西南证券 SOUTHWEST SECURITIES

AI专题·Agent

智能体基建厚积薄发，商业化应用曙光乍现

西南证券研究院
海外研究团队
2025年4月

核心观点

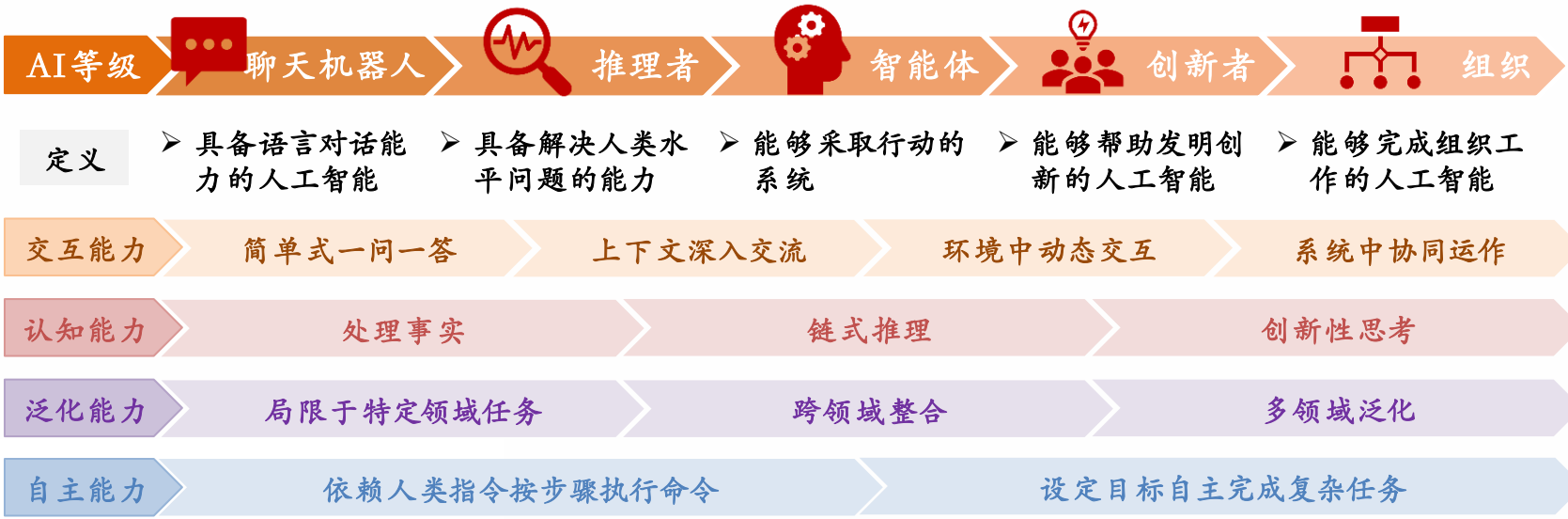
- **AI发展阶段从推理者走向智能体，模型底座智能水平提升。**目前，AI发展水平正从推理者向智能体Agent演进，AI产品逐步能够理解目标、具备外部记忆和推理能力，相关智能体产业链正经历从模型能力提升到应用商业化的系统性跃迁。AI大模型能力由预训练、后训练、测试时三条扩展曲线推进，其中，预训练奠定模型内部智能上限，后训练及测试时扩展分别释放模型在特定领域和推理方面的潜力，当前基础模型迭代放缓，逐步从训练扩展向测试时扩展转变，主次曲线迎来切换，从而对大规模集群依赖程度下降、推理算力需求攀升，更加聚焦AI产品的商业化能力和生态建设。
- **中间工具厚积薄发，开发者生态积极构建。**在中间层，智能体生态所需的通信协议与开发工具快速涌现，Anthropic MCP协议、谷歌A2A协议等代表性技术正助力构建智能体新型操作系统，为模型与工具、智能体与智能体之间建立统一的交互接口。其中，2025年3月MCP Server发现平台Smithery的服务器创建数量较2月实现3倍增长，A2A已得到50多家合作伙伴的支持，开发者生态加速繁荣。开发工具与底层框架的标准化，可类比为互联网时代移动手机的USB-C接口，或者类比为用于App和操作系统之间通信的Android API，将加速AI智能体商业化进程。
- **初代产品创收加速，商业化应用曙光乍现。**在应用层，智能体应用分为跨行业通用产品和垂类专业产品，前者发展相对成熟，部分产品已开始规模化应用，后者商业化起步略晚，但有望成为B端数智化转型的重要抓手。目前，智能体作为交互式AI产品开始快速落地，初代产品Cursor、Glean等已实现上亿美金年经常性收入（ARR），展现出较高成的长潜力，并出现基于实际交付成果、任务完成率等指标的新收费模式。整体来看，AI智能体产品正形成“底层模型能力升级+中间工具繁荣+商业场景落地”的基础设施与应用协同的演进路径，未来AI智能体应用还需进一步提升规划能力、具备更好的记忆、拥有更强的多模态理解能力，释放变现潜力。
- **相关标的：1）推理算力：**英伟达、博通；**2）中间工具和数据层：**谷歌、Snowflake；**3）下游应用：**Salesforce、SAP、Shopify；**4）云服务：**亚马逊、微软、谷歌。
- **风险提示：**AI技术进展不及预期；AI商业化进展不及预期；投资回报不及预期等风险。

目 录

1.1 AI等级：AI发展水平划为五大等级，当前正从推理者转向智能体

- **模型多维能力持续提升，AI从推理者转向智能体。**根据OpenAI对AI发展的理解和定义，AI水平可分为五大等级：一是聊天机器人(Chatbot)，能够用自然语言进行对话；二是推理者，基于推理模型，解决人类级别的智力问题；三是智能体(Agent)，能够代表用户采取行动；四是创新者；五是组织。过去，在ChatGPT等聊天机器人产品推出时，大模型通常采取一次性推理，用户与聊天机器人的交互形式呈现为简单的一问一答。而在推理模型的不断发展之下，AI模型逐渐能够与自己对话，实现内部思考，具备推理能力。当前，随着大模型在交互/认知/泛化/自主等多维度能力持续提升，AI正从推理者转向智能体，逐步具备采取行动及处理任务的能力，智能体产品加速推进。

AI发展水平划分为五大等级

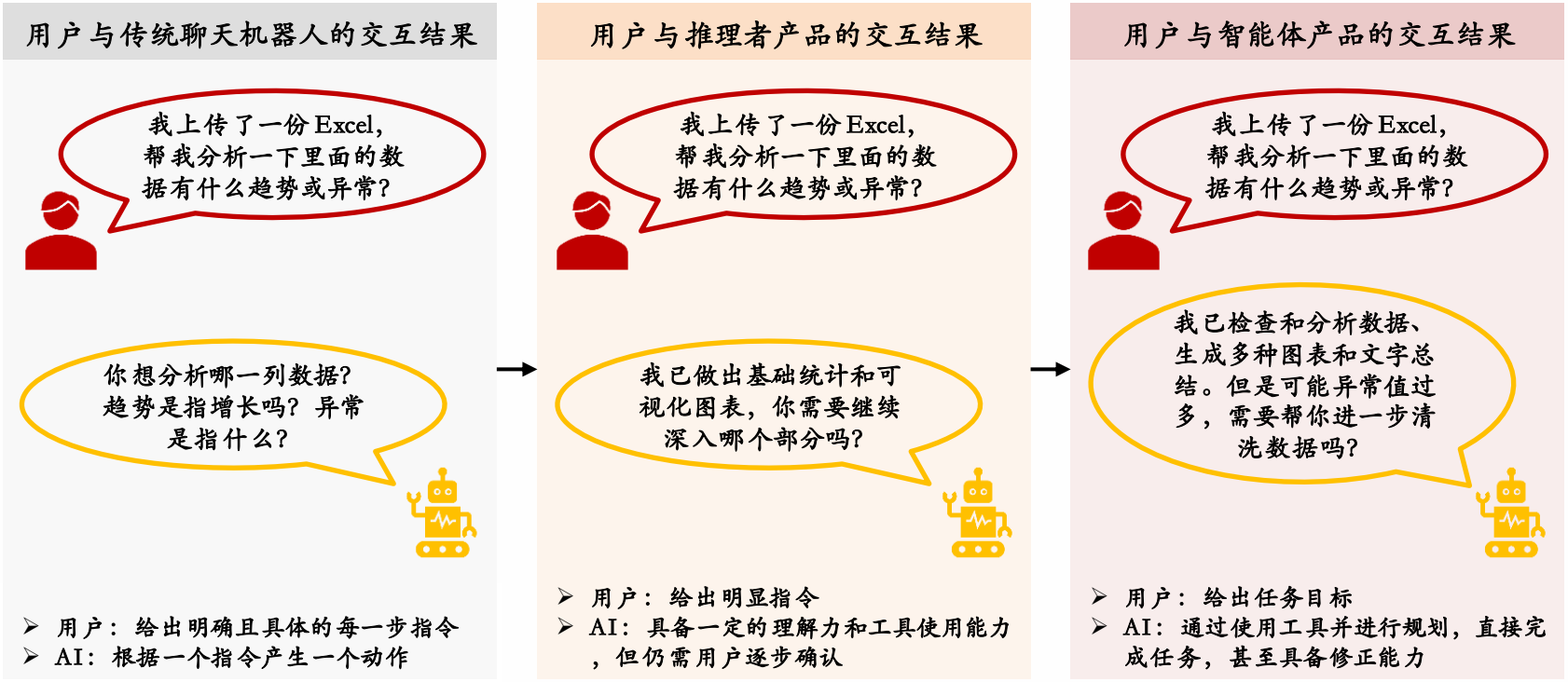


资料来源：Bloomberg reporting，OpenAI，西南证券整理

1.1 AI等级：AI发展水平划为五大等级，当前正从推理者转向智能体

- **AI产品目前处于中间过渡形态，智能体有望革新交互效率。**过去，传统聊天机器人只能执行明确指令，用户需要逐次下达任务指令，AI模型根据一个指令进行一个动作；当前，中间形态的AI产品已初步具备目标理解和推理能力，可以根据用户的模糊需求主动采取一部分行动，但仍然需要依赖用户反馈进行下一步操作；未来，真正的AI智能体将能够根据最终目标自主规划任务步骤、调用多种工具、识别错误并给出修正策略，具备完成任务的能力。

用户与不同AI产品形态的互动以及第一轮交互结果示例

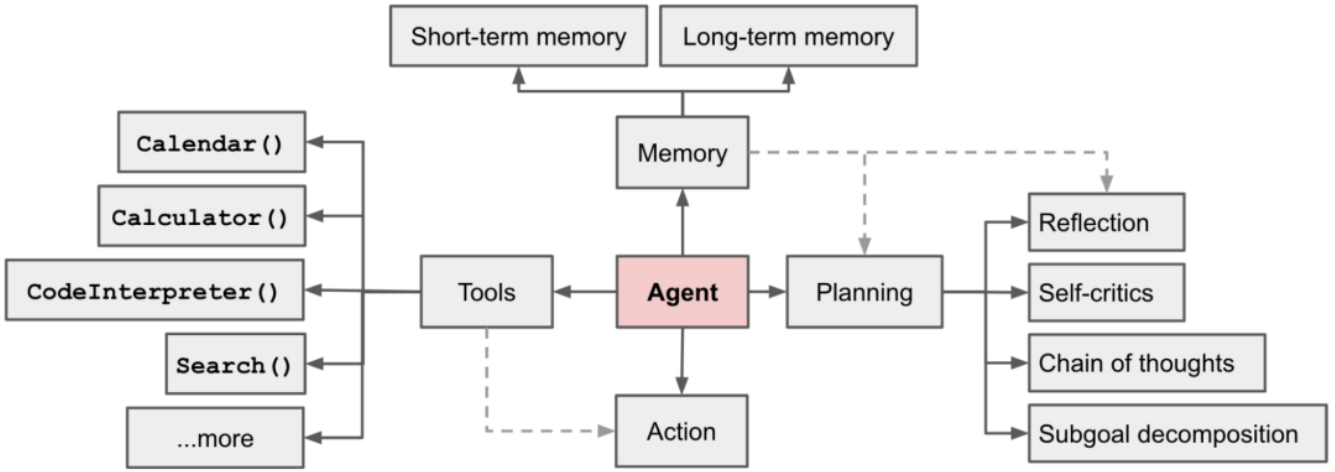


资料来源：西南证券

1.2 Agent等级：初阶能够使用工具，高阶可自主完成长时任务

- **智能体(Agent)=大模型(LLM)+记忆(Memory)+主动规划(Planning)+工具使用(Tool use)。**
- **大模型：**在基于LLM的智能体中，LLM充当智能体的大脑。
- **主动规划：**可以将大型任务分解为子任务，并规划执行任务的流程，同时能够对任务执行的过程进行思考和反思，从而决定是继续执行任务，或判断任务完结并终止运行。
- **记忆：**短期记忆指在执行任务的过程中的上下文，会在子任务的执行过程产生和暂存，在任务完结后被清空；长期记忆即可以长时间保留的信息，一般指外部知识库，可用向量数据库存储或检索。
- **工具使用：**为智能体配备工具API，如计算器/搜索工具/代码执行器/数据库查询工具等，从而与物理世界实现交互，解决实际问题。

