cursor；
  - DeepSeek的出圈也是因为其R1推理能力，其产品形态并没有过多优化；
  - OpenAI DeepResearch 并非在O3上做了套壳，而是基于CUA重新训练了专有模型。

AI应用的核心价值很大程度上就是让用户能够便捷地体验和使用这些前沿模型的基础能力。“模型的能力”几乎就是产品的全部吸引力。

DataBricks 公司生成式AI副总裁Naveen Rao 预测：在未来两到三年内，所有闭源的 AI 模型提供商都会停止销售 API 服务。

这表明，API 经济即将走向终结。**模型提供商与应用层之间原本的蜜月期可能已经彻底结束。我们已经看到了一些迹象：**

- **大模型公司开始停止提供最新模型的API：**DeepSearch 并未提供 API 接口，仅作为 OpenAI 高级订阅的增值功能出现；Claude Code 则只是一个极为简单的终端整合。模型厂商已开始跳过第三方应用层，直接创造用户价值。
- **应用层企业开始布局模型训练能力：**应用型公司也意识到了这种威胁，尝试转型。例如 Cursor 拥有一款自主开发的小型代码补全模型；WindSurf 内部开发了 Codium 这样一款低成本的代码模型；Perplexity 此前一直依靠内部分类器进行请求路由，最近更是转型训练了自己的 DeepSeek 变体模型用于搜索用途。
- “应用套壳商”（Wrappers）实际上处于困境之中：他们要么自主训练模型，要么就等着被上游大模型彻底取代。他们现在所做的事情，本质上都是为上游大模型厂商进行免费的市场调研、数据设计和数据生成。

# 模型即产品：浅层套壳产品终将被颠覆



**什么是“浅层套壳产品”？**

“浅层套壳产品”（Wrappers）指的是那些仅仅在强大的底层AI模型（通常通过API调用）之上增加了一个相对简单的用户界面（UI）、应用外壳或非常基础的功能封装，而没有提供显著附加价值的应用。这类产品可能包括：

- 提供特定Prompt模板的简单问答工具。
- 对模型输出进行非常有限的格式化或后处理的应用。
- 仅仅是换了个皮肤或交互方式来调用通用模型API的服务。
- 缺乏深度工作流整合、独特数据、复杂功能或差异化用户体验的应用。
- **它们的核心竞争力几乎完全依赖于底层模型的表现，自身的“护城河”非常浅。**

**我们已经看过了太多的失败案例。**AI墓地（AI Graveyard）网站统计了5046个AI应用，其中1210个已停止运行或停止服务（截至2025/4/28），其中许多是套壳产品，停运最多的类型是AI写作工具。这些套壳产品通常模仿大模型如ChatGPT的功能，但由于娱乐性大于实用性、难以应对复杂社交场景、用户留存和盈利能力不足等原因而失败。

**AI Graveyard各类型代表性项目**

| 类型 | 产品名 | 简介 |
|---|---|---|
| **Chatbot** | Addcontext.xyz | 为用户创建个性化聊天机器人的平台 |
| | Write-a-card | AI贺卡信息生成器 |
| | BibleGPT | 圣经GPT |
| **AI Writing** | Neuralcanvas | AI动漫生成平台 |
| | Postgeniusapp | 社交媒体推文生成器 |
| | Cluc | SEO优化内容生成工具 |
| **AI Image** | Photofix | AI照片编辑器 |
| | MakePose | 角色和动作生成器 |
| | Illustrate | AI插图生成工具 |
| **AI Design** | AI Designer | 可视化室内设计工具 |
| **Productivity** | ClipGPT | AI书签和笔记工具 |
| **Audio** | Whisper.ai | openAI开发的多语言转录、翻译和识别工具 |
| **Video** | Question Youtube | AI视频问答工具 |

资料来源：甲子光年，AI Graveyard，东吴证券研究所

# 模型即产品：浅层套壳产品终将被颠覆



**我们认为浅层套壳产品终将被颠覆。为什么？**根本原因在于这类产品缺乏可持续的竞争壁垒，极易受到快速迭代的AI技术和市场格局的冲击：

- **过度依赖底层模型，易受到模型迭代的降维打击。**浅层产品的能力上限完全由底层模型决定。一旦底层模型升级换代缓慢、API提价、调整服务策略甚至停止服务，这些产品将立刻失去竞争力甚至无法生存。它们的命运完全掌握在模型提供商手中。然而，基础大模型的能力迭代速度极快。今天需要一个“套壳”应用才能实现的功能，明天可能通过调用新一代模型的一个简单Prompt就能直接完成。模型能力的提升会不断“内化”原本属于应用层的功能，使得那些仅仅是对旧模型能力进行封装的浅层应用迅速变得多余和过时。
- **缺乏核心壁垒，易被复制：**由于没有构建真正的技术或产品壁垒，竞争对手可以轻易地通过调用相同或类似的底层模型API，快速复制出一个功能相近的产品，导致市场迅速陷入同质化竞争和价格战。
- **平台整合的挤压效应**：拥有强大基础模型和生态系统的平台公司（如微软、谷歌）倾向于将AI能力深度整合进其操作系统、办公套件、浏览器等核心产品中（如Windows Copilot, M365 Copilot）。用户在熟悉的、高频使用的平台内就能便捷地获得类似甚至更好的功能，这将极大挤压独立的、功能单一的浅层套壳应用的生存空间。

**平台方的核心优势**

1. **掌控基础大模型：**平台方如OpenAI、Google、微软、Anthropic等投入巨资研发和迭代基础大模型。这些模型是Agent能力的“大脑”，其性能、成本和功能直接决定了上层应用的天花板。应用开发商在很大程度上依赖平台方提供的模型API。
2. **控制算力：**Agent的训练和大规模推理需要庞大的算力、存储和网络资源。平台方通常也是主要的云服务提供商（如AWS、Azure等），它们不仅提供这些底层资源，还越来越多地推出专门的Agent开发、托管和管理平台。这使得应用开发商在基础设施层面也对平台方产生依赖。
3. **设定技术标准与构建生态：**平台方有能力推动和设定关键的技术标准和协议（如MCP）。它们通过提供SDK、开发者工具和应用市场（如GPT Store），吸引开发者围绕其平台构建应用。一旦生态形成，平台方可以通过分发、认证、服务抽成等方式进一步巩固其价值地位，并增加开发者的迁移成本。
4. **数据与研发的规模效应：**平台方拥有海量数据用于训练更通用的基础模型，并且具备更雄厚的资金实力进行前沿的AI研究。这种规模效应使得它们在提升模型通用能力和探索新技术方面具有显著优势。
5. **强大的分发渠道：**平台方通常拥有庞大的现有用户基础（操作系统、搜索引擎、办公软件、社交网络、企业客户群），可以将Agent能力和相关产品快速触达海量用户，这是初创应用开发商难以比拟的。

我们认为真正的壁垒来自于复杂工作流的可靠编排、高质量且持续维护的工具集成能力、难以通过通用模型获得的深度领域知识。这三大要素共同构成了超越底层模型能力的、真正可持续的产品壁垒和护城河。它们需要大量的工程投入、领域专长、产品设计智慧和持续运营维护。那些能够在这几个方面建立优势的 AI Agent 产品，才能在激烈的竞争中脱颖而出，避免沦为被轻易颠覆的“浅层套壳”。

**真正的技术和产品壁垒**

| 维度 | 能力 |
| --- | --- |
| 复杂的工作流编排 | 超越简单脚本：这不是指执行几个预设的线性步骤。真正的挑战在于设计和管理能够处理长链条、多分支、包含条件逻辑、循环、甚至并行处理的复杂任务流。Agent 需要能将用户的模糊意图或高级目标，分解为一系列具体、可执行的子任务，并规划它们的执行顺序和依赖关系。<br><br>驾驭现实世界的混乱：现实世界充满不确定性。API 可能临时失效、网页结构可能改版、外部服务可能返回预期外的数据或错误。一个强大的 Agent 必须具备复杂的错误处理机制（如识别错误类型、自动重试、切换备用方案）和动态重规划能力（在某个步骤失败或环境变化时，能调整后续计划以继续达成目标）。这需要深厚的工程实践和对失败模式的充分预估与处理。 |
| 高质量的工具集成与维护 | 超越简单调用：集成工具远不止是知道一个 API 端点。高质量集成意味着 Agent 能够在众多可用工具中，准确判断在当前任务的哪个环节、使用哪个工具最合适。 |
| 特定领域知识与优化 | 领域专业知识和数据的积累是重要的护城河，通用大模型难以直接具备。 |

