

洞见AI新机遇

中国AI行业 系列观察报告

» 穿越资讯迷雾，重塑AI认知 «

本月主题——全面拥抱AI生产力时代

出品机构：甲子光年智库

智库院长：宋涛

项目团队：刘瑶、努尔麦麦提·买合木提

发布时间：2025.6

并购家 免费行业报告网

IPOIPO.CN 每日更新 不用注册会员 无广告 永久免费

背景介绍

AI一天，人间一年。

人工智能（AI）领域，正以前所未有的速度向前发展。几乎每一天，都有新的突破、新的应用和大量的科技资讯涌现。

这种速度令人振奋，但也容易让人感到应接不暇，甚至产生焦虑——无论是关注行业动态的人，还是身处其中的研究者，都难以完全消化和理解这海量的信息。

与此同时，各种充满诱惑力的营销话语也扑面而来，声称AI将“颠覆一切”、“解决所有问题”。许多技术被冠以“革命性”“划时代”的标签，过度夸张的宣传与不切实际的预测比比皆是。这些声音有时模糊了技术的真实边界，也加剧了人们的不确定感。

正是基于这样的认识，我们希望做出一份系列报告，持续提供具有一定系统性和梳理性的文字，形成的一份阶段性观察与思考。

我们深知自己力量有限，但依然希望尽一份绵薄之力。这份《中国AI行业系列观察报告》，就是我们尝试给出的答案。它不是一份面面俱到的百科全书，也不敢保证绝对的客观。它仅仅是我们团队，基于对海量信息的梳理、分析和讨论，尝试理解技术发展的脉络、落地的挑战与机遇，提供一个相对冷静、务实的第三方视角。这份努力，是我们能献上的最大诚意。

我们非常清楚，AI的图景宏大而复杂，任何一份报告都难以完全涵盖。因此，我们更愿意将这份报告视为一个起点，一次邀请。如果您在阅读中有所启发，或者持有不同的见解，甚至更好的建议，我们真诚地欢迎您直接联系作者（联系方式见报告末页）。

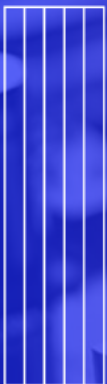
您的反馈，将帮助我们做得更好。

前路充满未知，变化是唯一的不变。希望我们这份虽然不完美但力求真诚的观察，能在快速变化的AI浪潮中，提供一点有价值的参考和思考的支点。AI的未来，需要所有参与者的共同探索与书写。