智能体公式：Agent = LLM + Memory + Planning + Tool use

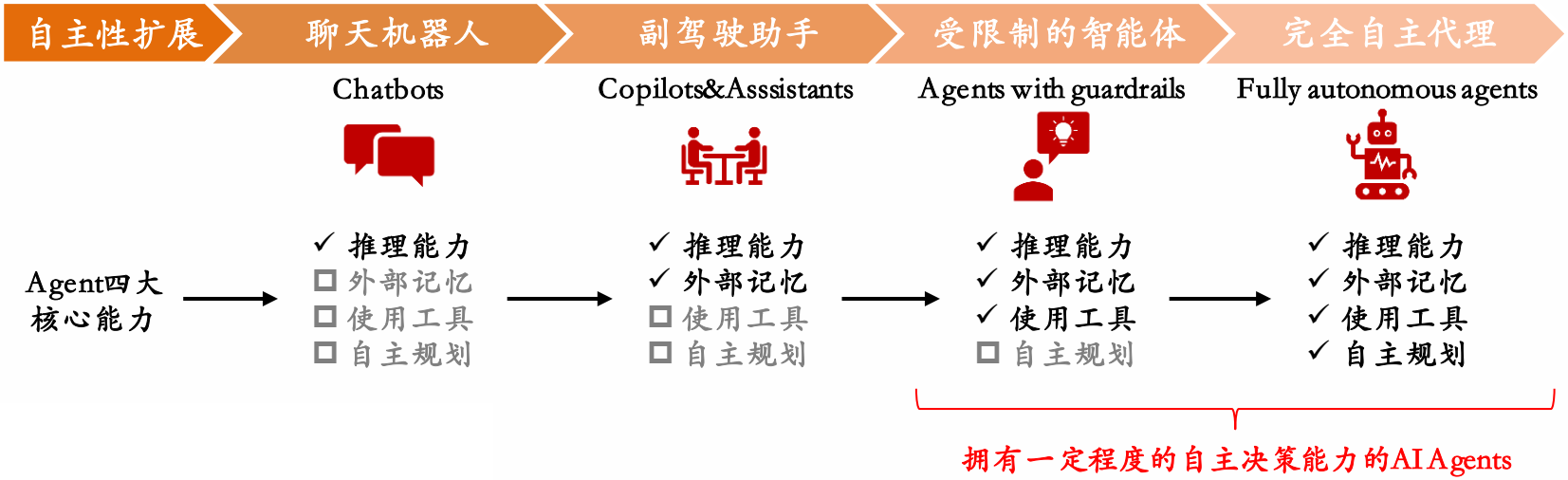


资料来源：OpenAI官网，西南证券整理

1.2 Agent等级：初阶能够使用工具，高阶可自主完成长时任务

- **自主决策能力是基础，解决长时任务是关键。**根据智能体“推理+记忆+使用工具+规划”的四大核心能力来看，截至目前，聊天机器人产品逐步具备推理能力，副驾驶和工具型助手可以建立外部记忆，但仍然不具备使用工具和自主规划的能力，只能根据用户指令按步骤执行，不属于能够自主决策的智能体。根据CBInsights研究，具备一定自主决策能力的智能体可分为两大等级。**1）初级Agent：**AI模型可以通过编排组合，把重复性高、需要一定灵活性的任务从人替换成数字员工，实现任务的自动化，在该阶段，Agent尚不具备完全开放的决策空间，自主决策范围主要局限于任务流程和有限选项之中，决策行为将受到安全条件、访问权限等限制，**是有限决策的智能体。2）高级Agent：**智能体不只是LLM calling的组合，而是能够更自主、更主动地规划，完成多步骤、长时任务，可以在多个选项之间做出自主选择，不需要人工指示，**实现高度自治。**

根据自主化程度划分AI Agent等级

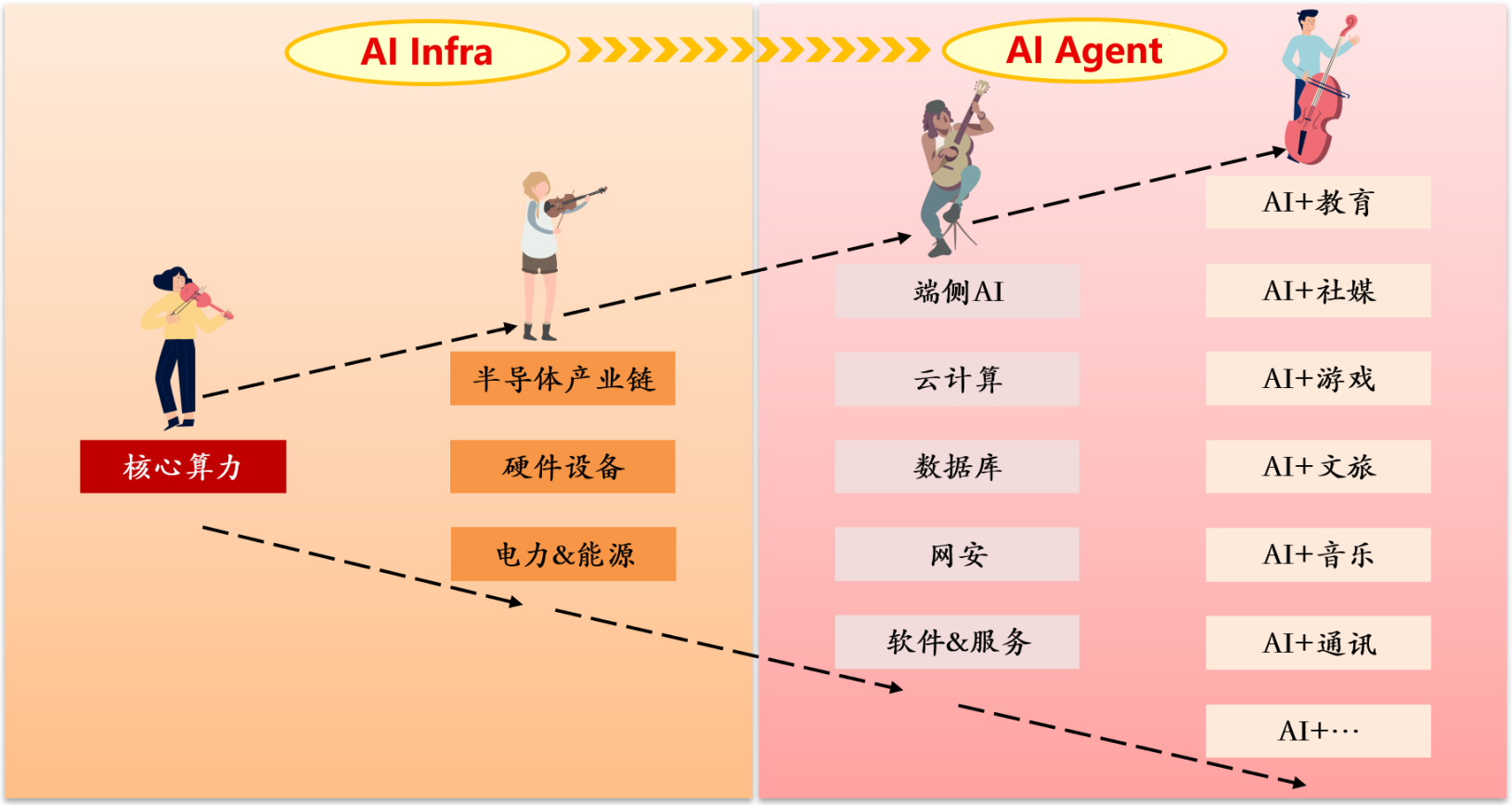


资料来源：CBInsights，西南证券整理

1.3 AI产业链：AI Infra奏响主旋律，AI Agent拉开新画布

- **AI Infra**：核心算力、半导体产业链、硬件设备、电力能源为AI大模型的训练与推理奠定硬件基础。
- **AI Agent**：B端软件、C端应用、端侧AI及具身智能等环节在Agent应用上蓄势待发。

AI产业链示意图

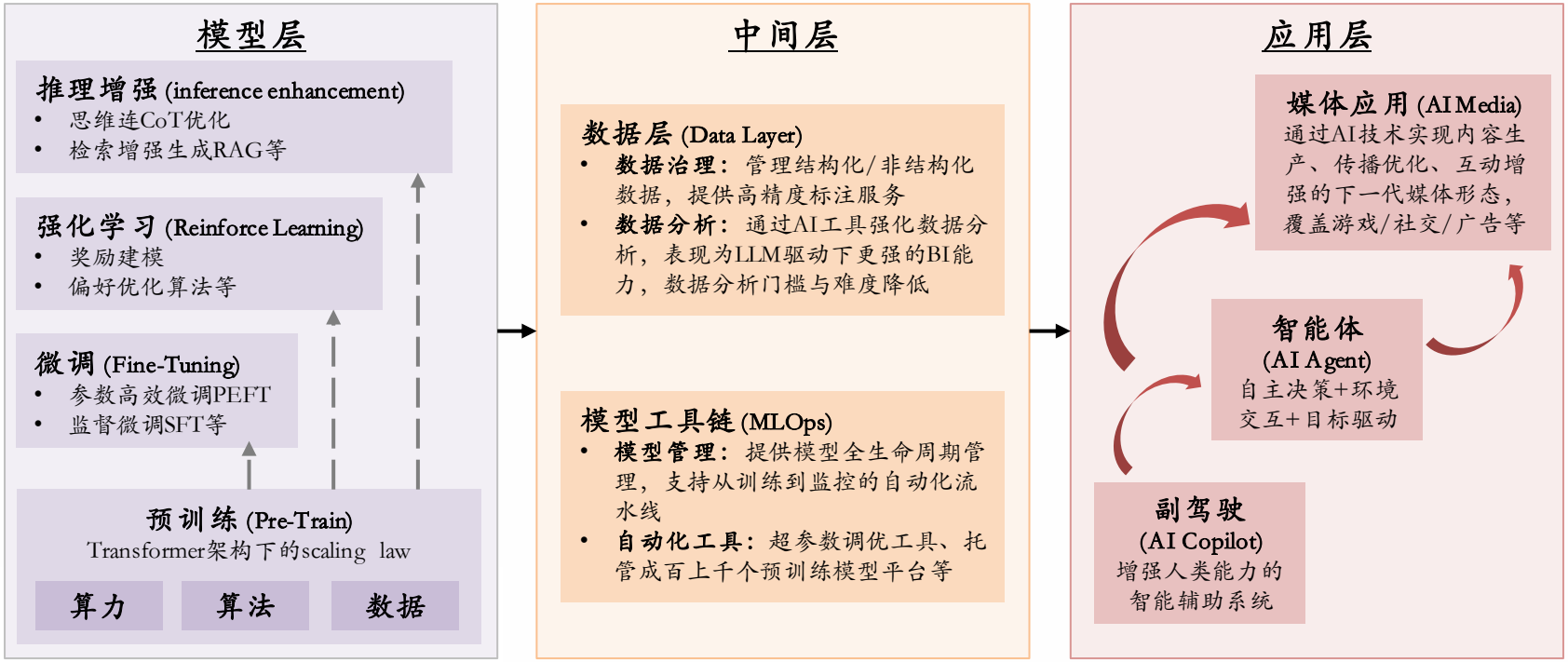


资料来源：西南证券

1.3 Agent产业链：智能体基建厚积薄发，商业化应用曙光乍现

- **模型层**：Agent本质是大模型能力的工程化载体，大模型智能水平仍是打造Agent的底层支撑，未来依旧需要通过预训练、后训练和测试时计算进行扩展。
- **中间层**：Agent产业链的中间层工具正加速构建，数据库、身份治理、通信协作等成为重要议题。
- **应用层**：Agent应用形态随着以上底层大模型和中间原生基础设施的发展逐步从构想更加贴近现实。