# Big Giants：角逐AGI、通用Agent和流量入口



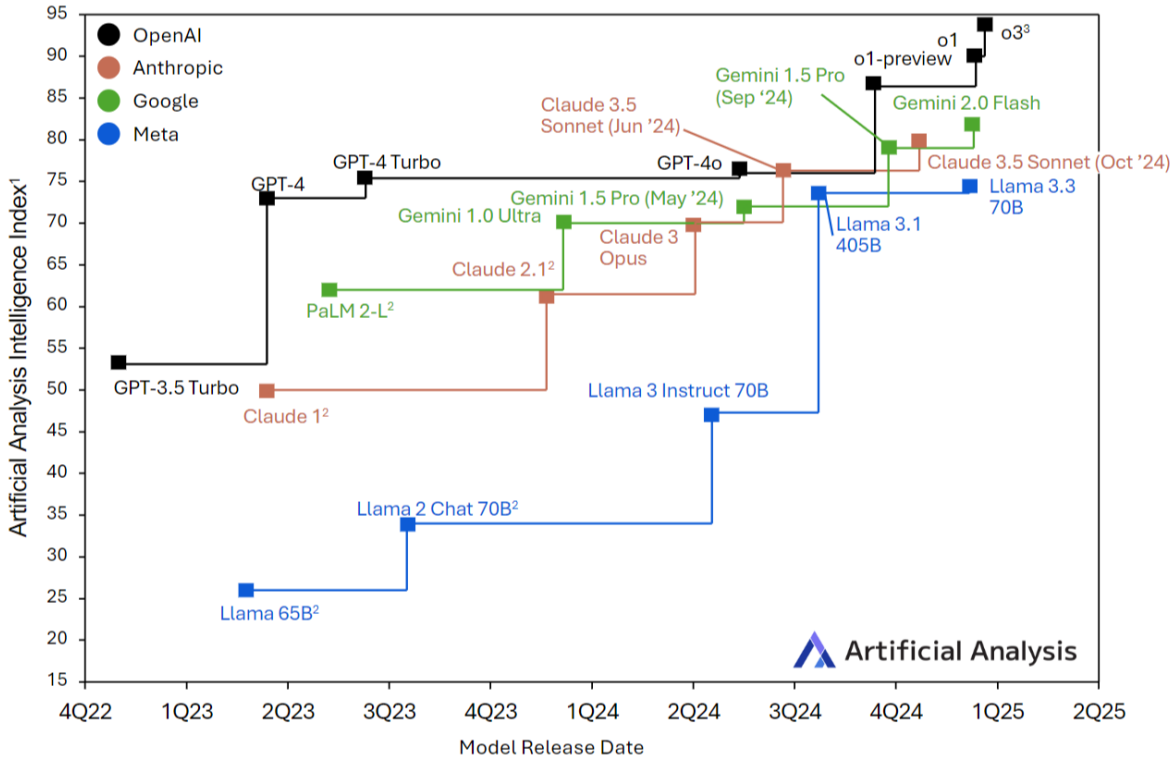
**我们将Agent分为两大类型：一是垂直型Agent，**具有预先设定好的prompt和workflow，通常融合了特定行业的know-how；**另一类是通用型Agent，**其智能程度更高，会根据用户目标自主生成执行计划并决定调用哪些工具，因此适用范围更广。

| | 垂直型Agent | 通用型Agent |
|---|---|---|
| **执行方式** | 类似于编程中的**“编译”**。用户在Agent执行任务之前，通过Prompt、拖拽界面或其他方式，预先设定好一个固定的 Workflow。它会在流程中调用大模型来完成特定步骤，但整体路径是固定的。 | 类似于编程中的**“解释”**。Agent在接收到用户任务之后，在运行时动态地、实时地进行思考、规划和决策，决定下一步该做什么。Agent根据当前情况和目标，自主生成执行计划（可能是内部的，也可能展示出来），并灵活调用工具（包括代码编写、网页浏览、用户交互等）来执行。没有完全固定的、预先编译好的针对该具体任务的工作流。 |
| **代表产品** | Devin, Cursor（早期或特定功能模式下），以及面向特定行业的智能体，它们往往内置了固化的行业知识和操作流程。<br>OpenAI Deep Research | Manus, OpenAI Operator, Genspark |
| **灵活性** | 灵活性低: 无法应对工作流之外的突发情况或新需求。一旦某个环节出错，可能整个任务就失败了。 | 灵活性高: 能够应对更广泛、更开放的任务，可以根据实际情况调整策略，甚至在遇到困难时（如虚拟机崩溃、搜索失败）能自主寻找变通方法（如请求用户协助）。<br>展现“计算思维”: 能根据任务需要，自主选择最高效的工具和方法，例如判断出写代码比纯粹“思考”更有效时就去写代码。 |
| **通用性** | 适用范围窄: 只能处理那些能被预定义流程覆盖的任务。 | 通用性强: 不受限于预设流程，理论上能处理更复杂、新颖的任务。 |
| **稳定性** | 稳定可靠: 因为流程固定，执行结果相对可预测，不易出错。 | 稳定性较低: 动态规划可能导致行为不可预测，任务失败率可能更高。 |
| **成本** | 成本较低: 可以优化模型调用，甚至某些步骤无需调用大模型，效率高。 | 成本较高: 需要更多的实时思考和规划，意味着更多的大模型调用，计算成本更高。<br>响应时间较长: 实时规划和执行复杂步骤需要时间，任务耗时可能较长。 |
| **行业know-how** | 易于融入行业Know-how: 开发者可以在设计工作流时直接嵌入行业数据。 | 领域知识融入相对困难: 主要依赖底层大模型的通用能力，深度垂直领域的Know-how不如编译型Agent那样容易直接嵌入。 |

# Big Giants： 角逐AGI、通用Agent和流量入口

**“通用性”在AI行业发展的主线，也是“皇冠上的明珠”。** L1阶段的大模型是通用大模型，到了L2的推理模型，我们也在追求通用推理模型。几乎没人去做垂直行业的专属模型，推理模型一出现就是通用的。我们认为在L3阶段，仍然会是大厂（例如美国“七姐妹”、OpenAI、Anthropic以及国内的腾讯、字节、阿里等）主导通用型Agent的格局。通用型Agent要求底层模型具有较强的智能，是几乎所有的大厂都在追求的皇冠明珠。大厂一方面在底层模型上持续迭代，另一方则布局Agent平台和生态，并将Agent能力集成到现有的产品和业务中。

**底层大模型的智能提升如同攀登珠穆朗玛峰**




注：纵轴为模型在多个测试集（包括MMLU, GPQA Diamond, MATH-500, HumanEval）的得分均值

**科技大厂主导通用Agent与生态**

| | 策略 | 例子 |
|---|---|---|
| **基础大模型** | 科技大厂基本都拥有或正在大力投入研发自己的基础大模型，这是构建Agent能力的核心 | OpenAI的GPT系列、Google的Gemini、Anthropic的Claude、Meta的Llama、X的Grok、字节的豆包、腾讯的元宝、阿里的通义千问等 |
| **平台与生态** | 科技大厂不仅开发自己的Agent应用，更致力于提供Agent开发平台，构建围绕自身模型的开发者生态系统，让第三方也能开发和部署Agent。 | 亚马逊的Amazon Bedrock Agents、OpenAI的 Agents SDK、微软的Microsoft Copilot Studio、谷歌的Vertex AI Agent Builder、阿里的百炼、百度的文心智能体平台等 |
| **应用集成** | 将Agent能力集成到现有的核心产品和业务中是普遍策略 | 操作系统（微软）、办公软件（微软）、社交平台（Meta的Facebook、X、字节的抖音和头条、腾讯的公众号和视频号）、电商（阿里）、云服务、搜索（百度、谷歌、微软Bing）等 |

资料来源：Artificial Analysis，东吴证券研究所

# Big Giants：谷歌——平台制胜、标准引领、生态融合

东吴证券 SOOCHOW SECURITIES

**谷歌正在构建一个全方位的 AI Agent 生态系统，其战略包括：**

- 平台制胜: 以 Vertex AI Agent Builder 为引擎，打造企业级 Agent 开发与部署的强大中枢，最大化整合其在 AI 研究、云计算及开发者生态的深厚积累。
- 标准引领: 采取**“采纳+主导”**的双协议策略 (MCP + A2A)，既确保当下兼容性，又力图塑造未来多智能体协作规则，巩固其平台和生态的战略优势。
- 生态融合：将 Agent 能力深度植入其庞大的现有产品矩阵（搜索、Android、Workspace、Cloud），利用 AI 赋能数十亿用户与企业客户，实现核心业务的巩固与扩张。

| 方向 | 布局 |
|---|---|
| **Agent产品** | **消费者端: Google Assistant** 深度整合 Android 及智能家居生态。<br>**企业端: Agent Assist** 聚焦呼叫中心，提供实时智能辅助 |
| **产品深度整合** | **Gemini** 驱动同名聊天机器人 (前 Bard) 。<br>Agent 能力**全面渗透** Google 核心服务：搜索、Workspace、Android 、Cloud 等<br>**Google Assistant SDK** 赋能第三方集成 |
| **基础模型** | 谷歌是 Transformer 架构的奠基者及 LaMDA 的开发者<br>谷歌的 Gemini 系列 (Nano/Ultra, 1.5 Pro/Flash, 2.5 Flash) 具有强大的多模态能力和开创性的超长上下文窗口 (如 1.5 Pro/Flash 支持 1M tokens) |
| **开发者生态** | **中心平台是Vertex AI Agent Builder** (基于 Google Cloud)，提供构建与部署**企业级多智能体系统**的**端到端**解决方案。**核心组件包括：Agent Garden**（发现与探索 Agent 范例及工具）、**ADK**（开源框架，简化复杂 Agent 构建）、**Agent Tools**（全面的工具库，涵盖**内置工具、RAG 引擎、Google Cloud、MCP 协议支持**及第三方）、**Agent Engine**（全托管运行时，赋能生产环境的 Agent 部署、管理与规模化扩展）、**辅助工具（Google AI Studio** 提供便捷的 Gemini API 访问，并与 Langbase 等第三方平台协同） |
| **协议领导力** | **采纳 MCP:** 拥抱**模型上下文协议 (MCP)**，确保与现有工具生态的广泛兼容。<br>**主导 A2A:** 牵头发起 **Agent2Agent (A2A) 协议**，旨在**定义未来 Agent 间通信与协作的标准**，发布即获 **50+ 伙伴**支持。 |