第一章 行业动态

第二章 大咖观点

第三章 产品解读



第一章

行业动态： 追踪发展脉络，剖析行业动向

行业动态

从全球标志企业看行业变化

AI智能体化加速推进，
改变软件生态

AI商业化路径面临调整，
创新才是关键

专有模型与开源模型共存，
双轨战略日益明显

谷歌AI发力，以模型为核心带动多款产品迭代

谷歌I/O 2025：Gemini进化、AI搜索革新与多款新品亮相

谷歌在2025年的I/O开发者大会上，全面展示了其以Gemini模型为核心的AI战略，并在搜索、医疗、教育、通信和内容创作等多个领域推出了创新产品与服务。

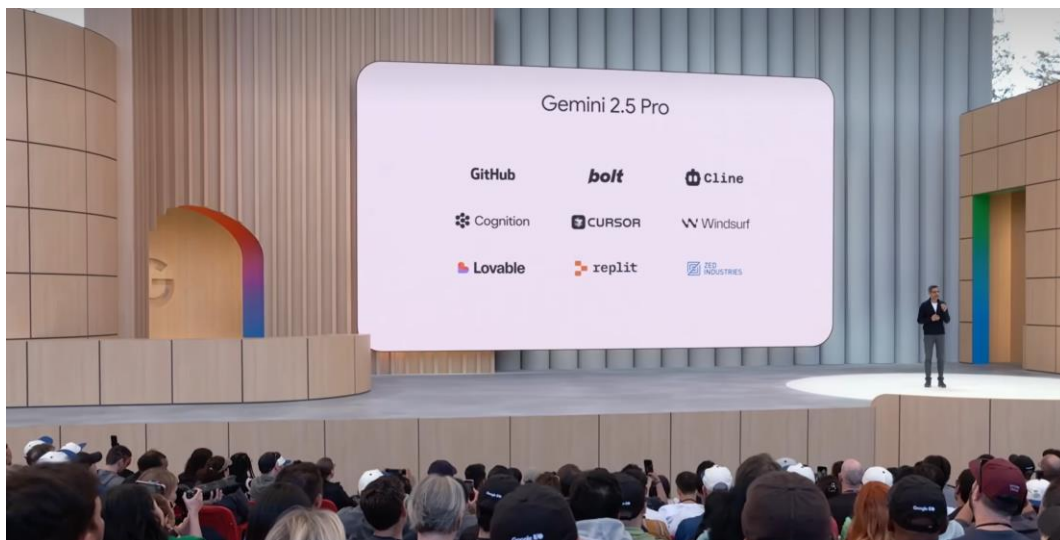
Gemini 2.5系列模型重磅升级：谷歌宣布对其旗舰AI模型Gemini 2.5系列（包括Pro和Flash版本）进行重大更新。新版本在推理能力上得到显著增强，特别引入了实验性的“Deep Think”模式，专为处理复杂的数学和编码任务而设计。此外，Gemini 2.5还具备了原生音频输出能力，以实现更自然的交互体验，并强化了安全特性以及对Project Mariner计算机使用能力的支持。

AI驱动搜索体验革新：谷歌进一步扩展了其搜索引擎中的“AI模式”，并向所有年满18周岁的美国用户开放。该模式的核心是“AI概览”（AI overviews），即在搜索结果顶部提供对话式的摘要信息，目前已开始在Search Labs环境之外进行测试。

智能体生态系统建设取得突破性进展

以A2A协议构建开放标准联盟，构建完整的Agent开发与应用体系。同时推出的Agent Engine UI提供了全生命周期管理工具，帮助开发者更好地管理和优化智能体应用。

行业应用落地成效显著：在金融合规场景，德勤全球智能体网络通过A2A协议实现了KYC流程的端到端自动化，审计效率提升40%。在制造业物联场景，西门子工业智能体集群借助优化的文件交换机制，将设备诊断报告生成时间从15分钟缩短至107秒，这一突破显著提升了产线运维效率。在医疗协作场景，梅奥诊所部署的影像分析智能体与电子病历智能体通过智能体卡自动发现机制，将误诊率降低了29%，同时将诊断时间缩短了40%。