AI Agent产业链示意图

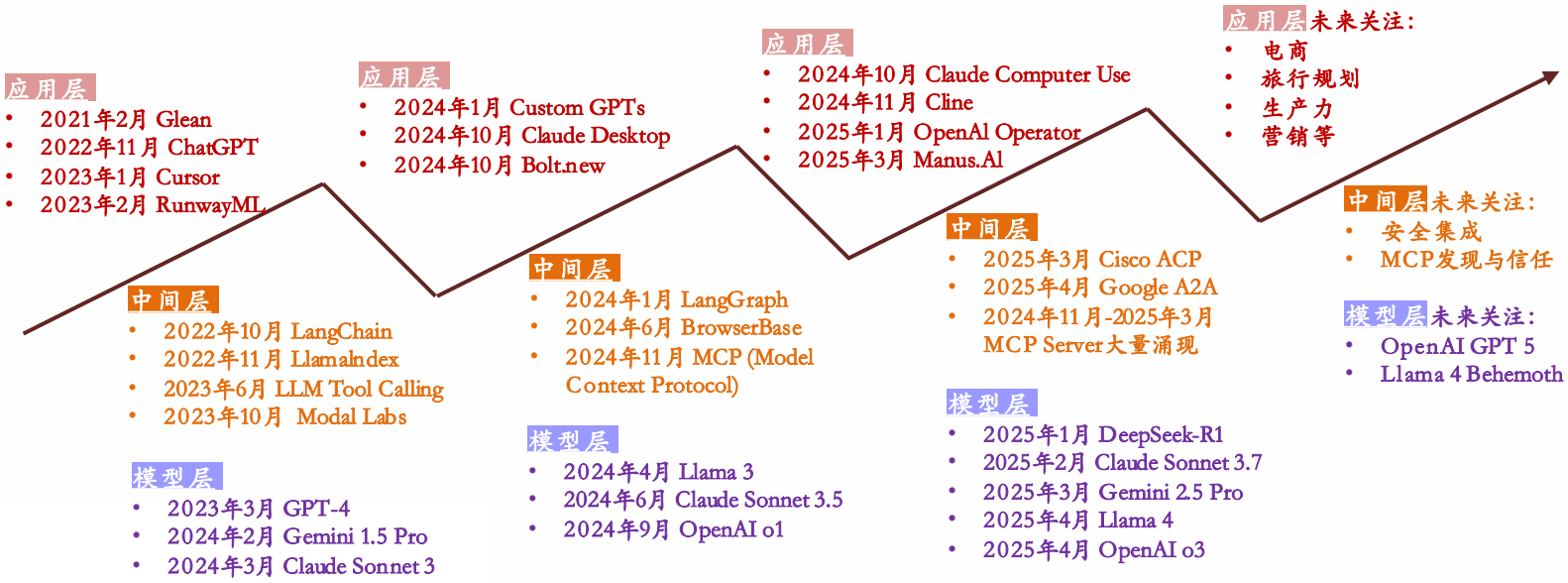


资料来源：Z Research，西南证券整理

1.3 Agent产业链：智能体基建厚积薄发，商业化应用曙光乍现

- **基础设施加速发展推动新应用诞生，新应用积极引导基础设施下一步健全方向。**智能体生态正在经历波浪式发展进程，每一波创新应用的诞生，都会带动基础设施的迭代升级，底层技术的进步又会进一步催生出更智能的应用，如OpenAI的GPT系列（从GPT-1到GPT-4）和o系列（从o1到o3）模型、Anthropic的Claude模型（Sonnet-3迭代至3.7）、谷歌Gemini模型（从1.5Pro迭代至2.5 Pro）。智能体中间层则陆续出现LangChain、Tool Calling、MCP和A2A等工具；应用层相继出现Cursor、Claude Desktop、OpenAI Operator等。新应用对基础设施提出更复杂的需求，基础设施的进步又将反哺新的智能体应用，两者相互塑造、共同演进，加速AI的商业化落地。

Agent基础设施建设与应用协同演进



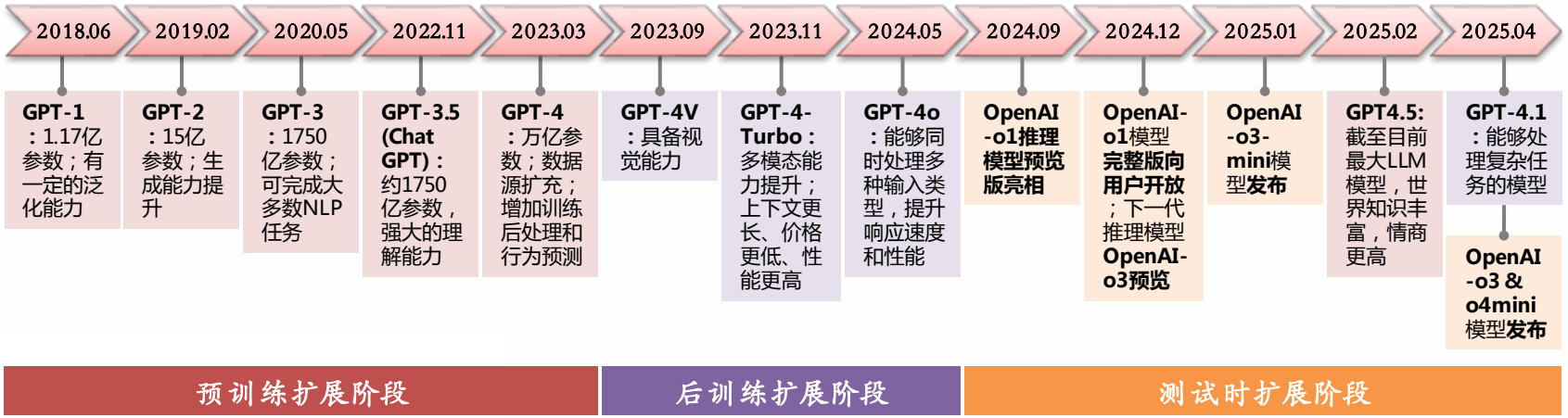
资料来源：Madrona，西南证券整理

目 录

2.1 AI模型扩展法则：扩展法则迎来范式转变，主次扩展曲线逐步切换

- **扩展法则迎来范式转变，推理模型迭代节奏加速。**从OpenAI旗下AI模型迭代进程来看：
  - **2018年6月至2023年3月——预训练扩展阶段：**OpenAI大模型预训练快速推进，在五年内从GPT-1迭代至GPT-4模型，模型基础能力2023年3月之后，预训练扩展进程逐渐放缓，截至目前仍未推出下一代预训练大模型GPT-5。
  - **2023年下半年至2024年5月——后训练扩展阶段：**基于微调技术开始打磨多模态、上下文等能力，提升特定指标性能。
  - **2024年9月至今——测试时扩展阶段：**2024年9月OpenAI-o1模型预览版亮相，标志正式进入推理模型时代；2025年4月17日，OpenAI推出完整版o3模型和o4-mini模型，截至目前，半年内已迭代多次，测试时扩展正加速发展。

OpenAI旗下模型迭代进程梳理



资料来源：OpenAI官网，西南证券整理

2.1 AI模型扩展法则：扩展法则迎来范式转变，主次扩展曲线逐步切换

- **规模法则从训练阶段延伸至推理阶段，推动计算需求持续提升。**预训练法则和后训练法则均与模型的训练阶段有关，而测试时扩展法则与推理阶段有关，深度推理有望对算力需求进一步增加。

① **预训练扩展法则(Pre-training Scaling Law)：**关注计算资源、模型大小和训练数据三大要素，当三要素同时增加时，模型性能将同步提升，打造优质基座模型。

② **后训练扩展法则(Post-training Scaling Law)：**关注在预训练完成后对模型的进一步优化和微调，可以针对特定任务进行改进，从而提升模型在特定领域的性能，有助于打造垂类模型。

③ **测试时扩展法则(Test-time Scaling Law)：**针对在模型的实际推理或应用中，根据问题的复杂程度实时分配计算资源，面对复杂问题能够进行分步骤、多阶段推理，在多个解法中寻求最优解。

AI三大扩展法则对比

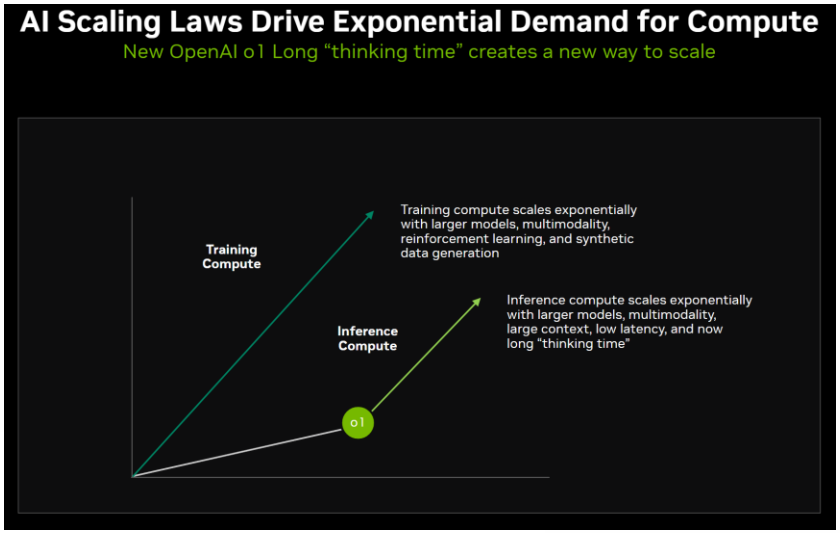
名称	预训练扩展法则 (Pre-training Scaling Law)	后训练扩展法则 (Post-training Scaling Law)	测试时扩展法则 (Test-time Scaling Law)
阶段	**训练（training）**阶段的scaling law	**训练（training）**阶段的scaling law	**推理（reasoning）**阶段的scaling law
定义	指在**训练过程中**，通过增加**训练数据**、**模型参数**和**计算资源**来提升模型能力。模型在预训练阶段需要通过大量数据进行训练、学习基础知识。	指在**预训练之后**，利用**强化学习**、**人工反馈**等技术对模型进行进一步优化，后训练通常涉及对模型的**精细化调整**，提高在**特定任务**上的表现。	指在模型的**实际应用中**，模型根据需要**动态分配计算资源来提升推理效率**，更加关注模型如何在实时推理时优化自身的计算策略。
特点	① 主要依赖大量的**数据**（多模态数据，如文本、图像、视频等）和**计算资源**。 ② 模型通过大规模数据集进行自我学习，获取广泛的知识。 ③ 是训练过程中的初步阶段，主要帮助模型建立**基座能力**。	① 在模型**初步训练完成后**，使用**人类反馈或强化学习**的方式帮助模型在特定任务上改进。 ② 强调通过模拟**“自我提升”**的方式逐步提升模型能力（如通过解决复杂的数学问题等）。 ③ 可以看作是**“训练后的进步”**或**“微调”**过程，帮助模型在**特定领域变得更精通**。	① 重点是在**实际使用**中，通过**调整计算资源的分配**来提升**决策过程的质量**。 ② 模型可以在推理时进行**“深度思考”**，将问题**拆解成多个步骤**，逐步推理或**产生多个解法，评估最优解**。 ③ 强调推理过程中的**灵活性和高效性**，目的是在实时环境中产生高质量的结果。
模型	OpenAI GPT系列模型：**GPT-1至GPT-4**	GPT-4 turbo	**OpenAI-o1至o3**系列模型

资料来源：英伟达CES大会，西南证券整理

2.1 AI模型扩展法则：扩展法则迎来范式转变，主次扩展曲线逐步切换

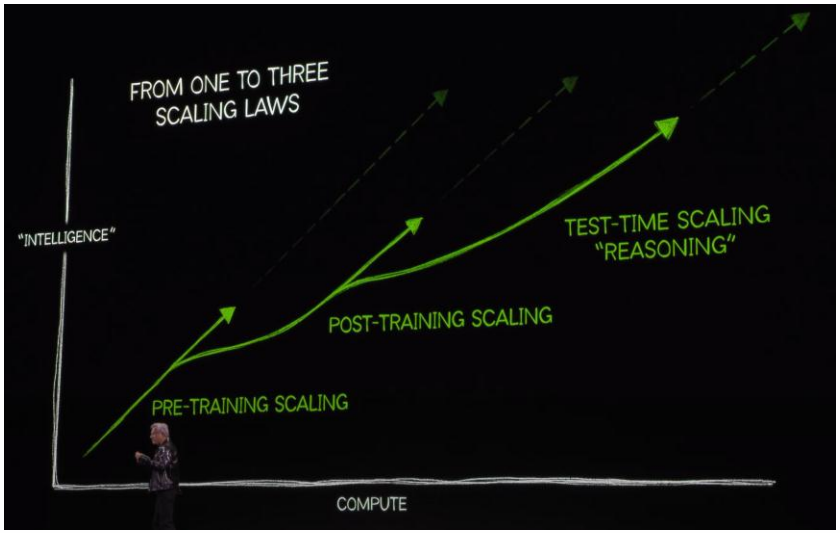
- **模型性能提升路径持续探索，主次增长曲线发生转变。**2020年1月和2022年3月，OpenAI和谷歌先后发布论文《Scaling Laws for Neural Language Models》和《Training Compute-Optimal Large Language Models》，两者认为**预训练阶段的扩展法则是提升大语言模型性能的有效路径。**2024年8月谷歌发表论文《Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters》，提出**在测试时对大语言模型的计算进行最优扩展，可能比扩展模型参数来提升模型性能更有效。**根据英伟达CES大会信息，除预训练和后训练扩展法则之外，测试时扩展法则同样推动算力需求持续增长，以OpenAI-o系列模型为代表的推理模型通过测试时计算，带动推理算力高增，Scaling Law持续有效。随着模型性能提升曲线从训练扩展转向推理扩展，投资方向也随之向推理侧转变。