资料来源：IBM，Google，Platform Engineering，东吴证券研究所

# Big Giants：腾讯——元宝嵌入微信生态，后续潜力值得期待

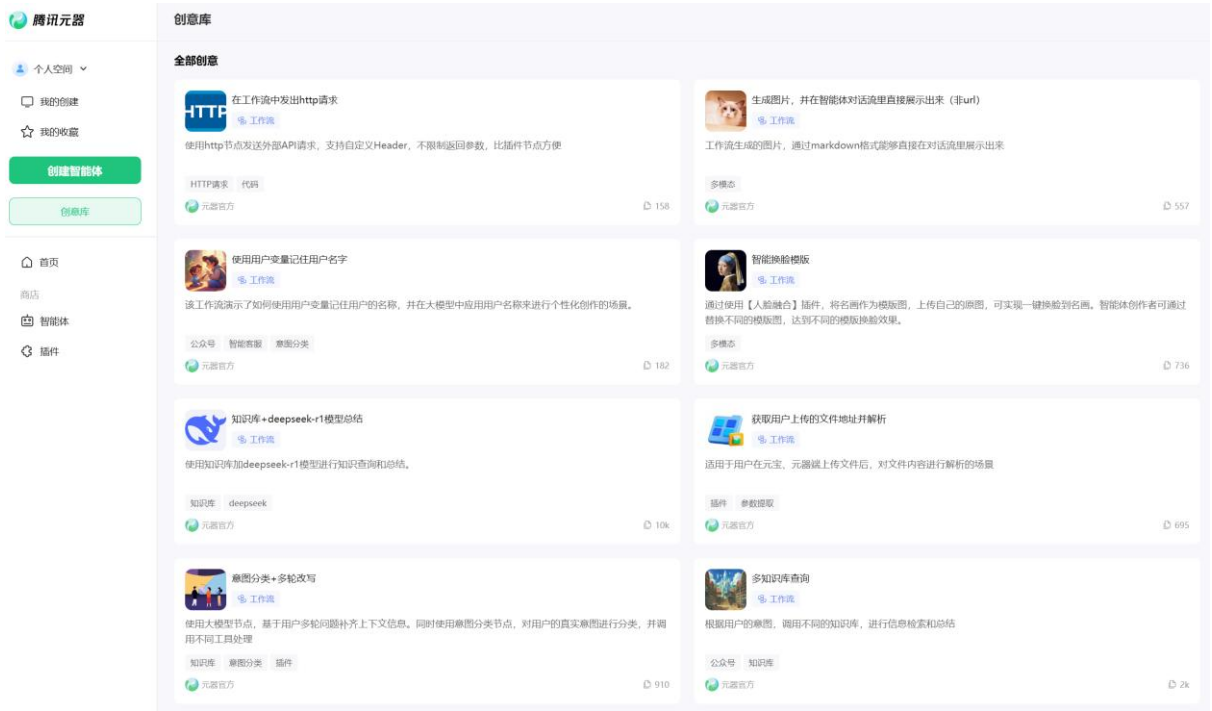
腾讯尚未推出独立的Agent产品与OpenAI等直接竞争，而是将其AI能力融入到现有生态中，旨在利用网络效应，通过AI增强核心产品的用户体验和粘性，进一步巩固其生态壁垒。例如，混元大模型已在腾讯内部支持超过700个业务场景。众多核心产品已接入混元或DeepSeek等模型，例如微信（测试AI搜索功能）、QQ、腾讯文档（AI辅助创作、润色、校阅）、QQ浏览器、QQ音乐、腾讯会议（会中问答、会议总结和待办事项整理）等。在广告和营销领域，AI被用于智能素材创作和构建智能导购。

在agent开发生态方面，开发者可以在腾讯的“元器”平台上搭建Agent。但是目前平台上的agent仍以workflow形式为主，功能也较为单一。

我们期待后续微信相关生态通过AI Agent进行打通，例如微信元宝作为个人助手，辅助用户整理消息、制定待办事项、调用小程序等。

微信添加“元宝”作为好友
微信元宝尚且不具备调用小程序的能力，
目前仅支持搜索网页和微信公众号


腾讯元器平台上的agent以workflow为主

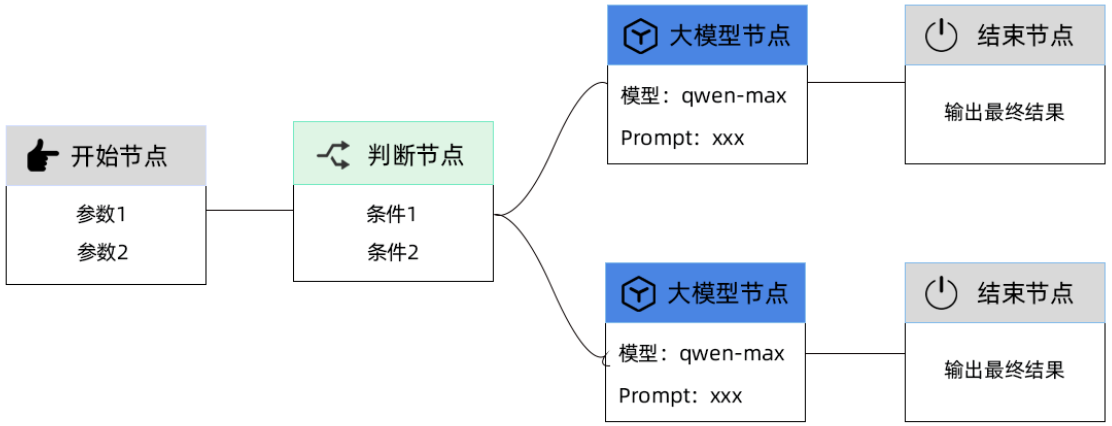



资料来源：壹娱观察，腾讯研究院，极客公园，证券时报，微信元宝，腾讯元器，东吴证券研究所

# Big Giants：阿里——agent能力嵌入钉钉生态

**和腾讯类似，阿里目前尚未推出独立的agent产品，而是将agent能力整合到现有产品中（例如钉钉），并推出低门槛的agent开发平台：**

- **以企业协作为核心突破口。**将AI Agent整合到钉钉这一拥有庞大企业用户基础的平台上，是其最核心的策略。通过提升钉钉用户的办公效率和自动化水平，阿里巴巴旨在巩固其在企业服务市场的地位，并将AI能力转化为实实在在的生产力工具。
- **构建低门槛的Agent开发平台。**阿里云的百炼平台提供的零代码/低代码Agent创建能力，降低了企业构建定制化AI应用的门槛，有助于推动通义模型和阿里云服务的普及。
- 3月11日，阿里巴巴通义千问与AI创业公司Monica旗下的Agent产品Manus达成合作，基于通义开源模型为中国用户打造Manus的功能。




**百炼大模型平台（Model Studio）**是一个支持零代码/低代码创建智能体应用的平台。开发者可以在百炼上选择通义千问等大模型，集成RAG能力（连接私有知识库）和各种插件（如图像生成、自定义插件）来构建面向特定业务场景的Agent应用。和腾讯元器类似，阿里的百炼平台仍然是workflow型，智能程度有待提高

# Big Giants：字节——展现强大的执行力

**2025年4月18日晚间，字节跳动开启了通用Agent平台——扣子空间的内测，采用邀请码制。平台上，用户拥有多样化的选择，既可以启用精通各项技能的“通用实习生”，也能够选择“用户研究专家”“A股观察助手”等“领域专家”，通过与AI互动完成各类工作任务。**

**极客公园对扣子空间进行了实测，在测试的多个任务中，制定旅游攻略和一周穿搭的任务完成情况良好，但是专家助手的任务测试出现诸多问题。**
以制定日本旅行攻略为例，扣子空间在10分钟以上完成任务，它运用“边想边搜”“边想边做”的模式，不仅规划出合理的行程安排，还生成了包含地图、景点介绍、必备日语短语以及旅行提示的HTML旅行手册，甚至还能根据用户需求提供个性化策划节目，实用性颇高。此外，在将旅行攻略转语音以及根据天气推荐穿搭并制图的任务中，扣子空间也展现出了一定的能力，虽然存在一些小问题，比如语音朗读时会读出符号，首次未按用户需求输出穿搭图片等，但整体功能方便好用。
然而，在专家助手的任务测试中，却出现了诸多问题，像Python脚本调用失败、API权限异常等Bug，导致任务执行受阻，一个晚上都未能成功完成。在使用“A股观察助手”时，任务执行也不稳定，出现数据未能获取、脚本调用失败的情况，且时长被大幅拉长。

**值得一提的是，字节在Agent赛道展现出强大的速度与执行力。**在扣子空间内测前一天，火山引擎面向企业市场推出OS Agent解决方案及AI云原生推理套件，助力企业构建和部署Agent应用。同时，字节发布豆包·深度思考模型，同步升级文生图模型3.0、视觉理解模型，为扣子空间的功能实现提供了有力的技术支撑。火山引擎总裁谭待强调，做好Agent在技术上需具备更强的多模态模型、更好的操作架构工具以及通过AI云原生降低模型推理成本和延迟，扣子空间的内测或许意味着字节已基本达成这些要求。

■ 扣子支持添加MCP扩展，接下来或有更多插件接入


■ 通用任务测试：根据天气预报提供穿搭建议


■ 专业任务测试：生成A股早报

# Niche market：垂直Agent长期面临通用Agent的威胁

通用Agent（如OpenAI Operator）的目标是具备广泛适用性，能够操作各种软件和完成多样化任务；而垂直Agent（如编程领域的Devin、销售领域的ElevenX）则专注于特定行业或职能，追求在特定领域的深度和专业性。