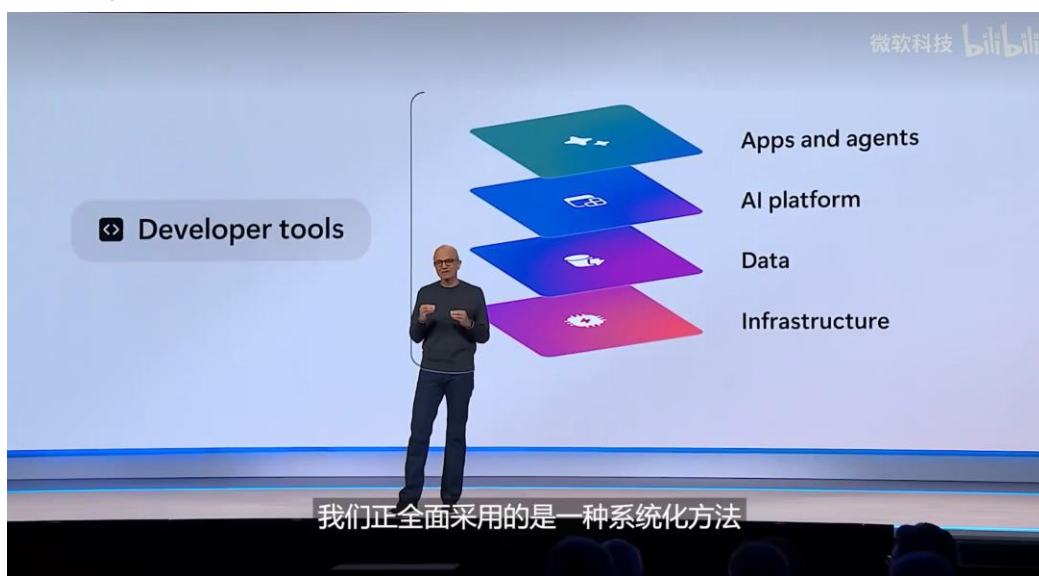
微软Build 2025：Agent为代表描述宏大愿景

微软Build 2025：Agentic AI愿景、Azure AI升级与Power Platform融合

微软在其Build 2025开发者大会上，重点阐述了其构建“开放式智能体网络”（open agentic web）的宏大愿景，旨在将操作系统转变为一个AI智能体平台，并开发能够执行复杂任务的自主系统。

AI 智能体战略全面铺开：

- ❑ **GitHub Copilot**：从编码助手进化为可独立处理 GitHub 问题、起草拉取请求并迭代的自主团队成员，已向 Copilot Enterprise 和 Copilot Pro + 客户提供。
- ❑ **ows AI Foundry 与模型上下文协议 (MCP)**：ows 11 原生支持 MCP，促进 AI 智能体与本地应用及系统服务通信；ows AI Foundry 支持开发者在 Copilot+ PC 上对开源或自定义 AI 模型进行设备端微调 and 部署（支持 CPU、GPU、NPU），提升速度与隐私保护。
- ❑ **Microsoft 365 Copilot Tuning 与 Copilot Studio**：集成于 Copilot Studio 的低代码工具，企业无需编码即可构建特定领域 AI 智能体（如用“配方”进行专家问答和文档生成，用企业专有数据训练），现支持多智能体编排。
- ❑ **Azure AI Foundry 更新**：扩展模型目录，新增 xAI 的 Grok 3、Black Forest Labs 的 Flux Pro 1.1 及 Hugging Face 超 10,000 个开源模型；推出 Foundry Agent Service，用于编排多智能体工作流，配备模板、操作及实时模型路由器。
- ❑ **Microsoft Discovery**：全新智能体平台，基于图形知识引擎，通过自动化构思、实验和协作改变研究流程。
- ❑ **Microsoft Copilot Studio/Copilot** 新增大量 Agent 实用功能：包括集成 Microsoft 365 Agents SDK 和 Azure AI Foundry 以简化企业级智能体开发并支持项目脚手架、测试和发布且已普遍可用的 Visual Studio 工具包，支持构建可扩展多渠道智能体并与 Copilot Studio 和 Visual Studio 集成且已全面上市的 Microsoft 365 Agents SDK 等
- ❑ **GitHub Copilot 新智能体开发功能**：GitHub Copilot 现在可以将想法在数小时内变成软件原型，并在数天内投入生产应用。
- ❑ **全新 SRE 智能体**：持续监控 Azure 上生产应用程序的健康状况和性能，提供故障排除见解。与 GitHub Copilot 中的 SWE 功能合作，主动识别问题并分配给 GitHub 中的智能体，推动问题解决。



OpenAI: AI多模态与智能体构建加速

o3和o4-mini模型发布: 4月16日, OpenAI推出o3和o4-mini两款推理模型。o3全面继承gpt-4级多学科水平, 首度把“视觉链式思考”与全套chatgpt工具整合到同一个模型中, 成为OpenAI当前最强“推理旗舰”;

gpt-4.1系列上线: OpenAI发布gpt-4.1系列, 支持百万token上下文窗口, 编程能力(swe-bench得分54.6%)显著提升, 价格低于gpt-4o。

视觉链式思考: o3/o4-mini模型可在隐式思维链内部对上传图片做局部放大、旋转、标注并将其中的结构化信息写入后续推理, 从“识别→回答”升级为“观察→思考→行动”。

OpenAI Agent相关动作:

推出typescript原生支持: OpenAI的agents sdk正式跳出python圈, 推出typescript版本sdk, 支持handoffs、guardrails、tracing、mcp以及agent所需的各类基础能力。