英伟达：扩展法则推动算力需求提升



资料来源：英伟达CES大会，西南证券整理

英伟达：扩展法则从一个扩展至三个

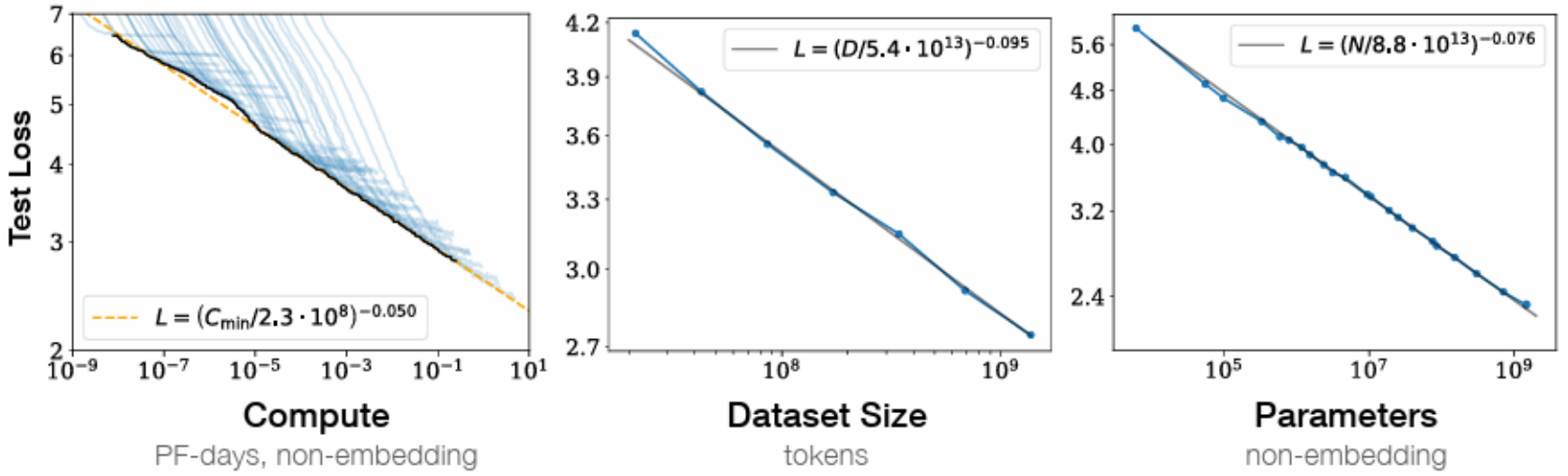


资料来源：英伟达CES大会，西南证券整理

2.2 预训练扩展：三要素影响模型性能，高质量数据成为瓶颈

- **算力决定Transformer模型性能上限，模型参数与训练数据比例影响模型最佳性能。**根据OpenAI和Google相关研究，模型性能随着模型参数大小、训练数据集大小、计算量的增加而提高。对于基于Transformer架构的大语言模型，模型性能三要素的关系为C≈6N*D，其中，**N**代表模型参数规模，**D**代表预训练数据集大小；**C**代表预训练算力资源。大语言模型若要获得最佳性能，需同时扩展三大要素。**当其中一个因素受限时，模型的智能表现可以随着另外两个因素的增加而变好，但边际效应会逐步递减。在给定预训练计算量的情况下，可以确定最佳的参数量和数据集之比，从而确定模型的最佳能力。因此，在总计算量越多的情况下，模型能力的上限会越高。**然而当前**模型参数量与数据量的扩展比例尚存争议**，OpenAI在论文中指出模型参数规模比数据集大小更重要，两者比例在0.73：0.27时计算效率最优；谷歌论文则认为模型参数和数据大小同等重要，随着预训练计算资源的增加，模型参数量和训练数据量应该等比例增长。

预训练扩展法则“算力+数据+模型”三要素

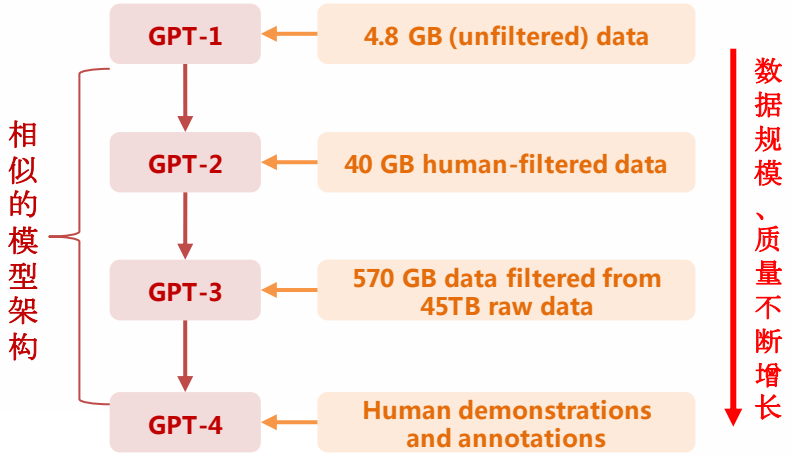


资料来源：OpenAI《Scaling Laws for Neural Language Models》，西南证券整理

2.2 预训练扩展：三要素影响模型性能，高质量数据成为瓶颈

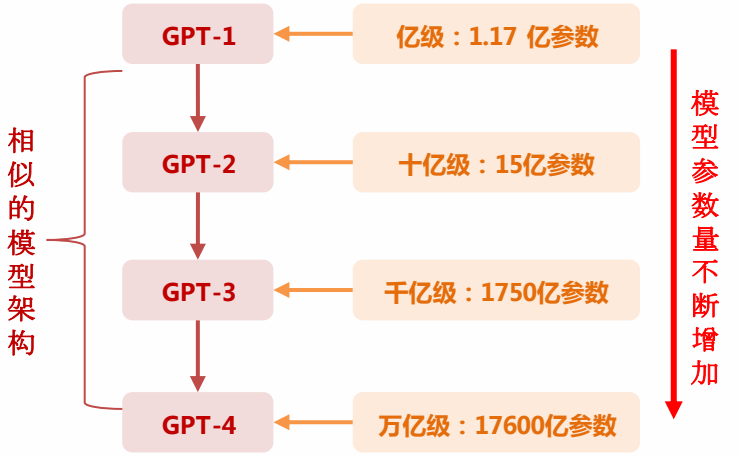
- **OpenAI GPT系列模型持续扩展，高质量语料成为瓶颈。**近年来，OpenAI GPT系列模型实现快速扩展，从2018年的GPT-1到2020年的GPT-3，模型参数量从1.17亿增长至1750亿，数据量从4.8GB增加至570GB，数据质量从原始数据提升至过滤后的高质量数据。2023年3月，GPT-4模型再次实现扩展，参数量达到万亿级别，数据规模和质量进一步提升，模型性能实现再次跃升。然而GPT-4发布至今，已将近两年，OpenAI仍未发布下一代GPT模型。由于模型能力的指数级增长离不开算力和数据资源的同步增加，目前OpenAI可能遇到数据增长跟不上模型性能提升诉求的问题，因此OpenAI在寻求更多预训练数据的同时，逐步转向结合Re-train、Post-train和Test-time compute，或研究更高效的模型架构等方式，以解决数据瓶颈导致模型性能提升放缓的挑战。

OpenAI GPT系列模型训练数据持续增长



资料来源：OpenAI官网，西南证券整理

OpenAI GPT系列模型参数规模持续增长



资料来源：OpenAI官网，西南证券整理

2.3 后训练扩展：微调技术持续创新，打造模型特定性格

- **后训练扩展：在后训练期间不断自我改进，强化特定任务能力。**后训练是在预训练之后，通过使用人工反馈、强化学习等方法来进一步提升模型响应能力。在后训练阶段，模型通常会根据反馈机制进行微调，相较于预训练阶段的全数据集训练，后训练的计算需求较低，不需要在庞大的训练样本上迭代，而是基于模型已经学习到的知识进行微调，根据反馈在后训练期间不断进行自我改进，选择性地对某些任务或场景实现强化。
- **微调技术持续创新：OpenAI于2024年12月发布会推出o1强化微调、偏好微调等技术。1）强化微调(Reinforcement Fine-Tuning)：**通过结合强化学习和监督微调，针对特定领域打造专家模型。在强化微调技术下，开发人员只需利用少量训练数据即可创建特定领域的专家模型，通过使用几十到几千个高质量数据，模型能够通过强化学习自行探索和学习如何推理复杂任务。**2）偏好微调(Preference Fine-Tuning)：**通过使用直接偏好优化（DPO）来比较模型响应对，教会模型区分偏好和非偏好输出，有助于在语气、风格和创造性等主观任务上尤为有效。

监督微调/强化微调/偏好微调方法对比

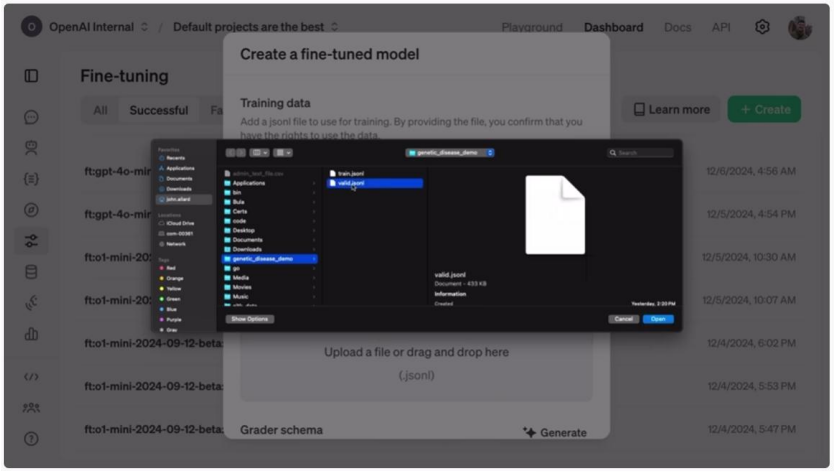
特点	监督微调	强化微调	偏好微调
定义	在已经预训练的模型基础上，使用标注好的数据集进行进一步的训练。模型通过**输入-输出对**的方式学习，从而调整权重和参数。	通过强化学习方法对预训练模型进行进一步训练。在强化微调中，模型与环境互动，**基于执行的动作获得奖励或惩罚。**	在预训练的基础上，通过用户反馈、偏好评分、**针对性的主观反馈来优化模型**，使其符合特定的偏好或需求。
训练数据	标注数据 (输入-输出对)	环境交互和人类反馈 (奖励信号)	人类偏好反馈 (选择/评分/建议等)
优化技术	监督学习，通过最小化预测误差优化	强化学习，通过奖励优化行为	基于人类反馈/选择或评分/偏好反馈优化输出，符合用户需求
目标	提高模型在特定任务上的准确性	优化模型行为，便其适应复杂环境	优化模型输出，便其符合用户的个性化需求
应用场景	分类任务、生成任务、回归任务、情题分析、机器翻译等任务	对话系统、游戏AI、对话系统	个性化对话系统、个性化推荐等

资料来源：OpenAI官网，西南证券整理

2.3 后训练扩展：微调技术持续创新，打造模型特定性格

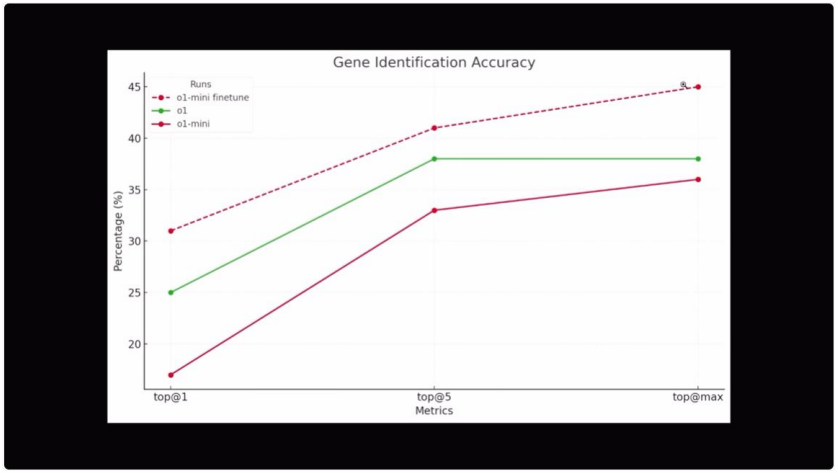
- **打造模型特定性格，丰富垂类交互体验。**在预训练环节，模型主要针对基础能力进行提升；而在后训练环节，模型专注打造自身特点和性格。根据OpenAI Day-2发布会，OpenAI通过演示生物医学微调模型案例，微调出o1-mini finetune模型，用于分析发病症状、鉴定基因，最终微调后的o1-mini模型得分提高80%，超越o1正式版，在生物医学方面专业性能提升明显，垂类模型应用前景广阔，B端科研领域有望受益。目前，OpenAI的强化微调技术已对企业、大学和研究院开放申请测试通道，并将于2025年春季作为产品发布并向用户开放。在OpenAI内部测试中，强化微调技术已在生物化学、安全、法律、医疗保健领域取得成功，未来有望助力泛科研实现突破性进展。此外，随着后训练技术的持续发展，垂类模型的打造有望更加专业和精细，也将要求产品经理能够更准确地理解AI产品特性，提升模型的后训练效率，实现用户与垂类模型的交互体验升级。