**这场博弈的关键在于：通用Agent不断提升的泛化能力，最终会在多大程度上蚕食甚至取代垂直Agent的市场？**

**短期内垂直Agent的优势所在（护城河）**

1. **深度领域知识与经验，在特定领域更高的任务成功率和可靠性：**这是垂直Agent最核心的壁垒。在特定行业里，往往沉淀了大量非结构化、未文档化、甚至只可意会的专业知识、术语、流程规范、决策逻辑和“潜规则”。这些深度知识很难通过通用模型的预训练完全覆盖，也难以仅通过简单的Prompt或少量示例教会通用Agent。垂直Agent在特定领域具有优势。由于深度优化和领域知识的加持，垂直Agent在其专长领域通常能达到比通用Agent更高的任务完成率和可靠性，尤其是在处理边缘情况和复杂细节时。
2. **定制化的工作流与集成：**垂直Agent可以针对特定业务流程进行深度定制，无缝集成行业常用的软件系统（如ERP、CRM、行业数据库、专用硬件接口等）。这种与现有生态的紧密耦合是通用Agent短期内难以实现的。
3. **优化后的性能与成本：**针对明确的任务范围，垂直Agent可能采用更轻量、更高效的模型或算法组合，或者通过大量领域数据进行优化，从而在特定任务上实现比通用Agent（可能需要调用昂贵的大模型进行每一步推理）更快的响应速度、更高的稳定性和更低的运行成本。
4. **数据隐私与合规性：**在金融、医疗等高度敏感和受监管的行业，客户更倾向于选择了解并能满足特定数据安全和合规要求的垂直解决方案提供商。通用Agent的数据处理流程可能不够透明或难以满足严格的行业规范。

**长期来看通用Agent的威胁（泛化能力的冲击）**

1. **基础模型能力的指数级提升：**通用基础大模型（尤其是多模态模型）的理解、推理、规划和学习能力正在快速提升（Scaling Law）。通用Agent的能力上限随之提升。今天通用Agent难以处理的复杂领域知识或任务，未来可能通过更强大的模型能力轻松解决。
2. **更强的学习和适应能力：**通用Agent天生设计用于适应各种环境和任务。随着上下文学习（In-Context Learning）、工具调用（Tool Use）、检索增强生成（RAG）等技术的成熟，通用Agent能够更有效地动态获取和利用外部知识（包括垂直领域的知识库），弥补自身知识的不足。用户或企业也可以对通用Agent进行轻量级微调或提供特定指令集，使其快速适应特定需求。
3. **“足够好”效应与便利性：**对于许多非核心或非极端复杂的垂直任务，用户可能并不追求极致的专业表现。一个“足够好”且能够处理多种任务的通用Agent，可能比管理多个独立的垂直Agent更加方便和经济。便利性往往是技术普及的关键因素。
4. **平台集成与分发优势：**通用Agent往往由平台型巨头开发，能够深度集成到操作系统、浏览器、办公套件等用户日常使用的环境中（如Windows Copilot）。这种无缝集成和巨大的分发优势，使得用户更容易接触和习惯使用通用Agent，从而挤压独立垂直Agent的入口机会。
5. **成本下降趋势：**随着模型效率提升和市场竞争加剧（尤其是开源模型的冲击），通用大模型的API调用成本预计将持续下降。这可能会削弱垂直Agent在成本上的一些优势。

**结论：我们认为，通用Agent的泛化浪潮对垂直Agent构成了长期且显著的威胁，但短期内垂直Agent凭借其深度领域知识和定制化能力仍有发展空间。**

# Niche market：谈谈cursor的壁垒

东吴证券 SOOCHOW SECURITIES

**以Cursor为例，其成长曲线极其陡峭：**ARR从23Q1的50万美金增长至25Q1的1.5亿美金，2年时间增长约**500倍**；估值从23年12月的4亿美金增长至25年3月的100亿美金（洽谈中），15个月里增长约**25倍**。**推动ARR增长的核心逻辑是：**接入更多的、更好的大模型（例如claude sonnet 3.5）；优化产品能力，例如稳定性、反应速度、准确率、跨文档能力等。cursor成员Aman曾说：“未来一年的Cursor需要让今天的Cursor看起来过时。”

**Cursor的成长曲线极其陡峭**

| 日期 | 事件 | ARR | 融资与估值 |
|---|---|---|---|
| **2023年1月** | 发布cursor | | |
| **2023年3月** | 迁移至VSCodium架构，提升IDE稳定性 | **23Q1的ARR为50万美元（免费用户为主）** | |
| **2023年6月** | **引入GPT-4模型**，代码生成质量显著提升 | | |
| **2023年9月** | 推出20美元/月的pro版 | 120万美金 | 天使轮：获得open startup fund 800万融资 |
| **2023年12月** | **获得GPT-4 API早期访问权限**，代码生成准确率提升至78% | | |
| **2024年3月** | 与anthropic达成模型合作，**集成claude 3.5 sonnet模型**，响应速度提升 | 450万美金（Pro版订阅量激增） | |
| **2024年6月** | 推出推测解码技术，实现1000 tokens/秒生成速度，使响应延迟降低 | | |
| **2024年8月** | | | **A轮融资6000万美金，估值4亿美金** |
| **2024年9月** | 引入跨文件修改功能，支持全项目上下文理解 | 24Q3的ARR 2400万美元 | |
| **2024年11月** | 推出企业级SSO功能；上线Codebase Agent，实现代码库智能问答 | ARR预计6500万美元，付费用户数4万人 | |
| **2024年12月** | | | B轮融资1亿美金，估值25亿美金 |
| **2025年1月** | 本地代码索引系统将上下文理解准确率提高至92% | | |
| **2025年2月** | 开放本地化部署选项 | | |
| **2025年3月** | **集成Llama3.1 405B模型**，推理成本降低40% | **25Q1的ARR突破1.5亿美元** | **C轮洽谈中，估值预计100亿美金** |

资料来源：AI洞见研究院，Z potentials，东吴证券研究所

# Niche market：谈谈cursor的壁垒

**如何理解Cursor的壁垒？**我们认为Cursor的壁垒更多来自于围绕着特定应用场景精心构建的“产品体验”和“集成工程”，而非技术垄断壁垒。这是一场关于产品、工程和速度的竞赛，而非单纯的技术竞赛。这个壁垒的“深度”取决于：

- Cursor在产品体验和功能创新上领先竞争对手的速度。
- 基础大模型将Agent能力“内化”的速度（例如github copilot）
- 竞争对手（例如windsurf、devin）整合AI能力的决心和效果

**如果基础模型的能力变得足够强大且易于调用，而竞争对手又能提供足够好的集成体验，那么Cursor的壁垒就可能被削弱。**

| | | Cursor | Windsurf | Devin | GitHub Copilot |
|---|---|---|---|---|---|
| ARR与估值 | ARR | 1.5亿美金（截至25年3月） | 约3000万美金（截至25年4月） | 未披露 | 未披露 |
| | 估值 | 预计100亿美金（截至25年3月） | 预计约30亿美金（截至25年4月） | 40亿美金（截至25年3月） | 微软旗下产品 |
| 核心能力 | 代码生成 | 支持多模型，实时建议与补全 | 预测用户意图，自动生成多文件 | 多任务并行代码生成与修改 | 实时代码补全，支持复杂函数实现 |
| | debug | 提供实时错误检测与自动修复 | Inline AI精确控制代码编辑 | 深度搜索与自动测试 | 自动检测错误并建议修复方案 |
| | 工作流支持 | 多文件上下文理解，跨文件修改 | Cascade模式自动化工作流 | Interactive Planning生成任务计划 | 集成开发环境（如VS Code） |
| | 文档生成 | 集成文档建议，提升开发效率 | 自动生成文档与注释 | Devin Wiki自动索引代码库 | 文档生成与代码解释 |
| | 协作能力 | 支持团队协作 | 提供Memories优化上下文响应 | 多实例并行任务处理 | 提供团队版，支持多人协作 |
| | 图像支持 | 无 | 可上传图像生成HTML/CSS/JS代码 | 无 | 无 |
| 用户体验 | 易用性 | 界面简洁，支持多模式切换 | 功能丰富但入门较难 | 专注任务规划与深度搜索 | 集成度高，易上手 |
| | 定制化能力 | 支持自定义模型选择 | 提供上下文优化规则 | 自动化任务规划与文档生成 | 对项目需求的适应性强 |
| | 性能 | 快速响应，多模型支持 | 自动化程度高，但需用户确认 | 深度分析但响应稍慢 | 补全速度快但偶有错误 |
| 市场定位 | 定价 | 个人版：免费、$20/月、$40/月<br>企业版：$36,096–$115,200/年 | 免费、$15/月、$30/月、$60/月 | $20/月起价+额外购买计算单元$500/月 | 免费、$10/月、$39/月 |
| | 主要面向 | 小型项目或个人开发 | 前端开发 | 大型团队协作 | 小型项目或个人开发 |

资料来源：Z potentials，cursor，windsurf，devin，Github，the decoder, signal hub, data camp, cognition, swimm, 东吴证券研究所