新增human-in-the-loop功能: OpenAI增强了human-in-the-loop功能, 允许开发者在agent调用工具或执行任务时, 主动暂停、保存当前agent状态, 让人工审核、确认或调整调用, 再决定是否恢复执行。

实时语音agent功能上线: 支持和文字agent一样的工具调用、任务移交、guardrails, 自动处理语音输入、输出、用户打断。

发布browsecomp基准测试: OpenAI正式开源浏览器智能体评测基准browsecomp。

OpenAI正式发布Codex。 这是一款基于云端运行的软件工程智能 Agent, 能够并行处理多个开发任务, 协助开发者高效完成编程工作。Codex提供一个输入框和两个按钮——Ask (提问) 和 Code (编码), 除了编写代码, 还包括理解代码结构、回答关于代码库的问题、修复 bug, OpenAI已在内部使用Codex处理代码重构、测试生成和错误修复等任务。



以大语言模型为核心积极布局 AI 智能体，聚焦自主性、生态构建及企业级应用

2025 年 6 月观察显示，国际科技巨头积极布局 AI 智能体领域，将其视为战略重点，以下是主要巨头布局概况及共同趋势。

谷歌：核心是构建“世界模型 AI”，企业级平台 Agentspace 支持开发部署智能体。产品包括 AI Agent 框架、Project Astra 原型智能体、Agent2Agent 协议等，还有 Deep Research 等具体应用。

微软：愿景是“开放式智能体网络”，产品有进化为智能体的 GitHub Copilot、AWS AI Foundry、Azure AI Foundry Agent Service 等，还支持 NLWeb 和 MCP 协议。

OpenAI：视 AI 智能体为通往 AGI 的关键路径，推出 Agents SDK 和 Responses API 等，CEO 预测 2025 年可能出现“L3 级别 AI 智能体”。

Salesforce：通过 Einstein Copilot 平台整合 AI 智能体到 CRM 各环节，产品有 Einstein Copilot，具备多种核心能力，提供可定制智能体。

IBM：强调“以人为本”，watsonx Orchestrate 平台旨在规模化自动化 workflows，关注高风险行业应用。

Anthropic：Claude 模型是构建智能体的优选基础模型，具备“工具使用”功能，还有 Claude Code 智能体，为企业伙伴提供技术支持。

亚马逊：通过“Nova Act”计划进入该领域，Nova Act SDK 允许开发者创建在浏览器执行任务的智能体。

科技巨头的Agent布局

Microsoft

战略定位

构建“AI Agent Factory”生态系统
推出开放协议 MCP，促进互操作性

核心产品

GitHub AI Coding Agent
Microsoft 365 Copilot
Azure AI Foundry
Microsoft Discovery

创新亮点

Copilot Studio 支持多 Agent 编排和企业级定制

Google

战略定位

推动跨 Agent 通信标准化
简化 Agent 开发流程

核心产品

Agent2Agent (A2A) 协议
Agent Development Kit

创新亮点

A2A 协议支持多模态交互和安全通信

OpenAI

战略定位

聚焦企业级应用，推动 ChatGPT 向“企业大脑”转型
加强推理模型研发，支持复杂任务自动化

核心产品

ChatGPT 企业版 GPT-4.1
API 服务

创新亮点

计划推出支持多任务执行的自主 Agent 产品

Anthropic

战略定位

解决 AI 模型与数据源的统一接入问题
推动 MCP 成为行业标准，促进生态互联

核心产品

MCP 协议
Claude 4 系列模型

创新亮点

Claude 3.7 Sonnet 支持“标准模式”与“扩展模式”动态切换

行业动态

从国内厂商动态观察 技术变化

技术浪潮已至，AI正从
“工具”迈向自主“智能体”
消费端商业模式尚在探索

风口之下，是充满无限
可能的起点

阿里云：模型迭代与服务生态升级，逐步构建智能体完善生态

阿里巴巴的Qwen系列大模型在开源与商业化方面均有显著进展。开源层面，Qwen3-Embedding系列模型的发布，特别是其8B版本在MTEB多语言榜单的领先表现，以及Qwen2.5和Qwen2.5-VL系列模型从限免到商业化的转型，是两大核心动态。