OpenAI利用强化微调技术创建模型



资料来源：OpenAI官网，西南证券整理

OpenAI o1强化微调后专业性能提升

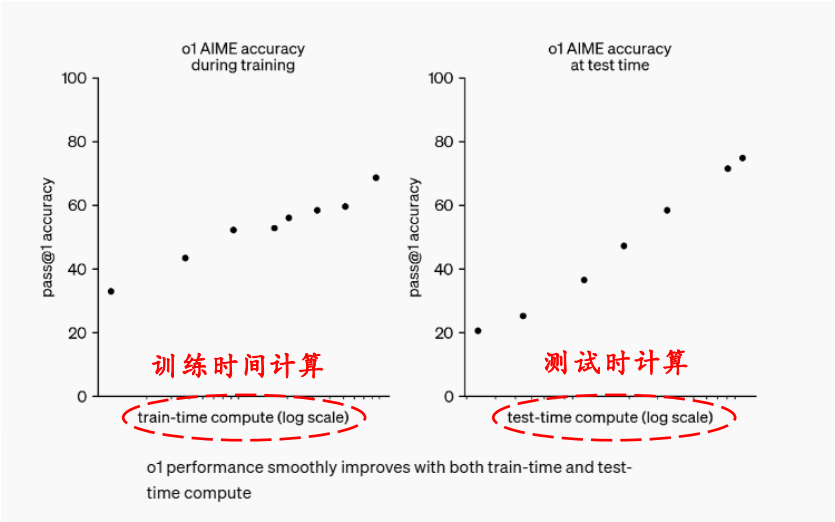


资料来源：OpenAI官网，西南证券整理

2.4 测试时扩展：模型实现深度推理，Agent落地未来可期

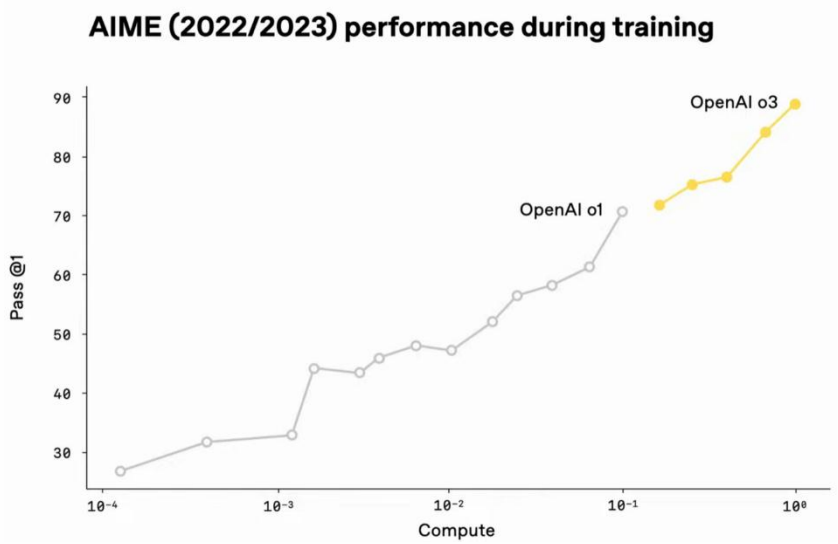
- **大模型推理能力持续提升，奠定智能体规划能力基础。**在数据瓶颈挑战下，大模型厂商从训练扩展转向测试时计算，打造AI新扩展曲线。在此背景下，OpenAI推出推理模型，于2024年9月公开介绍OpenAI-o1模型；2025年12月向用户开放完整版o1模型，并预告下一代OpenAI-o3模型；2025年4月推出完整版o3模型。**在o系列推理模型中，OpenAI引入测试时计算（Test-time compute），使模型能够根据用户提问调节思考行为、分配计算资源、优化输出结果、提升模型性能。**2024年9月，OpenAI-o1推理模型的推出标志着测试时Scaling Law开启；2025年4月，OpenAI-o3模型公开信息表示Scaling Law持续，**AI模型仍在继续扩展Train-time和Test-time，o3模型的训练计算量是o1的10倍**，其中，RL范式中的算法优化进一步推动模型性能提升。

OpenAI-o1模型性能随测试时计算实现性能提升



资料来源：OpenAI官网，西南证券整理

OpenAI-o3模型随强化学习扩展实现性能提升

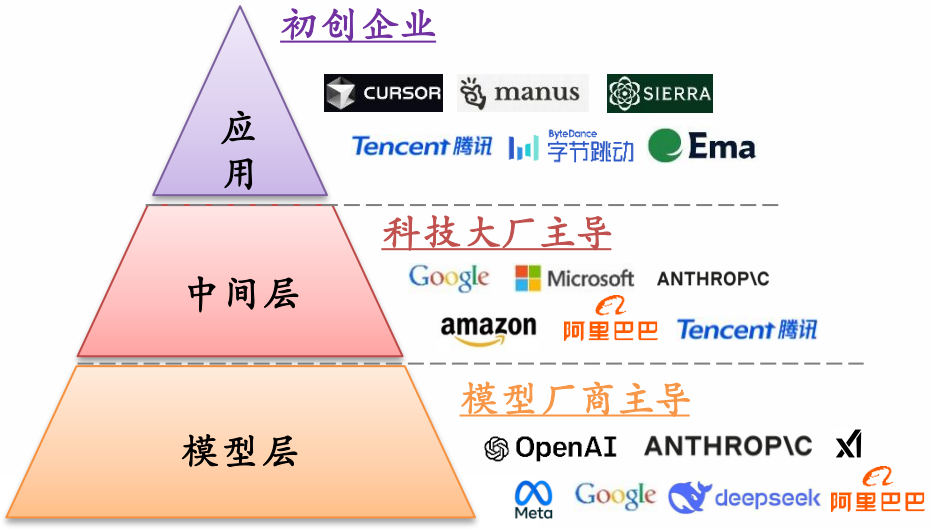


资料来源：OpenAI官网，西南证券整理

2.5 AI模型扩展循环：智能水平仍需提升，大模型扩展持续进行

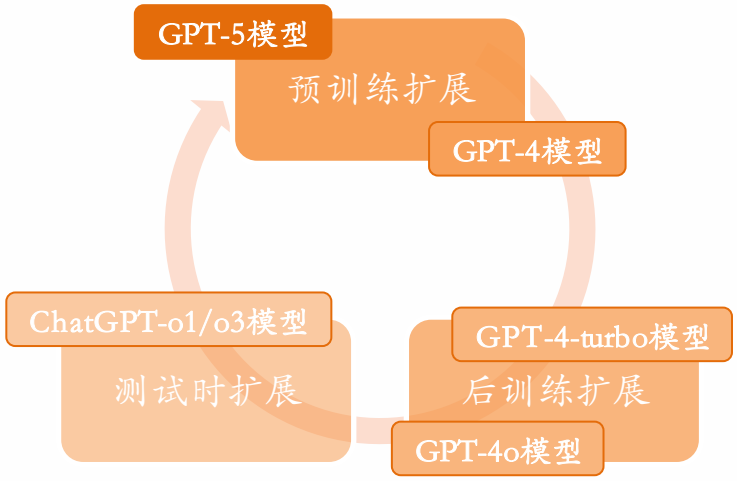
- **预训练奠定模型内部智能上限，后训练及测试时扩展释放智能潜力。**未来值得关注的大模型：
- **Meta Llama-4模型：**2025年4月5日，Meta推出Llama-4系列模型，其中，小模型Scout和中模型Maverick目前已对外发布，大模型Behemoth仍在训练中。根据Meta公布，大模型Llama-4-Behemoth总参数接近2万亿，将成为Meta最大预训练扩展模型，预计于未来数月内发布。
- **OpenAI GPT-5模型：**2025年2月28日，OpenAI推出GPT-4.5语言模型，并表示将于数月内发布GPT-5模型，GPT-5模型将成为下一代大模型，并有望整合o系列模型测试时扩展能力。

AI Agent产业链三大环节核心玩家



资料来源：Z Research，西南证券整理

大模型循环扩展提升智能水平



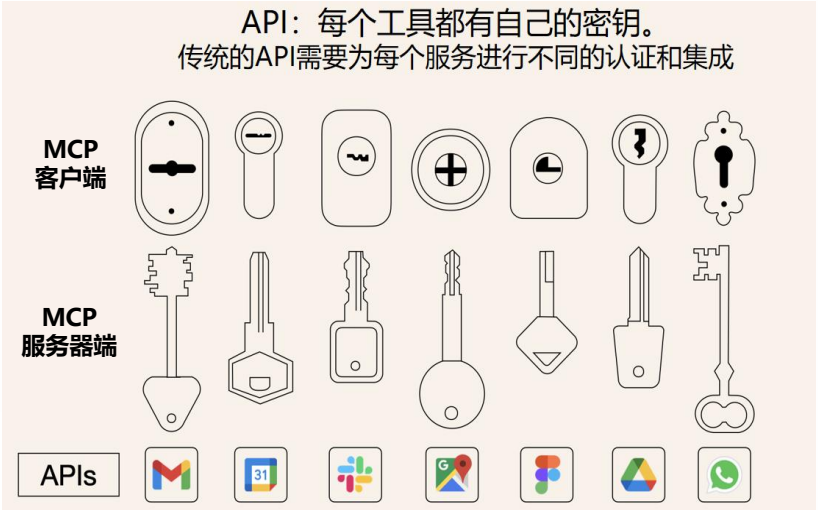
资料来源：西南证券

目 录

3.1 MCP：定义工具接口标准，打造新一代上下文通信协议

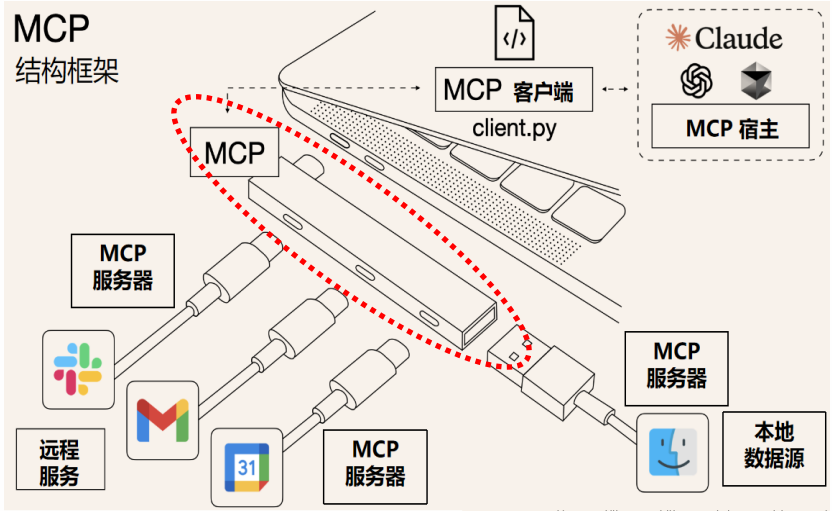
- **MCP：链接模型上下文信息与智能工具，建立上下文协议行业标准。**2024年11月，Anthropic发布Model Context Protocol（MCP），MCP是一种用于在AI系统中管理和交换模型上下文信息的协议，旨在不同的AI模块、系统或模型之间共享环境、状态和上下文数据。自推出以来，MCP迅速成为AI原生应用的重要基础设施，从结构框架层面来看，传统API与MCP之间存在显著差异：**1）传统API：**基于经典的**“客户端-服务端”架构**，客户端发起请求，服务器处理并返回响应，传统API充当二者之间的中介，开发者通常需要分别集成多个服务接口，单独处理认证、数据格式和通信协议，带来较高的集成与维护成本，易出现响应机制不一致等问题。**2）MCP：**遵循**“客户端-服务器”**架构，由**MCP主机/MCP客户端/MCP服务器**三个核心组件组成，专为AI系统设计，通过标准化协议传递模型所需的上下文数据，使模型能够高效调用工具，提升AI模型的理解与执行能力。

传统API技术路线示意图



资料来源：Z Research，西南证券整理

Anthropic MCP技术路线示意图

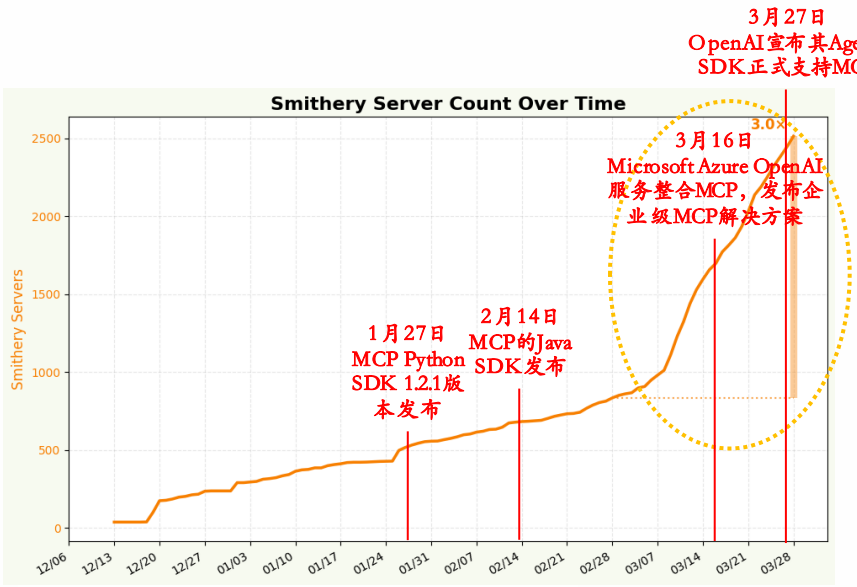


资料来源：Z Research，西南证券整理

3.1 MCP：定义工具接口标准，打造新一代上下文通信协议

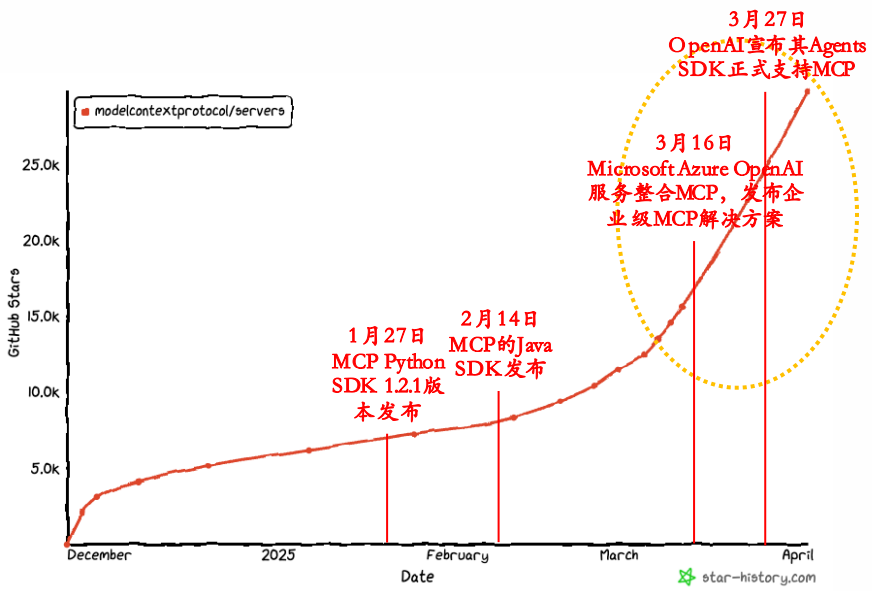
- **MCP Server供给快速增加，开发者生态加速繁荣。**根据Madrona信息，截至2025年3月28日，MCP Server发现平台Smithery的服务器创建数量较2月同期实现3倍增长；截至2025年4月初，MCP Server的GitHub star数已突破2.5万，曲线呈现加速上升趋势；同时，MCP TypeScript SDK也在快速增长，每周在npm上的下载量已接近70万次，**SDK下载量远超服务器包，表明更多的人选择自己搭建MCP服务器或者在自己的应用中支持MCP，开发者更倾向于为未来的使用场景做准备，而非为现有用户需求去部署MCP工具。**从原始数据来看，受益于新服务器和开发工具的不断涌现，MCP呈现爆炸式增长，供给侧生态逐步繁荣。MCP协议自发布以来，海内外众多企业陆续采用，应用场景不断拓宽，MCP向AI工具集成行业标准加速迈进。