# Niche market：垂直Agent的价值在于深耕领域知识



**我们梳理了30家上市公司在AI Agent领域的布局（具体内容见于后文表格），有四点结论：**

**1. AI Agent 已成为众多行业公司的战略重点，**包括：
①**企业软件智能化：** ERP、OA等领域的公司普遍将 Agent 深度集成到核心系统中，用于自动化流程、辅助决策、提升效率。
②**生产力工具增强：** 办公、PDF、创意类软件公司利用 Agent 提供智能写作、编辑、设计、出版、分析等高级辅助功能。
③**垂直领域深耕：** 许多公司针对金融、医疗、教育、电商、外贸、司法、政策服务、软件开发等具体行业的痛点，开发高度专业化的 Agent 解决方案。
④**平台化赋能：** 部分公司不仅开发应用，还致力于构建 Agent 开发平台，赋能自身及客户或第三方开发者。

具体来说，这些公司有：综合科技平台（昆仑万维）、企业软件（用友、金蝶、泛微、致远、汉得）、办公与文档处理（金山办公、福昕）、创意设计（万兴、美图）到金融（同花顺、新致）、医疗（润达、卫宁）、教育（佳发、科大讯飞、豆神教育）、体育（舒华体育、Keep）、电商（焦点、值得买、光云）、司法（金桥）、出版（果麦）、编程（卓易）、中小企业服务（创业黑马）、招聘（科锐国际、北京人力、外服控股、同道猎聘）等多个领域。

**2. 这些产品更接近于Agent，而不仅仅是Chatbot，原因在于：**
①**可以自动完成（特定行业的）任务：**可以自动完成特定的、多步骤的工作流程（如报销审批、报告生成、代码编写、客户开发、政策匹配等）。
②**可以调用（有限的）工具：**与软件（ERP、OA、Office、PDF 编辑器、行业软件）、数据库、API 等进行交互和操作。
③**可以解决（相对）复杂的问题：**能够处理需要一定规划、推理和信息整合才能完成的任务。

**3. 这些公司的Agent产品具有相对明确的商业价值。例如：**大量公司，特别是 ERP、OA、医疗 IT、企业服务领域的厂商，致力于利用 Agent 自动化内部工作流（如审批、报告生成、数据处理、客户服务、人力资源管理等），实现降本增效。**再如**，软件工具类公司普遍将 Agent 作为提升核心产品（如 Office、PDF 编辑器、创意软件）智能化水平的关键手段，提供智能辅助、内容生成、数据分析等功能。

**4. 我们认为这些公司在Agent领域的壁垒来自数据、客户基础、产品打磨和工程化能力。**

# Niche market：垂直Agent的价值在于深耕领域知识

东吴证券 SOOCHOW SECURITIES

**4. 我们认为这些公司在Agent领域的壁垒来自数据、客户基础、生态构建能力、工程化能力：**

**①数据壁垒+行业 know-how。这是垂直领域 Agent 最核心的壁垒之一。拥有独特、高质量、大规模的行业数据，并深刻理解该行业的运作逻辑和痛点，才能训练出真正好用的 Agent。例如：**同花顺（金融数据+用户行为）、科大讯飞（语音+多行业数据）、焦点科技（外贸 B2B 数据+流程）、卫宁健康（医疗 IT 流程）；用友/金蝶（企业经营数据+流程）、新致/金桥（特定行业如金融/司法流程+数据）、润达（IVD+医疗流程）、佳发（教育考试场景数据）、值得买（消费数据+口碑）。

**②客户基础 + 应用场景/工作流绑定。庞大的现有用户群是 Agent 产品推广、获取反馈、迭代优化的基础。将 Agent 深度嵌入客户难以替代的核心工作流中，能建立极高的用户粘性和转换成本。**例如：金山办公（海量用户+办公场景）、用友/金蝶（大量企业客户+核心 ERP 流程）、泛微/致远（大量企业客户+OA 日常工作流）、卫宁健康（医院客户+核心诊疗流程）；汉得信息（大型企业客户+多行业流程实施经验）、同花顺（金融用户）、福昕软件（PDF 用户）、万兴/美图（创意/影像用户）、焦点科技（外贸用户）。**品牌信任与合规壁垒：**在金融（同花顺、新致）、医疗（卫宁、润达）、司法（金桥）等高敏感、强监管领域，已有的品牌信誉、客户信任以及满足合规要求的能力是重要壁垒。

**③技术平台与生态构建能力。具备构建易用、开放的 Agent 开发平台或拥有核心底层技术（如自研大模型）的公司，能吸引开发者、聚合应用，形成网络效应。**例如：昆仑万维（天工平台+多模态+开源）、科大讯飞（星火平台+AI 全栈能力）、同花顺（Agent Studio+金融插件生态）、致远互联（低代码 Agent 平台）、彩讯股份（Rich AIBox 平台）

**④产品打磨与工程化能力。将 Agent 技术转化为稳定、可靠、用户体验良好的产品，需要强大的工程化和产品设计能力。**例如：金山办公（WPS AI 用户体验）、福昕（PDF 工具成熟度）、万兴/美图（创意工具易用性）、卓易信息（Multi-Agent 架构实现）

# Niche market：垂直Agent的价值在于深耕领域知识



| 类型 | 公司 | 股票代码 | Agent | Agent功能 |
|---|---|---|---|---|
| 文档处理 | 金山办公 | 688111 | WPS AI | 提供智能伴写(内容生成/补全/改写)、数据分析、PPT美化、文档理解等功能，目标是成为集成化的智能办公助理，提升办公效率。 |
| 文档处理 | 福昕软件 | 688095 | AI Assistant | 集成于PDF编辑器，通过自然语言处理提供智能编辑、内容提取、格式调整、文档摘要、信息安全监控等功能，提升PDF文档处理效率与安全性。 |
| 教育 | 佳发教育 | 300559 | 教育Agent | 聚焦教育垂直领域，构建Agent模型应用于智慧考试(智能监考/分析)、个性化教学辅导、体育训练等场景，提升教育效率与质量。 |
| 教育 | 豆神教育 | 300010 | “超拟人”AI导师 | 打造“超拟人”AI导师，具有引导写作、作文点评、陪伴学习等功能 |
| 金融 | 新致软件 | 688590 | ACE产品线 (Agent/Client/Ent.) | 面向金融、保险、司法等行业，提供智能营销代理、智能客服、风险评估等定制化Agent解决方案。 |
| 金融 | 同花顺 | 300033 | 同创智能体平台 (Agent Studio) | 金融垂直领域的低代码Agent构建平台，提供大量金融插件，赋能投研、投顾、量化、风控等场景，服务金融机构与个人投资者。 |
| 医疗 | 润达医疗 | 603108 | “良医小慧”医疗Agent | 面向医护(辅助诊疗)和患者(健康咨询/计划)提供服务；结合IVD业务，利用AI优化诊断服务流程。 |
| 医疗 | 卫宁健康 | 300253 | WiNEX Copilot 智能助手 | 嵌入医院信息系统(HIS/EMR)，通过智能化任务流框架，辅助医生进行病历书写、临床决策、流程管理等，提升医疗工作效率。 |
| ERP | 用友网络 | 600588 | 企业级AI Agent（基于YonBIP） | 深度融入ERP系统，赋能财务、供应链、人力资源等场景，提供智能决策支持(预算优化/风险预警)，推动企业管理智能化。 |
| ERP | 金蝶国际 | 0268（港股） | AI Agent + 云ERP (苍穹平台) | 集成AI能力于云ERP，覆盖财务自动化、供应链预测、客户洞察等，通过自然语言交互简化系统操作，降低中小企业数字化门槛。 |
| OA | 泛微网络 | 603039 | AI Agent + OA (集成e-cology) | 嵌入OA系统，实现智能流程审批、会议纪要自动生成、内部知识问答等，提升日常办公协同效率，已有较多客户落地。 |
| OA | 致远互联 | 688369 | 低代码Agent开发平台 | 提供低代码平台，支持企业便捷构建定制化的Agent，应用于HR、法务、采购等垂直场景. |
| 招聘 | 科锐国际 | 300662 | AI招聘 Agent | 开发了MatchSystem匹配系统、PC端寻访自动化Agent等工具，提升招聘效率，计划2025年进一步推出AI招聘产品。 |
| 招聘 | 同道猎聘 | 6100（港股） | AI面试官“猎聘·Doris” | 自主研发AI大模型“同道汇才”及AI面试官“猎聘·Doris”，实现精准岗位匹配，通过国家生成式AI服务备案。 |
| 营销 | 迈富时 | 2556（港股） | Agentforce | 推出了营销领域大模型——Tforce和AI智能体中台——AI-Agentforce。 |
| 编程 | 卓易信息 | 688258 | AI编程Agent | 采用Multi-Agent架构，内置多种开发角色Agent，实现从需求到代码的自动生成(AI Coding)、任务分解、智能调试等，旨在革新软件开发模式，大幅提升开发效率。 |