阿里云百炼平台不仅推动了上述模型的商业化，还通过资源包优惠、上线新模型（如视频编辑模型wanx2.1-vace-plus）和更新现有模型（如qwen-vl-plus）来增强其商业产品矩阵。

阿里巴巴于2025年6月6日正式开源了Qwen3-Embedding及Reranker系列模型。这一系列模型是基于Qwen3的基础架构构建的，代表了国产大模型在文本表征与排序技术方面的突破。

Qwen3-Embedding模型的推出，旨在提升语义理解与检索领域的性能，此次开源不仅为开发者社区提供了强大的工具，也进一步推动了国内大模型生态的发展。值得注意的是，Qwen3-Embedding系列模型的开源，紧随谷歌更新其Gemini 2.5 Pro模型之后，显示出全球AI领域竞争的激烈程度以及各大厂商在模型性能和应用场景上的持续探索。

Qwen3-Embedding系列模型

Model Type	Models	Size	Layers	Sequence Length	Embedding Dimension	MRL Support	Instruct Aware
Text Embedding	Qwen3-Embedding-0.6B	0.6B	28	32K	1024	Yes	Yes
	Qwen3-Embedding-4B	4B	36	32K	2560	Yes	Yes
	Qwen3-Embedding-8B	8B	36	32K	4096	Yes	Yes
Text Reranking	Qwen3-Reranker-0.6B	0.6B	28	32K	—	—	Yes
	Qwen3-Reranker-4B	4B	36	32K	—	—	Yes
	Qwen3-Reranker-8B	8B	36	32K	—	—	Yes

排序模型评测结果

Model	Param	MTEB-R	CMTEB-R	MMTEB-R	MLDR	MTEB-Code	FollowIR
Qwen3-Embedding-0.6B	0.6B	61.82	71.02	64.64	59.26	75.41	5.09
jina-multilingual-reranker-v2-base	0.3B	58.22	63.37	63.73	39.66	58.98	-0.68
gte-multilingual-reranker-base	0.3B	59.51	74.08	59.44	66.33	54.78	-1.84
BGE-reranker-v2-m3	0.6B	57.03	72.16	58.36	59.51	41.38	-0.81
Qwen3-Reranker-0.6B	0.6B	65.80	71.31	66.36	67.28	73.42	5.41
Qwen3-Reranker-4B	4B	69.76	75.94	72.74	69.97	81.20	14.84
Qwen3-Reranker-8B	8B	69.02	77.45	72.94	70.19	81.22	8.05

同时，阿里云百炼平台也在不断丰富其模型矩阵，推出了新的视频编辑模型wanx2.1-vace-plus，该模型支持多模态输入，能够实现图文视频和视频重绘等功能。此外，视觉理解模型qwen-vl-plus-2025-05-07快照版也得到更新，显著增强了数学和推理能力。图像生成模型aitryon-plus也提升了图片清晰度和细节还原能力。

阿里云百炼平台推出了业界首个全生命周期MCP服务，覆盖从服务注册、云托管，到Agent调用、流程组合的全生命周期。这使得AI应用开发从“手工作坊”迈入“工业化生产”阶段。

2025年5月则新上架30个云部署 MCP Server，其中一方服务包含全妙、Hologres、DMS等。



百度：应用驱动，模型为基，打造AI新范式，Robotaxi成绩斐然

在2025年第二季度，百度AI的战略核心清晰地体现为“模型的世界，应用的天下”，这一理念由百度创始人李彦宏在Create 2025百度AI开发者大会上明确提出。该理念标志着百度AI战略重心的显著转移，即从过去可能更侧重于基础大模型的研发和性能提升，转向更加注重AI技术的实际应用和商业化落地。

随着大模型技术的逐渐成熟和普及，**单纯追求模型参数的庞大和benchmark得分的领先已不再是唯一的竞争焦点，如何将强大的模型能力转化为触手可及、能够解决用户痛点的应用，成为决定AI企业成败的关键。**