Smithery平台上MCP服务器创建数量



资料来源：Madrona，西南证券整理

MCP Server的GitHub star数量

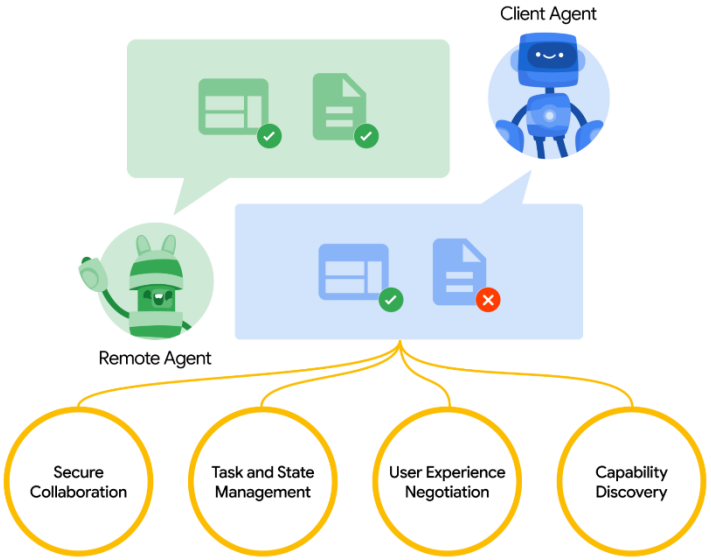


资料来源：Madrona，西南证券整理

3.2 A2A：开放智能体互联通信，优化新一代智能体网络协议

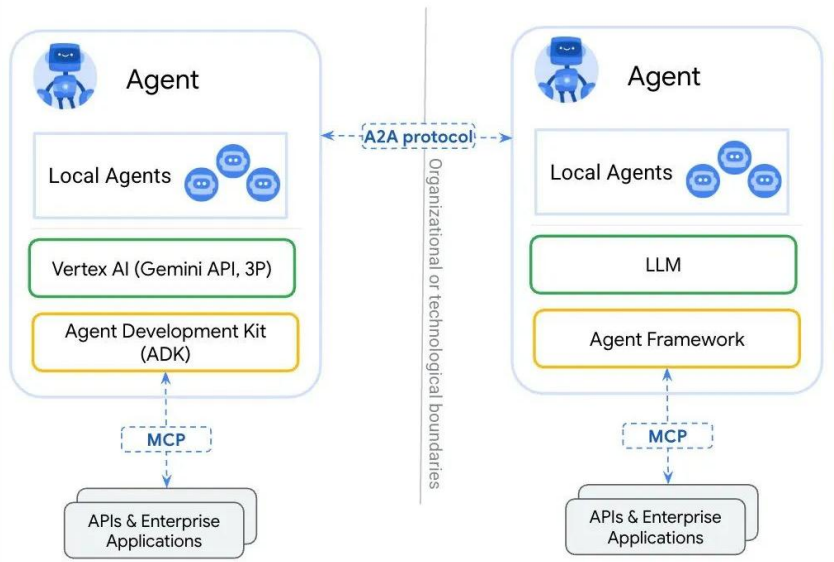
- **A2A：链接客户端与远程智能体通信，加速企业内部系统协同工作。**2025年4月9日，谷歌推出全新开放协议Agent2Agent(A2A)，允许AI代理跨生态系统协作。A2A协议旨在让不同来源、不同技术的AI智能体能够安全高效地交换信息，并协同执行跨企业平台或应用的复杂任务。**A2A协议是对Anthropic上下文协议MCP的补充，MCP为智能体提供链接工具与上下文的标准，A2A则侧重于智能体之间的交互与协作，帮助智能体之间的高效沟通。**工作原理方面，A2A用于促进客户端智能体和远程智能体之间的通信，其中，用户存在于协议中，主要的作用是用于认证和授权；客户端智能体负责制定和传达任务；远程智能体则负责执行任务或采取行动。

谷歌A2A开放协议工作原理



资料来源：谷歌官网，西南证券整理

谷歌A2A与MCP协同工作



资料来源：谷歌官网，西南证券整理

3.2 A2A：开放智能体互联通信，优化新一代智能体网络协议

- **A2A：合作伙伴阵营日益壮大，专业知识强化企业内部智能体。**根据谷歌官网展示的招聘软件工程师案例，**A2A协作能够大幅简化流程：**在统一界面Agentspace中，HR可以指派客户端智能体根据职位描述、地点和技能要求，寻找匹配候选人；客户端智能体随后会与其他远程智能体进行交互，以获取潜在候选人；HR收到推荐结果后，可以进一步指示客户端智能体安排面试，从而简化人才筛选流程；面试流程结束后，可以再启用其他远程智能体进行背调等行动，从而实现智能体的跨系统合作，帮助寻找合适的候选人。根据谷歌披露信息，A2A的发布已得到包括Atlassian、Box、Cohere、Intuit、Langchain、埃森哲、BCG、Capgemini、Cognizant等在内的**50多家技术合作伙伴和服务提供商的支持**，生态系统日益壮多样，合作伙伴的专业知识对塑造智能体的相互协作创造更大，企业内部智能体有望跨系统工作，释放工作效率和创新潜力。

谷歌A2A招聘应用示例

资料来源：谷歌官网，西南证券整理

谷歌A2A合作伙伴生态

- 技术&平台合作伙伴：提供构建和运行A2A智能体系统的技术与平台。
- 服务合作伙伴：将技术应用到具体的业务场景中。

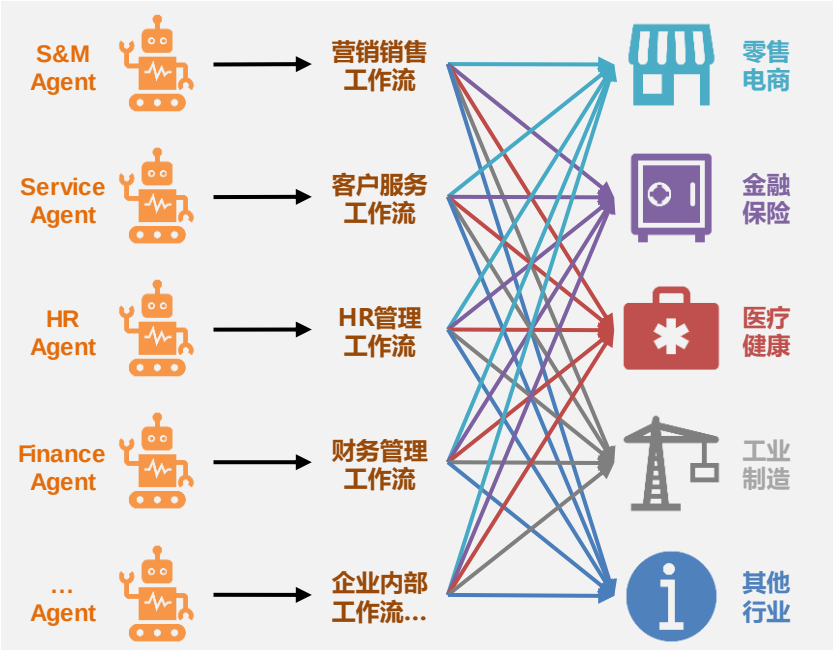
资料来源：谷歌官网，西南证券整理

目 录

4.1 Agent类别：把握“通用”与“垂类”两大应用方向

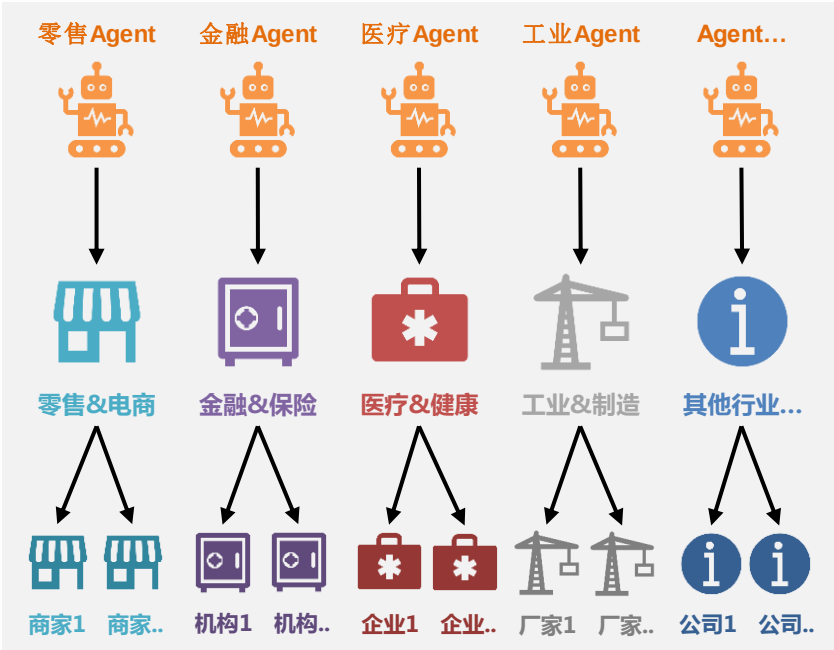
- **同步打磨通用与专业能力，交叉渗透横向与垂直市场。1）跨行业AI代理(Horizontal AI Agent)：**随着AI技术的发展、以及基于过去通用SaaS产品的历史经验，当前已有众多初创企业针对多种行业或领域打造AI通用智能体，如Sierra（客服代理）、Cursor（软件开发智能体），可提供跨行业智能服务。**2）垂类AI代理(Vertical AI Agent)：**针对某个具体行业或垂直市场提供专业智能体，如AI金融研究创企Boosted.ai、工业控制创企Composabl等，为特定客户类别制定个性化解决方案。当前，以上两类智能体正加速发展，有望向各市场横纵渗透。

横向跨行业通用AI代理 (Horizontal AI Agent)



资料来源：西南证券

垂类行业专业AI代理 (Vertical AI Agent)