资料来源：公司公告，财联社，新浪财经，智通财经，芥末堆，金融界，中国物流信息网，和讯网，大江网，新浪网，东吴证券研究所

# Niche market：垂直Agent的价值在于深耕领域知识

| 类型 | 公司 | 股票代码 | Agent | Agent功能 |
|---|---|---|---|---|
| 出版 | 果麦文化 | 301052 | AI校对/内容辅助工具 | 主要探索利用AI进行文字校对、辅助内容生成与选题策划等，优化图书出版流程，提升编辑效率与内容质量。 |
| 创意软件 | 万兴科技 | 300624 | 集成式AI Agent / 数字人 | 将AI融入视频剪辑、绘图、文档等创意工具，辅助内容创作(自动脚本/智能排版)；推出交互数字人应用于展厅讲解、演示等场景。 |
| 创意软件 | 美图公司 | 1357（港股） | 影像/设计/美业AI Agent | 在美颜、修图、视频编辑等应用中集成AI，提供智能美化与创作功能；面向B端提供AI设计工具和美业解决方案；探索个性化推荐与服务。 |
| 电商 | 焦点科技 | 002315 | 外贸AI Agent (AI麦可/Mentarc) | 专注于“AI+外贸”，提供Agent工具（如AI麦可、Mentarc、Sourcing AI），自动化处理客户开发、选品、营销、采购、订单管理等跨境贸易核心流程。 |
| 电商 | 光云科技 | 688365 | 电商运营/客服Agent | 面向中小电商企业，提供智能客服（自动应答咨询/处理订单）、营销自动化等工具；探索 AI 在电商数据分析（选品/库存预测）中的应用，旨在帮助商家降本增效。 |
| 电商 | 值得买 | 300785 | “小值”AI购物助手Agent | 基于自研消费大模型和数据库，提供商品智能推荐、多维度对比、口碑总结、全网比价等功能，旨在优化用户购物决策流程，成为需求驱动型电商入口。 |
| 多智能体 | 科大讯飞 | 002230 | Agentic AI驱动多重智能体协同 | Director Agent作为主控智能体，负责整体策略规划和任务分配。Audience Agent基于多模态大模型，实现受众刻画，以及新市场用户分层。Creative Agent 支持创意生成-创意衍生-创意结构-创意资产沉淀。Optimization Agent，助力企业实现高效冷启动，快速提升拿量规模。 |
| 企业服务 | 汉得信息 | 300170 | 企业流程自动化Agent | 开发大量面向企业具体业务场景的Agent，覆盖ERP、CRM、供应链、制造、营销、财务等，提供端到端的智能化解决方案，为企业客户降本增效。 |
| 企业服务 | 彩讯股份 | 300634 | Rich AIBox / 应用Agent | 提供多智能体开发平台(Rich AIBox)，降低Agent开发门槛；推出AI邮箱、AI云盘、智能客服、数字员工等Agent应用，服务企业通信、协作与客户服务。 |
| 企业服务 | 创业黑马 | 300688 | 政策通Agent | 针对中小企业政策查找难、理解难、申报难的痛点，提供政策智能解读、精准匹配、辅助申报等服务；结合生态伙伴提供云服务和版权服务。 |
| 司法/政务 | 金桥信息 | 603918 | 司法/政务AI Agent | 聚焦司法与政务领域，利用AI Agent技术构建多元解纷平台、智慧法院解决方案（如智能庭审、文书分析），提升公共服务智能化水平。 |
| 综合平台 | 昆仑万维 | 300418 | Skywork.ai | 公司预计将于 2025 年 5 月中旬发布生产力场景通用 Agent 平台 Skywork.ai，构建由五大 AI Agent 组成的智能体系，分别针对专业文档、数据表格、演示文稿、播客及网页内容进行深度优化。 |
| 体育 | 舒华体育 | 605299 | AI健身助手 | 在舒华智能触屏跑步机上推出新一代“AI健身助手”，能即时输出包含训练频次、动作组合、强度分级的周计划，并支持按场景筛选居家徒手训练、健身房器械方案等场景化课程。 |
| 体育 | KEEP | 3650（港股） | AI 教练卡卡（Kaka） | AI教练卡卡（kaka）可以精准分析用户运动相关数据，集成了个性化运动计划、可以动态调节训练计划，为用户打造专属智能教练。 |

资料来源：公司公告，每日经济新闻，AI校对王公众号，大江网，科大讯飞，东吴证券研究所

# 四、Agent将最先落地于知识工作（尤其是代码）

# Agent最先落地的行业和场景可能是知识工作（尤其是代码）

**我们认为，AI Agent早期落地场景将是通用办公场景、专业开发领域或特定垂直行业。在这些场景中，Agent可以发挥其不断增强的推理、规划及工具使用能力（尤其是编码和操作数字界面），去自动化目前由人类执行的相对标准化、流程化的数字任务。而且在这些场景中，Agent提升效率、解放生产力的价值能最快体现。而更复杂、需高度创造力（L4）或涉及复杂物理世界交互的任务，是更长远的目标。**

- **自动化重复性、流程化的数字/知识工作：**这是Agent最容易发挥核心价值的场景，也是我们最看好的早期落地场景。因为：这类任务通常流程相对固定，适合当前 Agent 的规划和工具使用能力；能显著提升效率、解放人力（节省注意力），价值明确。
- **软件开发与编程辅助：**这是Agent能力（特别是代码能力）能展现突出优势，且已经有成功产品的领域。因为：编程环境本就是高度结构化、规则明确的数字环境，非常适合Agent发挥作用；且模型在编码任务上进步显著，对开发者生产提升价值巨大。
- **垂直领域的专业Agent（如营销/人力资源）：**对于创业公司而言，这是更容易切入和建立壁垒的方向。因为：任务边界相对清晰，更容易整合领域知识；相比通用 Agent，技术门槛和投入相对较低；商业需求明确。

| | 自动化重复性、流程化的数字/知识工作 | 软件开发与编程辅助 | 垂直领域的专业Agent |
|---|---|---|---|
| 代表产品 | OpenAI Deep Research, Perplexity | GitHub Copilot, Cursor, Devin | 如营销/人力资源等行业的Agent |
| 具体能力 | • 信息研究与报告生成：自动搜集、整理、分析信息并生成报告，辅助研究人员、分析师等知识工作者。<br>• 操作软件和网页：自动执行需要与软件界面或网页交互的任务，如填写表单、预订差旅、处理邮件、管理日程、关闭广告、计算退款等。<br>• 数据处理与分析：自动执行数据提取、清洗、初步分析等任务。 | • 代码生成、补全、调试: 提升开发者效率。<br>• 复杂开发任务执行: 能够理解需求、规划步骤、编写代码、配置环境、测试、修复 Bug 等更完整的开发流程。<br>• API 调用与集成: Agent 利用编码能力与其他系统或服务交互。 | • 客户服务: 处理标准化的客户请求，如查询订单、处理退款等。<br>• 销售/市场营销: 自动化部分销售流程，如潜在客户筛选、邮件营销等。<br>• 人力资源: 辅助处理简历筛选、安排面试等流程化任务。<br>• 特定行业: 如法律文书辅助、医疗信息查询与初步分析等 |

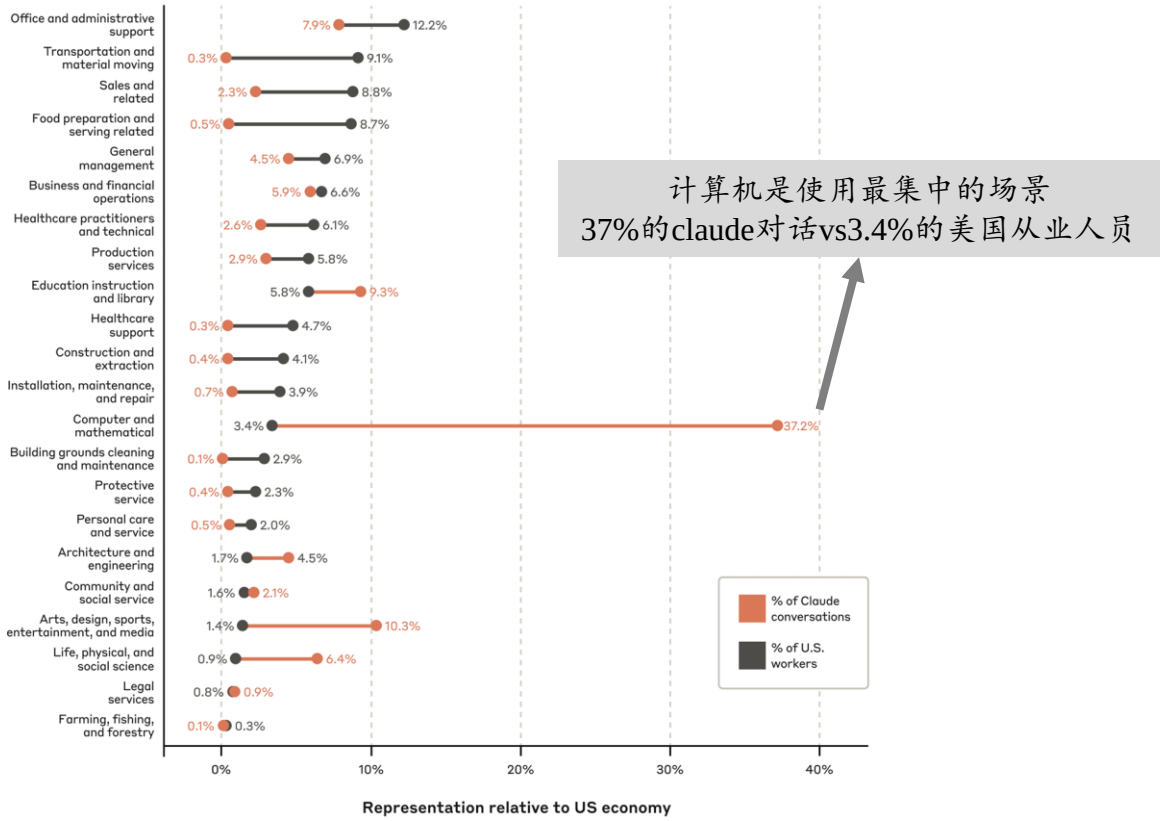
资料来源：东吴证券研究所

# Agent最先落地的行业和场景可能是知识工作（尤其是代码）

**根据Anthropic在2025年3月发布的论文，Claude AI的使用主要集中在软件开发（如编程、调试、维护）和写作任务（如技术写作、文案写作、内容编辑）上，这两者合计占了近一半的总使用量。原因在于：**编程和写作都需要依赖大量的背景知识（语法规则、API文档、特定领域的知识、写作风格指南等）。LLM可以作为强大的知识库和应用工具。