“心响”APP是百度在2025年第二季度推出的一款重磅AI原生应用，被定位为百度Agent战略的前锋部队，以及通用超级智能体。与传统AI助手主要作为“对话式工具”不同，“心响”APP真正具备了复杂任务拆解与执行能力，集成了浏览器、代码执行器、图表生成器等组件，能够通过自然语言驱动端到端的任务流。

百度在2025年第二季度还推出或迭代了其他一系列创新应用，共同构成了其AI Native应用矩阵。其中，**无代码编程工具“秒哒”已向全社会开放，用户可以通过自然语言描述，快速生成应用程序。**

同样，无人驾驶是“AI在物理世界的典型应用场景”，是“人工智能竞争”的关键领域，百度旗下萝卜快跑的在Robotaxi领域表现不俗，**作为“全球无人驾驶三大家”之一，正凭借安全初心、技术能力、落地经验，创造无人驾驶出海的“中国范式”，代表中国无人驾驶和AI的最高水平，参与国际竞争。**

成本控制驱动创新，萝卜快跑 2024年5月发布全球首个支持L4 级无人驾驶的大模型Apollo

ADFM：大模型重构自动驾驶，萝卜快跑能够以更低成本、更快速度在中国复杂的城市道路提供服务。萝卜快跑第六代无人车比特斯拉cybercab更早落地两年，成本不到Waymo无人车的1/7。

国内经验沉淀是出海的安全保障和基础，萝卜快跑技术和环境适应性和泛化性更强。未来基金管理合伙人Gary Black：由于中国复杂的交通状况，萝卜快跑的行駛环境是最具挑战性的环境之一。

	萝卜快跑	特斯拉Robotaxi	Waymo
⑤ 类别			
④ 车辆成本	2.8万美元	3万美元 (Cybercab目标)	20万美元
③ 技术路线	多传感器融合 优先考虑安全性	纯视觉	多传感器融合
② 全球布局	全球15座城市 即将进入 欧洲、新加坡、马来西亚	初期奥斯汀推出 年底走出奥斯汀	已落地凤凰城、旧金山 洛杉矶、奥斯汀 已在东京测试 即将在纽约测试
① 服务单量	1100万	试运营	1000万

火山引擎Force大会聚焦Agentic AI时代

2025年火山引擎Force原动力大会的核心主题被明确地定位为“乘风破浪共赴Agentic AI时代”。这一主题精准地捕捉了当前人工智能领域发展的核心趋势，并彰显了火山引擎在推动AI技术向更高阶、更实用阶段迈进的决心和战略方向。

在为期两天（6月11日至12日）的会议中，火山引擎围绕其最新发布的豆包大模型1.6、多模态视频生成模型Seedance 1.0 Pro，以及全面升级的AI云原生产品体系，系统性地展示了从底层大模型研发到上层Agent应用开发的全链路解决方案。



火山引擎在本次Force大会上清晰地传递出其核心愿景：致力于将最新的AI技术和模型能力开放给企业客户，推动人工智能在企业业务中的深度应用和价值创造。

火山引擎的目标是让小模型不再仅仅局限于信息的识别、处理和生成，而是能够进一步实现自主推理、规划行动并执行复杂任务，从而成为构建企业级复杂业务闭环的智能体。

火山引擎在Force大会上重磅发布了其AI云原生全栈服务，该服务的核心导向明确为“Agent构建”，旨在帮助企业基于强大的大模型能力搭建自主的Agent，进而实现规模化商业应用。这一全栈服务可以概括为“模型 + AI云原生产品”的组合，针对当前Agent在落地过程中普遍存在的“思考深度不够”、“多模态适配不完善”、“工具调用能力缺失”以及“调用成本高昂”等四大核心痛点，火山引擎计划从模型、云平台、开发工具以及垂直应用等多个层面协同发力，提供一套主打“生产级”的Agent解决方案。