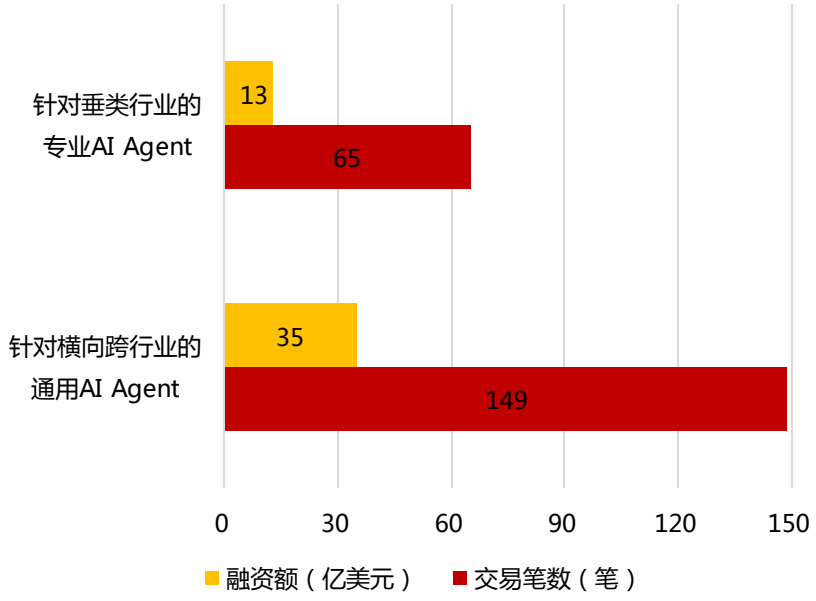


资料来源：西南证券

4.1 Agent类别：把握“通用”与“垂类”两大应用方向

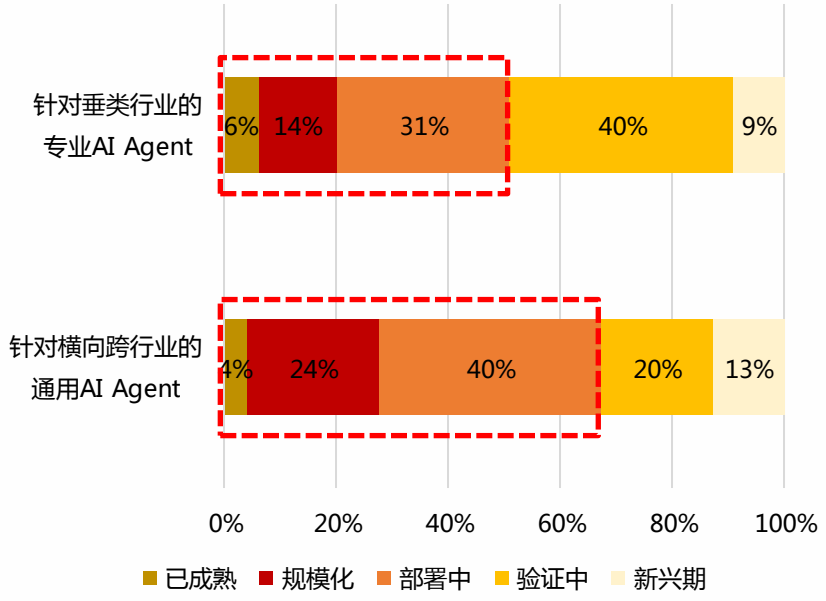
- **跨行业通用智能体发展相对成熟，垂类行业智能体逐渐兴起。**近年来，跨行业通用智能云引领创企风投活动，根据CBInsights统计数据，2020年至2025年2月，该领域的AI Agent创企融资额及交易数均远超垂类行业领域，融资额实现35亿美元，交易达成149笔，垂类行业的创企融资额仅为13亿美元，达成交易65起。在商业成熟度方面，跨行业通用应用在商业上更为成熟，超过2/3的市场正在部署或扩展AI解决方案，其中，客户支持、软件开发、销售和通用企业工作流程等赛道较为活跃；而垂类智能体仍处于新兴和验证阶段，预计未来垂直行业智能体将向部署阶段推进。

2020-2025年2月AI Agent创企融资额及交易数



资料来源：CBInsights，西南证券整理

截至2025年2月通用和专业AI Agent成熟度份额



资料来源：CBInsights，西南证券整理

4.2 Agent赋能：把握“降低成本+提高效率+增强体验”三项赋能

- **从AI智能体赋能作用来看，主要集中于“降低成本+提高效率+增强体验”三大方向。**其中，“降低成本”可以帮助企业提升**盈利能力**，“提高效率”有望加强企业**经营竞争力**，“增强体验”则将提升用户留存、扩大**潜在市场规模**，从而构建Agent的商业价值闭环。
- **赋能垂直业务：深入行业场景，扮演专业助手。**垂类行业智能体主要用于处理专业复杂任务，垂类行业涵盖电商、金融、医疗、工业等领域，扮演行业专家角色。在降低成本方面，Agent可以替代部分专业岗位中的重复性或规则明确的劳动，例如保单文件初审、医学影像识别等，节省人力开销；在提高效率方面，能够帮助流程节点的自动执行、加速专业信息流转；在增强体验方面，聚焦提升客户互动体验，例如问诊智能引导、客户服务个性化推荐。相较于跨行业通用智能体，垂类行业智能体的部署门槛相对更高，但赋能程度更深，是企业实现数智化转型的核心抓手。

Vertical Agent赋能垂直业务

垂类行业	降低成本	提高效率	增强体验
零售&电商	自动化产品目录，合成虚拟人节省模特费用	基于LLM搜索提高转化率	更智能、更相关的搜索，个性化头像
金融&保险	合成训练数据提高金融模型准确性并确保合规，在非结构化索赔文件中识别模式以最小化损失	AI助手能够大规模分析和合成财务数据，实现自动化承保决策	AI聊天机器人简化日常财务任务，保险销售过程中进行个性化互动
医疗&健康	AI药物发现与设计缩短上市时间，自然语言处理支持临床决策	自动化繁琐任务，改善电子健康记录文档，去噪放射扫描	AI伴侣关注健康和心理健康，合成患者数据保护患者隐私
工业&制造	自动化质检、预测性维护	智能调度系统提高生产效率，AI监控质量标准提升良品率	生产流程优化，实时故障预测与提醒

资料来源：CBInsights，西南证券整理

4.2 Agent赋能：把握“降低成本+提高效率+增强体验”三项赋能

- **赋能企业工作流：提升员工生产力，办公提效是核心。**跨行业智能体主要面向企业内部工作流场景，要求普遍适用于各类岗位和部门，具备覆盖面广、上手快、易部署等特点，本质在于打造办公提效工具、甚至形成AI办公操作系统。在降低成本方面，Agent通过自动化通用事务，减少基础人力资源投入；在提高效率方面，能压缩日常事务所耗费的时间；在增强体验上，为员工提供更高效、更流畅的协作环境，使工作节奏更轻盈，工作体验更智能。

Horizontal Agent赋能企业工作流

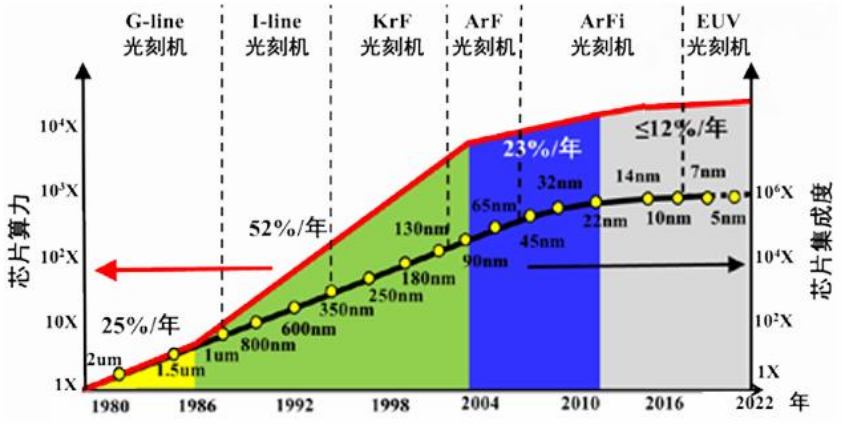
企业内部工作流	降低成本	提高效率	增强体验
营销&销售	自动化销售流程，降低销售人员成本	实时客户洞察，提高转化率	提供个性化推荐与互动，提升客户满意度
客户服务	减少人工客服数量和培训成本	快速响应，多轮对话处理能力强	24/7在线，支持多语言，提升响应体验
人力资源管理	节省简历筛选和入职流程的人力资源	自动化招聘、自动化绩效管理流程	提供个性化职业发展建议，提升员工满意度
财务管理	自动报账、对账，减少财务人员投入	实时报表生成，预测与预算更高效	更直观的财务仪表板与智能分析，便于决策
软件开发	自动生成代码、测试用例，降低开发与测试成本	智能体辅助调试、代码审查，加快交付周期	开发者获得实时建议和代码提示，优化开发体验
网络安全	自动监控与响应降低安全团队人力成本	实时检测威胁并自动响应，降低入侵风险	安全团队可通过智能助手快速定位并理解威胁来源

资料来源：CBInsights，西南证券整理

4.3 Agent摩尔定律：处理任务长度每7月翻一倍，性能增长且成本下降

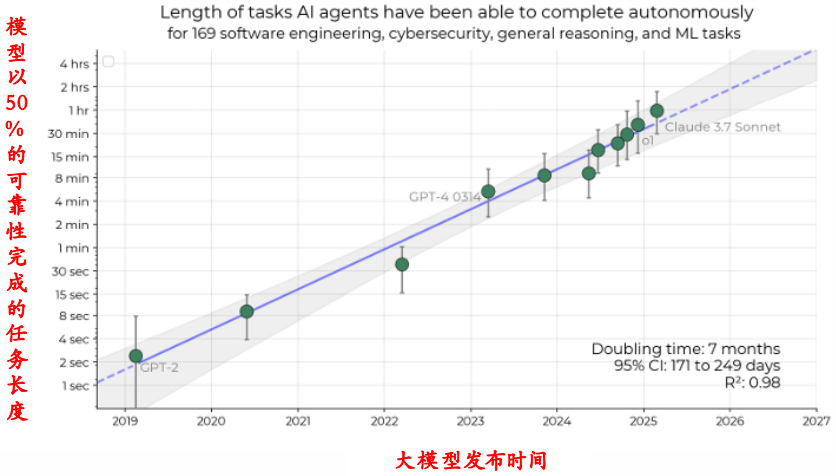
- **芯片领域的摩尔定律：每18-24个月，芯片中的晶体管数量翻倍，芯片算力性能增长、成本下降。**
- **Agent的“摩尔定律”：每7个月，AI能够处理的任务长度翻倍，模型性能提升、成本下降。**根据机构METR的研究《Measuring AI Ability to Complete Long Tasks》，2019年至2025年，AI在50%的可靠性标准下，完成的任务长度（以人类专业人士完成任务所需时间衡量）大约每7个月翻一番，若以该“摩尔定律”线性外推，到2029年AI或许能处理需1个月的复杂任务。2025年4月17日，OpenAI发布o3 & o4-mini模型，且初步具备主动调用外部工具的能力，在其官网演示的科研海报应用案例中，从耗时角度，o3模型可以在20s内完成人类研究员可能需数天完成的任务。

芯片中的晶体管数每18-24个月翻一倍



资料来源：《集成芯片与芯粒技术白皮书》，西南证券整理

AI能够处理的任务长度每7个月翻一倍

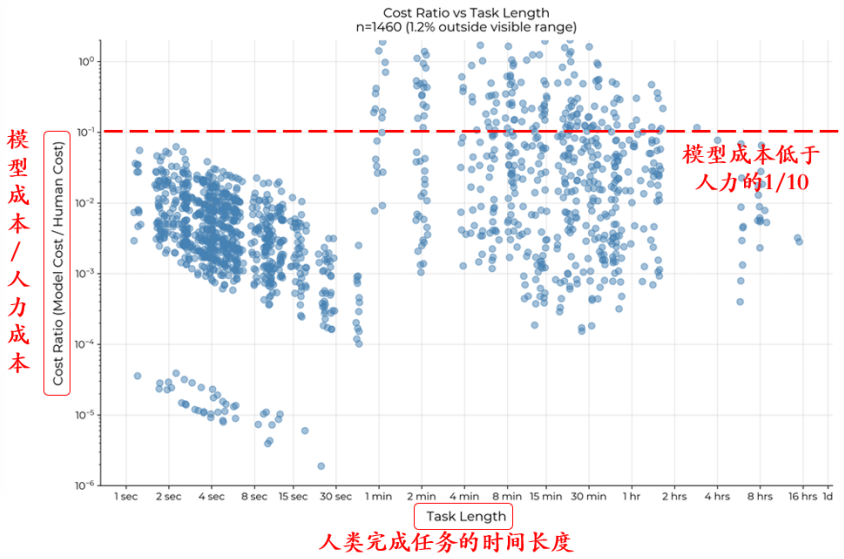


资料来源：《Measuring AI Ability to Complete Long Tasks》，西南证券整理

4.3 Agent摩尔定律：处理任务长度每7月翻一倍，性能增长且成本下降

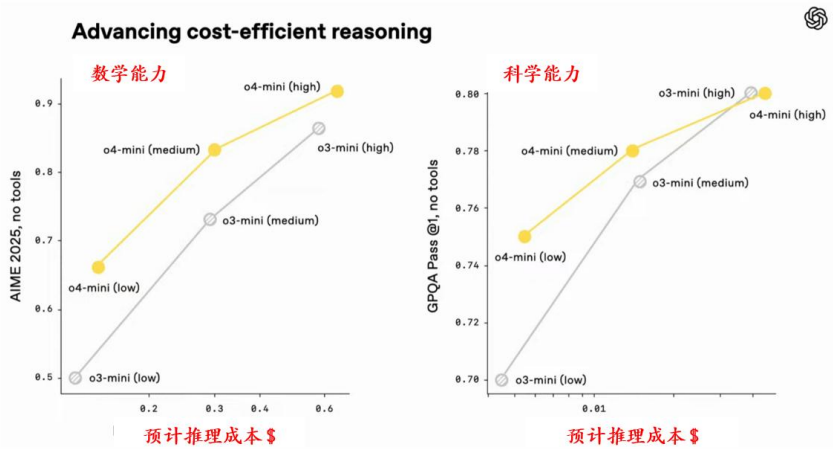
- **智能体处理任务长度正大幅提升，完成简单任务已具性价比。**根据机构METR的研究《Measuring AI Ability to Complete Long Tasks》，**目前超过80%由智能体成功完成的任务中，AI的推理成本仅为人类专家的10%**，其中，对于人类专家在30秒内完成的任务，使用AI的经济优势显著，目前已可完成16个小时的软件任务，但在完成长时序现实世界任务方面，人类在整个工作循环中仍然需要发挥较大作用和价值。在2025年4月17日OpenAI发布的o3和o4-mini模型中，o3和o4-mini在很多情况下比各自的前代o1与o3-mini更高效、更节省成本，在AME2025基准测试中，性价比远超过前代模型。未来随着AI技术持续发展，AI在复杂任务面前有望同样具备性价比，从而推动AI智能体进一步应用与渗透。