**用户与 Claude 交互时的TOP6场景**


**Claude使用场景分布vs美国从业人员分布**




计算机是使用最集中的场景
37%的claude对话vs3.4%的美国从业人员

注：研究方法：anthropic 对2024年12月和2025年1月的Claude.ai免费版和专业版用户对话数据进行聚合分析，使用Clio（Claude的内部工具）对应用场景进行分类统计。

资料来源：《Which Economic Tasks are Performed with AI? Evidence from Millions of Claude Conversations》，东吴证券研究所

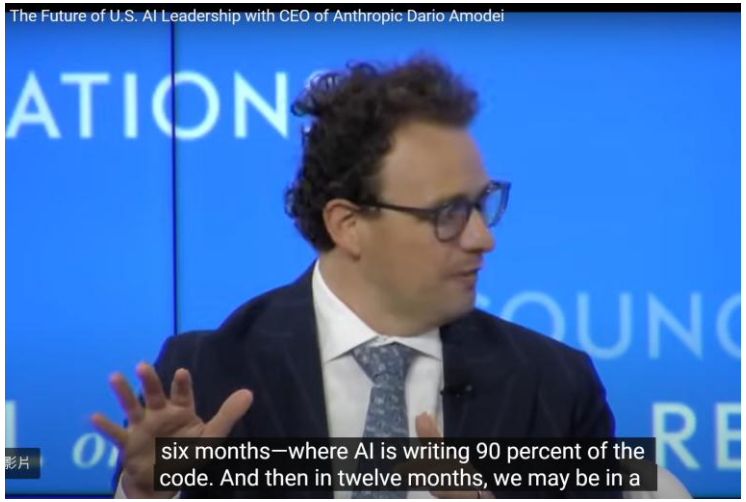
# AI Coding：预计是最快落地的方向

东吴证券 SOOCHOW SECURITIES

- **2024年，谷歌**CEO表示，谷歌已经有超25%新代码，都是由AI生成。
- **2025年1月**财报电话会议上，**Meta**的CEO扎克伯格预测，“2025年软件工程AI将具备中级工程师的编码和问题解决能力。”尽管这场巨变可能不会立即发生，但他希望Meta能够引领这场创新。
- **2025年3月，Anthropic** CEO Dario Amodei 预测，“未来3-6个月，AI将编写90%的所有代码。12个月后，AI可能会包揽几乎全部的代码。” Amodei还认为：尽管程序员仍需要负责设定目标和参数，但这些任务很快也会被AI取代；只要还有AI不擅长的“小块任务”，人类程序员的生产力会得到提升；但最终，这些“小岛”都会被AI系统逐一攻克。
- **25年3月17日，OpenAI**首席产品官 Kevin Weil 在接受采访时预测：“到2025年底，AI在编程领域的能力将全面超越人类程序员，到2025年底，AI编码将实现99%自动化。o1-preview发布后，其实力远超GPT-4，已经能媲美世界排名前百万的竞赛程序员。要知道，全球大概有3000-4000万程序员，o1-preview已经达到了前2-3%的人类level了。等到o1满血版正式版发布时，在竞赛编程方面，已经能达到世界TOP 1000工程师的水平。现在我们正在训练新一代模型，它们的表现更加出色。所以，我认为就在今年，至少在竞赛编程这个领域，AI就会实现超越。就像70年前AI在数学运算上超过了人类，15年前在国际象棋上战胜了人类，今年将是AI在编程能力上永久超越人类的一年。这个发展趋势已经不可逆转了。”

We're also using AI internally to improve our coding processes, which is boosting productivity and efficiency. Today, more than a quarter of all new code at Google is generated by AI, then reviewed and accepted by engineers. This helps our engineers do more and move faster.

I am energized by our progress, and the opportunities ahead. And we continue to be laser focused on building great products.

# Coding Agent：预计是最快落地的方向



**我们认为，编程会是Agent领域最快实现PTF和PMF、最先商业化、迭代最快的领域。原因在于Coding能够为Agent提供理想的环境、成熟的工具、清晰的反馈，且具有明确的价值主张：**

- **理想的环境**（Environment）：Agent 的设计深受强化学习影响，其核心要素是**状态 (State/Context)、行动 (Action/Tool Use) 和激励信号 (Reward Signal)**。一个好的 Agent 应用需要一个能够提供清晰反馈的环境。而集成开发环境 (IDE) 是天然的优质环境: 代码本身是高度结构化的文本，有严格的语法和语义规则，便于模型理解和生成。IDE (如 VS Code) 提供了结构化的项目信息 (状态/上下文)，明确的行动空间 (代码编辑、编译、运行、调试等工具)，以及最重要的——即时且明确的反馈。
- **相对成熟的工具**（tools）：IDE、版本控制系统 (Git)、包管理器、测试框架等构成了成熟的开发工具链，为 Agent 提供了丰富的Tools。
- **清晰的反馈（reward）：**大模型的输出一定得是结构化的。因为只有结构化了，才能够去用代码或者规则去校验。代码作为输出，其结构化特性使得自动化验证（编译、测试）成为可能。在代码领域，软件开发任务（如“修复这个 bug”、“实现这个功能”、“编写单元测试”）虽然复杂，但往往可以被分解为更结构化的子任务。代码执行结果（成功、失败、错误信息、测试结果）是天然的、客观的“激励信号”。Agent 执行一个操作（写代码、修改代码）后，可以通过编译运行或跑测试立刻知道这个操作是好是坏（离目标更近还是更远），这对于 Agent 的学习和迭代至关重要。这解决了通用领域中难以定义“好坏”和提供及时反馈的问题。
- **明确的价值主张和用户痛点：①效率提升:** 编码工作中有大量重复、繁琐、易出错的任务（如写样板代码、调试、测试、文档生成），Agent 在这些方面能显著提升开发者效率。**②降低门槛:** Agent 可以辅助初级开发者或进行跨语言/框架开发，降低学习曲线。**③市场需求：**软件开发是高价值行业，对生产力工具的需求巨大，为 Coding Agent 提供了明确的市场切入点和商业化潜力。

# Agent最先落地的行业和场景可能是知识工作

东吴证券 SOOCHOW SECURITIES

**在法律AI Agent中，Harvey AI是目前估值最高的公司**，25年1月完成3亿美元融资，估值达30亿美元；截至25年初，ARR达5000万美元。

相较于其他法律Agent，Harvey AI全面且高效，适用于大型律所（可以处理复杂的跨司法辖区案件）。其他产品则有不同侧重点，例如CoCounsel擅长文档总结或问答功能；Spellbook经济实惠，在合同工作方面表现出色，适合中小型律所；而Tucan.ai则以“完全符合GDPR合规要求”为卖点，主要面向需要严格数据保护的欧洲客户群体。

**法律AI Agent对比**

| 维度 | Harvey AI | CoCounsel (Thomson Reuters) | Spellbook | Tucan.ai |
|---|---|---|---|---|
| 核心功能 | • 法律研究、文件分析、合同起草<br>• 诉讼支持 | • 文档总结、问答和条款修改<br>• 性能稳定 | • 合同起草、条款审查、错误检测<br>• 法律文档自动化 | • 面向GDPR合规的文档分析、转录和合同管理 |
| 准确性 | • 文档Q&A准确率达94.8%<br>• 擅长时间线生成和转录分析 | • 文档Q&A准确率89.6%<br>• 总结准确率77.2% | • 在合同审查和条款修改中表现出色<br>• 缺乏诉讼相关工具 | |
| 定制化能力 | • **提供定制化法律LLM**，可根据律所需求调整。 | • 定制化能力有限，但在通用法律任务上表现可靠。 | • 专注于简单易用的界面，适合合同相关任务，但定制化能力较弱。 | • **提供高度可定制**的合同和文档模板 |
| 集成能力 | • 无缝集成 Microsoft Word，支持多文档数据洞察。 | • 与 Thomson Reuters 生态系统集成良好，但在其他平台上的灵活性较差。 | • 界面友好，但缺乏高级集成功能，如对 MS Word 的支持。 | • 集成选项有限，仅限核心功能范围内使用。 |
| 速度 | • **响应速度最快**，大多数查询可在1分钟内完成处理 | • 速度较快，通常也在1分钟以内完成，但某些任务上略慢于 Harvey AI | • 合同相关任务响应迅速，但在全面法律工作流中不够全面 | • 速度适中，在转录任务中表现出色，但多步骤流程较慢。 |
| 安全性与合规性 | • 高安全标准（基于 Microsoft Azure）<br>• 但欧盟客户可能存在 GDPR 合规性疑虑。 | • 数据安全措施可靠<br>• 未明确提及 GDPR 合规性 | • 标准安全功能<br>• 未特别提到 GDPR 等合规框架支持 | • **完全符合 GDPR 合规要求**，非常适合欧洲律所使用 |
| 市场定位 | • 面向全球大型律所，处理复杂的跨司法辖区案件；已被 Allen & Overy 等顶级律所采用。 | • 适用于需要强大文档处理能力的律所 | • 适合专注于合同工作的中小型律所。 | • 主要面向需要严格数据保护的欧洲律所客户群体。 |