特性	doubao-seed-1.6 (综合模型)	doubao-seed-1.6-thinking (思考强化版)	doubao-seed-1.6-flash (极速版)
定位	All-in-One综合模型	深度思考能力强化	极速响应，兼顾能力
核心能力	深度思考、多模态理解、图形界面操作	代码生成、数学解题、逻辑推理等基础能力提升	深度思考、多模态理解、图形界面操作
上下文窗口	256K (国内首个支持)	256K	256K
思考模式	开启、关闭或自适应思考 (根据提示词难度自动决定是否启用深度思考)	-	-
延迟 (TOPT)	-	-	10ms
视觉理解	-	-	可媲美其他旗舰模型

豆包大模型1.6系列主要版本特性对比

腾讯：混元大模型的持续进化，AI Agent战略的加速推进

腾讯在2025年第二季度持续加大AI技术研发投入，其核心的混元大模型能力得到显著提升，研发投入持续加码，为技术突破奠定了坚实基础。根据财报数据，腾讯在2025年第一季度的研发投入已达到189.1亿元人民币，同比增长21%，这一高投入态势在第二季度得以延续。**自2018年至今，腾讯累计研发投入已超过3026亿元人民币**。混元大模型作为腾讯AI战略的核心，其总体性能相比上一代提升了50%，在国内的基础大模型中处于领先地位。这一性能提升不仅体现在通用能力的增强，更在多模态能力上取得了突破。

例如，腾讯混元自研推理模型Thinker（T1）在年初上线元宝App后持续快速迭代，同时腾讯还发布了主打更快任务处理能力的新一代快思考模型混元Turbo S。

此外，混元图像2.0、混元3D v2.5及混元游戏视觉生成等一系列多模态模型也同步“上新”，其中混元图像2.0率先实现“毫秒级”生图，混元3Dv2.5凭借业界首创的稀疏3D原生架构，实现了可控性与超高清生成能力的代际飞跃

2025年第二季度，腾讯将AI Agent（智能体）作为产品创新的核心战略方向，并为此推出了腾讯云智能体开发平台（TCADP），旨在降低企业构建智能体应用的门槛，加速AI在产业中的落地。 腾讯CEO马化腾详细阐述了对Agentic AI的战略思考，其当前战略是“两条腿走路”：一类是“通用”Agentic AI，任何人都可以创建代理，让它到外部世界替用户完成任务；另一类则是嵌入微信生态、基于微信独特环境的Agentic AI。腾讯云在2025年5月21日举办的AI产业应用峰会上，将大模型知识引擎全面升级为“腾讯云智能体开发平台”（Tencent Cloud Agent Development Platform, TCADP）。该平台整合了腾讯云行业领先的检索增强生成（RAG）技术、全面的Agent能力以及经过实战打磨的、贴合用户需求的功能，旨在帮助企业快速激活私域知识，构建专属智能体。



MiniMax：开源自主研发 MiniMax-M1 系列模型及推出Agent

MiniMax 近期推出了其 Agent 产品及开源的 MiniMax-M1 系列模型，展现了其在通用人工智能领域的深厚积累。其 Agent 技术架构围绕处理长程复杂任务、提升模型效率和多模态能力展开，核心创新包括混合架构模型 MiniMax-M1、高效的强化学习算法 CISPO，以及构建开放生态的多智能体协作协议 MCP。这些技术使得 MiniMax Agent 在文本理解、代码生成、工具调用等方面表现出色，并在办公自动化、内容创作、软件开发及视频制作等多个应用场景展现出巨大潜力。

MiniMax Agent 被赋予了“反思”能力。在任务的每一个关键节点，它都会主动审视当前进展，发现潜在问题，并自主调整策略，确保最终交付的成果满足甚至超越用户的预期。这种反思机制结合长程记忆，使得 Agent 能够进行复杂的多步推理和规划。

MiniMax Agent 在工具调用（Tool Use）方面展现出极高的效率和稳定性。MiniMax 通过其 MCP（Multi-Agent Collaboration Protocol）协议构建了一个开放的应用生态——MCP Market。这使得 Agent 能够无缝集成各种外部工具和服务，如 GitHub、Google Maps、Figma 等，极大地扩展了其应用场景和能力边界。