AI完成简单任务已具备经济优势



资料来源：《Measuring AI Ability to Complete Long Tasks》，西南证券整理

OpenAI-o系列模型性价比持续提升



资料来源：OpenAI官网，西南证券整理

4.4 Agent初代产品：产品ARR迅速增长，爆发潜力可期

- **智能体打造交互式应用，未来爆发潜力可期。**智能体作为可交互的应用产品，能够快速触达用户，付费渗透空间较大。根据Sacra数据，Agent初代产品Cursor已成为年经常性收入(ARR)从0增长至1亿美元最快的初创企业之一，耗时约12个月、于2024年底达成1亿美元ARR里程碑，并于2025年3月ARR迅速达到2亿美金。根据当前AI智能体创企公司及产品来看，布局领域主要集中于编码、法律、招聘、客服、医疗等行业领域，商业模式持续探索，**已出现基于结果（AI交付实际成果、任务完成率等指标）定价的AI应用产品，**或根据资源消耗量收费，也可采用常见的SaaS产品订阅方式对商业模式进行补充。基于智能体产品爆发快、轻资产、强用户触达等特点，未来有望凭借更快的成长速度成为AI应用黑马，建议后续关注ARR实现快速提升的AI公司及产品。

AI创企ARR增速及2024年估值情况

AI初创企业	产品定位	ARR增长速度	2024年ARR ($亿)	2024年估值	2024年估值倍数
Cursor	AI**编程**助手	**12个月内**从0做到**1亿**美元	1	25	25
Lovable	非程序员的AI**软件开发**平台	**2个月内**从0做到**1000万**美元	0.07	/	/
Glean	**企业级AI搜索**平台	**21个月内**从0做到**1亿**美元	1.1	46	42
Codeium	AI**编程**助手	/	0.12	12.5	104
Harvey	AI**法律**助手	**26个月内**从0做到**5000万**美元	0.5	30	60
Hebbia	AI驱动的**金融/法律**助手	/	0.13	7	54
Bolt.new	AI驱动的**网页构建**平台	**2个月内**从0做到**2000万**美元	0.25	/	/
Mercor	AI驱动的**招聘**平台	**2年内**从0做到**5000万**美元	0.5	2.5	5
Decagon	AI驱动的**客服**代理	/	0.06	6.5	108
Sierra	AI驱动的**客服**软件	/	0.2	45	225
Commure	AI驱动的**医疗**软件	/	0.4	12	30

资料来源：Sacra，Tullopy，西南证券整理

4.5 Agent流量入口：AI入口尚处于早期阶段，或将呈现中心化特点

- **互联网流量入口从中心化转向分散化，AI时代入口可能处于早期收敛阶段。1）互联网1.0时期：**门户网站Yahoo、Sina成为信息获取的主要入口，流量入口高度中心化；**2）互联网2.0时代：**搜索引擎如百度、谷歌成为主流入口；**3）互联网3.0阶段：**随着移动互联网和智能终端的就是普及，以及各类垂类APP的崛起，逐渐打破过去流量入口的垄断格局，呈现明显多元化趋势，入口价值逐渐向场景和服务转移。类比互联网时代来看，当前AI时代的入口尚处于早期阶段，ChatGPT、DeepSeek等少数大模型产品主导用户心智，呈现出类似于互联网时代早期的门户式中心化状态。

互联网流量入口由中心化转向多元化

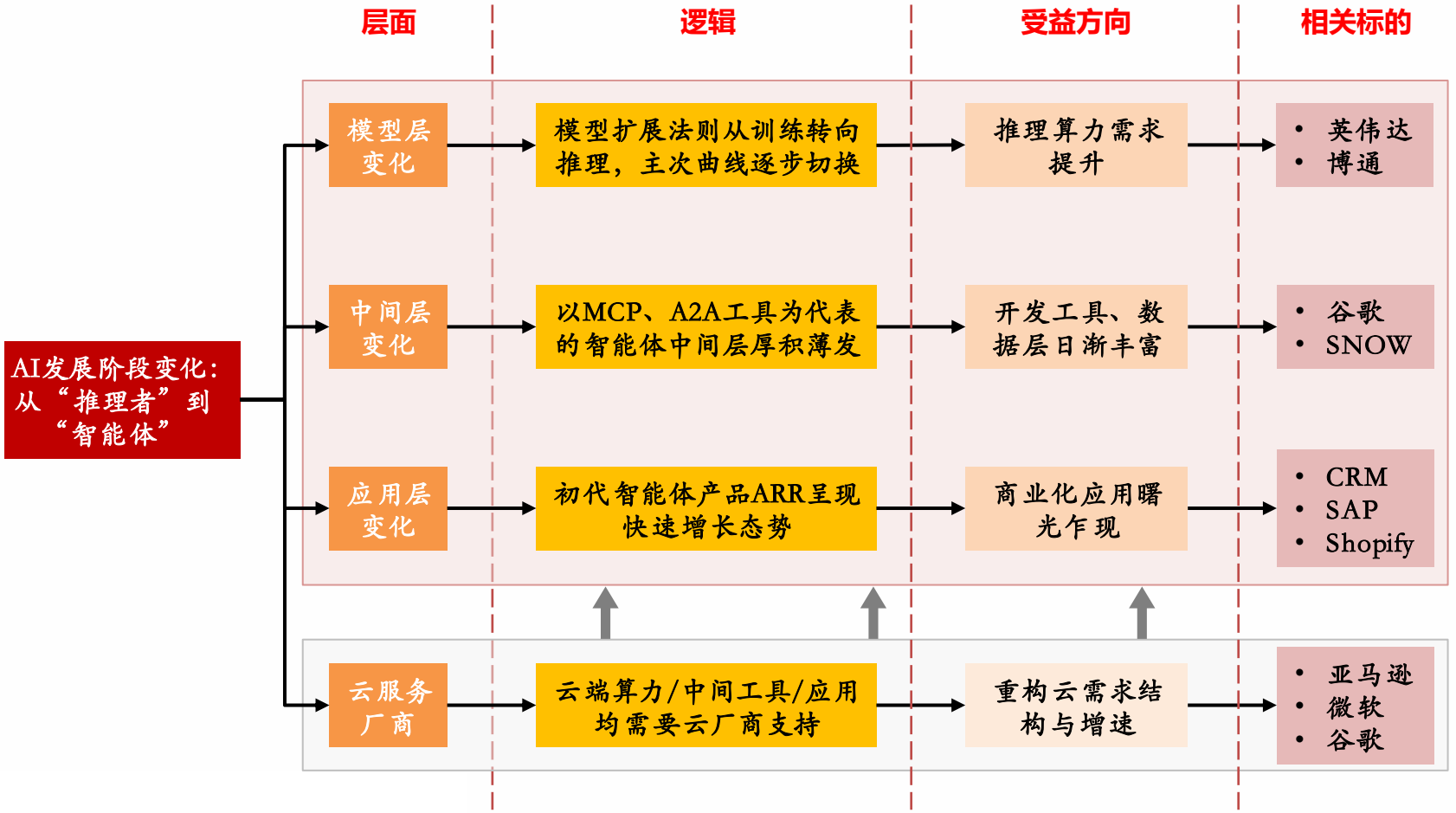
1990s → 2000s → 2014 → 2022

互联网时代	互联网 1.0	互联网 2.0	互联网3.0
➢ 产品入口	YAHOO! sina	Baidu 百度 Google	微信 淘宝 京东 抖音 Facebook Amazon Netflix YouTube
技术逻辑	基于HTML静态网页，用户仅能通过浏览器被动获取信息，内容由专业开发者单向灌输	1）动态网页和数据库驱动，用户既是消费者也是生产者；2）PC在家庭和个人用户中广泛普及	1）信息找人；2）智能手机渗透率攀升；3）区块链实现数据确权与跨链互通
商业模式	1）初代流量逻辑，点击率吸引广告；2）用户作为数据源被商品化，不参与价值分配	1）搜索引擎竞价排名；2）PC在家庭和个人用户中广泛普及	1）流量经济；2）用户行为数据成为核心生产资料，通过精准推荐和增值服务变现
权力结构	门户网站形成入口，入口中心化特点明显	搜索引擎形成入口，入口中心化特点明显	各垂类APP形成场景化生态（微信社交/淘宝电商/美团本地生活等），入口价值出现分散

资料来源：Z Research，西南证券整理

目 录

相关标的


层面
逻辑
受益方向
相关标的
AI发展阶段变化：从“推理者”到“智能体”
模型层变化
模型扩展法则从训练转向推理，主次曲线逐步切换
推理算力需求提升
• 英伟达
• 博通
中间层变化
以MCP、A2A工具为代表的智能体中间层厚积薄发
开发工具、数据层日渐丰富
• 谷歌
• SNOW
应用层变化
初代智能体产品ARR呈现快速增长态势
商业化应用曙光乍现
• CRM
• SAP
• Shopify
云服务厂商
云端算力/中间工具/应用均需要云厂商支持
重构云需求结构与增速
• 亚马逊
• 微软
• 谷歌

风险提示

- AI技术进展不及预期；
- AI商业化进展不及预期；
- 投资回报不及预期等风险。

西南证券
SOUTHWEST SECURITIES

分析师：王湘杰
执业证号： S1250521120002
电话：0755-26671517
邮箱：wxj@swsc.com.cn

联系人：尤品柯
邮箱：ypk@swsc.com.cn

西南证券 SOUTHWEST SECURITIES

西南证券研究院

西南证券投资评级说明

报告中投资建议所涉及的评级分为公司评级和行业评级（另有说明的除外）。评级标准为报告发布日后6个月内的相对市场表现，即：以报告发布日后6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。其中：A股市场以沪深300指数为基准，新三板市场以三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的）为基准；香港市场以恒生指数为基准；美国市场以纳斯达克综合指数或标普500指数为基准。

公司评级	买入：未来6个月内，个股相对同期相关证券市场代表性指数涨幅在20%以上
	持有：未来6个月内，个股相对同期相关证券市场代表性指数涨幅介于10%与20%之间
	中性：未来6个月内，个股相对同期相关证券市场代表性指数涨幅介于-10%与10%之间
	回避：未来6个月内，个股相对同期相关证券市场代表性指数涨幅介于-20%与-10%之间
	卖出：未来6个月内，个股相对同期相关证券市场代表性指数涨幅在-20%以下
行业评级	强于大市：未来6个月内，行业整体回报高于同期相关证券市场代表性指数5%以上
	跟随大市：未来6个月内，行业整体回报介于同期相关证券市场代表性指数-5%与5%之间
	弱于大市：未来6个月内，行业整体回报低于同期相关证券市场代表性指数-5%以下

分析师承诺

报告署名分析师具有中国证券业协会授予的证券投资咨询执业资格并注册为证券分析师，报告所采用的数据均来自合法合规渠道，分析逻辑基于分析师的职业理解，通过合理判断得出结论，独立、客观地出具本报告。分析师承诺不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接获取任何形式的补偿。

重要声明

西南证券 SOUTHWEST SECURITIES

西南证券研究院

西南证券研究院

上海

地址：上海市浦东新区陆家嘴21世纪大厦10楼

邮编：200120

北京

地址：北京市西城区金融大街35号国际企业大厦A座8楼

邮编：100033

深圳

地址：深圳市福田区益田路6001号太平金融大厦22楼

邮编：518038

重庆

地址：重庆市江北区金沙门路32号西南证券总部大楼21楼

邮编：400025

西南证券机构销售团队

区域	姓名	职务	手机	邮箱	姓名	职务	手机	邮箱
上海	蒋诗烽	总经理助理/销售总监	18621310081	jsf@swsc.com.cn	张玉梅	销售经理	18957157330	zymyf@swsc.com.cn
	崔露文	销售副总监	15642960315	clw@swsc.com.cn	欧若诗	销售经理	18223769969	ors@swsc.com.cn
	李煜	资深销售经理	18801732511	yfliyu@swsc.com.cn	李嘉隆	销售经理	15800507223	ljlong@swsc.com.cn
	田婧雯	高级销售经理	18817337408	tjw@swsc.com.cn	龚怡芸	销售经理	13524211935	gongyy@swsc.com.cn
	汪艺	高级销售经理	13127920536	wyyf@swsc.com.cn	蒋宇洁	销售经理	15905851569	jyj@swsc.com.c
北京	李杨	销售总监	18601139362	yfly@swsc.com.cn	张鑫	高级销售经理	15981953220	zhxin@swsc.com.cn
	张岚	销售副总监	18601241803	zhanglan@swsc.com.cn	王一菲	高级销售经理	18040060359	wyf@swsc.com.cn
	杨薇	资深销售经理	15652285702	yangwei@swsc.com.cn	王宇飞	高级销售经理	18500981866	wangyuf@swsc.com
	姚航	资深销售经理	15652026677	yhang@swsc.com.cn	马冰竹	销售经理	13126590325	mbz@swsc.com.cn
广深	郑龑	广深销售负责人	18825189744	zhengyan@swsc.com.cn	陈韵然	销售经理	18208801355	cyryf@swsc.com.cn
	龚之涵	高级销售经理	15808001926	gongzh@swsc.com.cn	林哲睿	销售经理	15602268757	lzr@swsc.com.cn
	杨举	销售经理	13668255142	yangju@swsc.com.cn				