资料来源：tools for humans, Thomson Reuters, AI Tech suite, future pedia,东吴证券研究所

# 五、投资建议

# 5.1 展望：Agent迭代的Roadmap

东吴证券 SOOCHOW SECURITIES

**我们预计Agent可能按照以下roadmap进行发展，总体趋势是向着更强能力、更高可靠性、更好适应性和更广阔应用范围发展。**

| | 近期（现在-2年） | 中期（2-5年） | 长期（5+年） |
|---|---|---|---|
| | 垂直领域深耕与可靠性提升 | 能力泛化与适应性增强 | 迈向更强的自主性与通用性 |
| 通用性 | **近期内将是“垂直 Agent”的时代。**焦点会集中在那些拥有良好“环境”和清晰“反馈”机制的领域，例如 Coding Agents、软件自动化 Agents（如 Office 套件、CRM、ERP）、结构化数据处理 Agents。 | **领域知识的深化与学习**：Agent 开始积累特定领域的“经验”。通过与用户的交互、成功/失败案例的学习，Agent 在其垂直领域内的表现会持续优化，更懂用户的意图和偏好。（基于 Fine-tuning 或 In-Context Learning 的增强）<br>**“领域通用” Agent 的雏形:** 出现一些能够在一类相似垂直领域（例如，多种不同的编程语言或多种数据分析工具）工作的 Agent，展现出初步的跨领域适应性。 | **更通用的问题解决能力**：Agent 能够理解和处理更开放、更模糊的任务目标，自主进行信息检索、知识学习、复杂规划，并适应全新的环境和工具。这更接近 AGI 的愿景，但可能仍会在某些方面受限。<br>**自主学习与自我改进:** Agent 不仅能从外部反馈中学习，还可能具备一定程度的自我评估和模型/策略优化能力，实现更快的迭代和能力提升。 |
| 工具调用 | **工具调用的精细化:** Agent 将更擅长理解和调用现有的、标准化的工具 (MCP 理念)。重点在于提高工具选择的准确性、参数传递的正确性以及对返回结果的理解。 | **跨工具/跨应用协作:** Agent 将能够编排更复杂的跨应用工作流。例如，一个 Agent 可以从邮件中提取需求，在项目管理工具中创建任务，调用代码生成工具完成部分代码，然后通知相关人员。 | **复杂的多 Agent 协作:** 多个拥有不同专长的 Agent 能够动态地、高效地协同工作，共同解决极其复杂的问题，可能涉及协商、任务分配、知识共享等复杂交互。 |
| 规划能力 | **基础规划能力的提升:** 能够处理相对线性的、步骤明确的任务分解。Agent 可以将一个中等复杂度的目标拆解成一系列顺序的工具调用或代码生成步骤。 | **更强的规划能力:** 能够处理包含分支、循环、更复杂依赖关系的任务。 | |
| 记忆能力 | **有限的上下文记忆:** 主要依赖于单次任务会话中的上下文信息。 | **更长程的记忆与个性化:** Agent 能记住跨会话的上下文、用户偏好、项目特定知识，提供更个性化、更连贯的辅助。 | |
| 可靠性 | **可靠性与抗“幻觉”能力的增强:** 更加重视执行结果的验证。利用代码执行、API 调用结果校验、结构化输出检查等方式，确保每一步操作的准确性，避免错误累积。 | **更强的纠错能力：** 当某个步骤失败或遇到预期外情况时，Agent 能具备一定的自主纠错、尝试替代方案或向用户寻求澄清的能力。 | **伦理与安全框架的成熟:** 随着 Agent 自主性和能力的增强，健全的伦理规范、安全约束和可解释性机制将变得至关重要，并深度集成到 Agent 的设计中。 |
| 其他 | | **人机协作界面的优化:** 从简单的指令接收变为更具交互性的协作模式。Agent 可能会主动提问、提供选项、解释其推理过程，与人类共同完成任务。 | **对物理世界或更复杂环境的理解:** 如果与机器人技术等结合，Agent 可能具备理解和操作物理世界的能力，或者在极其复杂的模拟环境中运行。 |

资料来源：海外独角兽，东吴证券研究所

# 5.2 投资建议



**1、重视 Agent 投资窗口：**AI 技术正经历从 L2（推理者）向 L3（智能体/Agent）的关键进化，标志着 AI 从“思考”走向“行动”，这是继大模型之后的下一个重要突破口和投资风口 。Agent 的发展由技术成熟、标杆产品验证和市场需求共同驱动。 Agent 作为下一代 AI 应用形态，具备深度自动化复杂任务、指数级提升效率、解放人类生产力的潜力，并将重塑互联网入口格局，是通往 AGI和具身智能的关键环节 。我们预计 2025 年下半年可能开启入口级通用 Agent 的竞争。需密切跟踪基础模型（尤其多模态、推理、规划）、强化学习、工具调用可靠性、推理成本优化以及标准化协议（如 MCP）的进展。

**2、长期配置平台巨头：**拥有强大基础大模型、算力、数据和生态系统的大型科技平台公司是 Agent 时代的核心受益者，最有可能主导通用 Agent 的发展，并能整合或取代单一功能应用，具备长期配置价值 。 例如海外的 Google、微软，以及与 OpenAI、Anthropic 深度绑定的公司；国内的阿里、腾讯、字节（未上市）。

**3、关注垂直领域领跑者：**在通用 Agent 能力尚未完全成熟之前，那些在特定垂直赛道（如软件开发/编程、法律、金融、特定行业 B2B 服务等）已经建立深厚领域知识壁垒、拥有清晰商业模式和客户基础的垂直 Agent 提供商具有较高的短期增长潜力 。 重点评估这些垂直厂商的领域知识独特性、工作流整合能力，以及其相对于通用 Agent 的护城河深度 。

我们认为编程领域会是最快落地、最先实现PMF和商业化的领域：尤其值得关注，已有成功案例（如 Cursor、Devin）。需关注相关公司的产品市场契合度（PMF）、增长速度及竞争格局（如 Devin、Windsurf、GitHub Copilot 的迭代）。

其他垂直应用也值得关注：我们总结了30家上市公司在垂类Agent方面的布局，其产品基本符合Agent定义且具有垂直领域的比较优势。例如出版校对（果麦文化）、电商外贸（焦点科技）、企业服务（创业黑马）、美学设计（美图公司）等。关注其利用 AI Agent 解决具体行业痛点的能力和商业化进展 。

**4、警惕“浅层套壳”风险：**避免投资那些仅仅是对底层大模型进行简单封装、缺乏核心技术或产品壁垒（如复杂工作流编排、高质量工具集成、深度领域知识）、容易被复制或被平台能力取代的公司。

# 六、风险提示



**1.技术成熟度风险：**AI Agent 在执行复杂、多步骤任务时的可靠性、稳定性仍不足，幻觉问题可能被放大，与外部工具/环境交互的失败率较高，可能无法达到预期效果。

**2.高成本风险：**Agent 执行任务需要多次调用大模型进行推理和规划，导致高昂的算力成本和 API 调用费用，可能成为商业化和大规模推广的瓶颈。

**3.商业模式不确定性风险：**如何定价以平衡成本和价值创造，找到可持续的商业模式仍是挑战。企业客户可能因 ROI 不明确或集成困难而犹豫部署。

**4.竞争加剧风险：**科技巨头凭借资源优势可能挤压初创公司空间。通用 Agent 能力的提升可能威胁垂直 Agent 的生存空间。技术快速迭代可能导致现有产品迅速过时。

# 免责声明

东吴证券股份有限公司经中国证券监督管理委员会批准，已具备证券投资咨询业务资格。

本研究报告仅供东吴证券股份有限公司（以下简称“本公司”）的客户使用。本公司不会因接收人收到本报告而视其为客户。在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议，本公司及作者不对任何人因使用本报告中的内容所导致的任何后果负任何责任。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。

在法律许可的情况下，东吴证券及其所属关联机构可能会持有报告中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投资银行服务或其他服务。

市场有风险，投资需谨慎。本报告是基于本公司分析师认为可靠且已公开的信息，本公司力求但不保证这些信息的准确性和完整性，也不保证文中观点或陈述不会发生任何变更，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。



**东吴证券投资评级标准**

资评级基于分析师对报告发布日后6至12个月内行业或公司回报潜力相对基准表现的预期（A 股市场基准为沪深 300 指数，香港市场基准为恒生指数，美国市场基准为标普 500 指数，新三板基准指数为三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的），北交所基准指数为北证50指数），具体如下：

公司投资评级：

买入：预期未来6个月个股涨跌幅相对基准在15%以上；

增持：预期未来6个月个股涨跌幅相对基准介于5%与15%之间；

中性：预期未来 6个月个股涨跌幅相对基准介于-5%与5%之间；

减持：预期未来 6个月个股涨跌幅相对基准介于-15%与-5%之间；

卖出：预期未来 6个月个股涨跌幅相对基准在-15%以下。

行业投资评级：

增持： 预期未来6个月内，行业指数相对强于基准5%以上；

中性： 预期未来6个月内，行业指数相对基准-5%与5%；

减持： 预期未来6个月内，行业指数相对弱于基准5%以上。

我们在此提醒您，不同证券研究机构采用不同的评级术语及评级标准。我们采用的是相对评级体系，表示投资的相对比重建议。投资者买入或者卖出证券的决定应当充分考虑自身特定状况，如具体投资目的、财务状况以及特定需求等，并完整理解和使用本报告内容，不应视本报告为做出投资决策的唯一因素。

东吴证券研究所

苏州工业园区星阳街5号

邮政编码：215021

传真：（0512）62938527

公司网址： http://www.dwzq.com.cn

东吴证券 SOOCHOW SECURITIES

# 东吴证券 财富家园