日期	Minimax发布内容
2025-06-17	发布全球 首个开源大规模混合架构推理模型MiniMax-M1 。该模型总参数4560亿，激活参数459亿，支持100万上下文输入及8万Token推理输出，在多个生产力场景中性能出色，成本效益高。同时，MiniMax-M1的技术细节在ArXiv平台公布，模型在Hugging Face和魔搭社区等平台开源。
2025-06-18	发布视频模型Hailuo-02。该模型采用创新的Noise-aware Compute Redistribution(NCR)架构，在将模型参数和训练数据量分别提升3倍和4倍的同时， 实现了2.5倍的效率飞跃 ，在知名测评榜单Artificial Analysis视频竞技场中拿下全球第二。
2025-06-19	发布MiniMax Agent。该Agent基于MiniMax-M1模型，具备强大的长程记忆与复杂推理能力，能够处理长文本、复杂代码库等任务，并 支持开发者以简单的XML格式描述工具功能 ，模型能够自动理解工具的输入输出、参数约束，并生成符合要求的调用代码。
2025-06-20	发布国内首个成片创作助手Hailuo Agent。该工具通过自然语言驱动全局、全流程工具集调用，实现 “零门槛专业级视频生成” ，用户输入“制作一段城市延时摄影”，即可一键生成包含分镜、剪辑、配音的完整成片。
2025-06-21	发布音色设计Voice Design。 用户可以通过自然语言描述来精准控制音色的多个维度。

MiniMax Agent 被赋予了“反思”能力。在任务的每一个关键节点，它都会主动审视当前进展，发现潜在问题，并自主调整策略，确保最终交付的成果满足甚至超越用户的预期。这种反思机制结合长程记忆，使得 Agent 能够进行复杂的多步推理和规划。

MiniMax Agent 在工具调用（Tool Use）方面展现出极高的效率和稳定性。**MiniMax 通过其 MCP（Multi-Agent Collaboration Protocol）协议构建了一个开放的应用生态——MCP Market**。这使得 Agent 能够无缝集成各种外部工具和服务，如 GitHub、Google Maps、Figma 等，极大地扩展了其应用场景和能力边界。

全球AI发展三大趋势：技术落地、商业化探索与开源协同

“甲子智库解读”

加速布局

软件与服务的“智能体化”正在加速

进化跃迁

AI商业化模式面临压力并持续演进

双轨AI

专有与开源并行的双轨AI战略日趋普遍

技术加速落地，重构人机交互模式

全球科技巨头正推动AI智能体从概念迈向规模化应用。谷歌在I/O大会上展示了集成于Workspace和Android的任务自动化智能体，以及具身交互项目Astra；微软通过Build大会构建“开放式智能体网络”，覆盖开发、终端与云端，如GitHub Copilot和Dynamics AI Foundry。

商业化路径分化，传统模式承压

AI商业化面临传统数字模式的挑战。谷歌AI概览功能虽提升搜索体验，却导致网站点击率下降，威胁广告收入；百度以AI云服务实现营收增长，验证企业级解决方案的可行性；腾讯通过广告能力升级与生态融合变现；阿里则通过电商AI工具创造核心业务价值。行业探索路径呈现多元化：企业级付费服务、AI增强功能订阅及生态效率提升成为新方向，而企业需通过清晰ROI证明AI投入的合理性，传统广告模式或加速迭代。

专有与开源双轨并行，平衡创新与生态

科技巨头采用“专有+开源”混合策略构建竞争壁垒。谷歌在推进Gemini等闭源模型的同时，开源MedGemma和Gemma；阿里自研通义千问但开源衍生模型，腾讯计划开源多模态模型；微软Azure AI Foundry兼容Hugging Face开源模型与专有系统。开源策略可吸引开发者、加速技术渗透与反馈收集，而尖端技术仍保留于闭源体系以维持差异化优势。双轨模式既推动生态扩张，又确保核心技术的控制权，成为AI竞争的新常态。

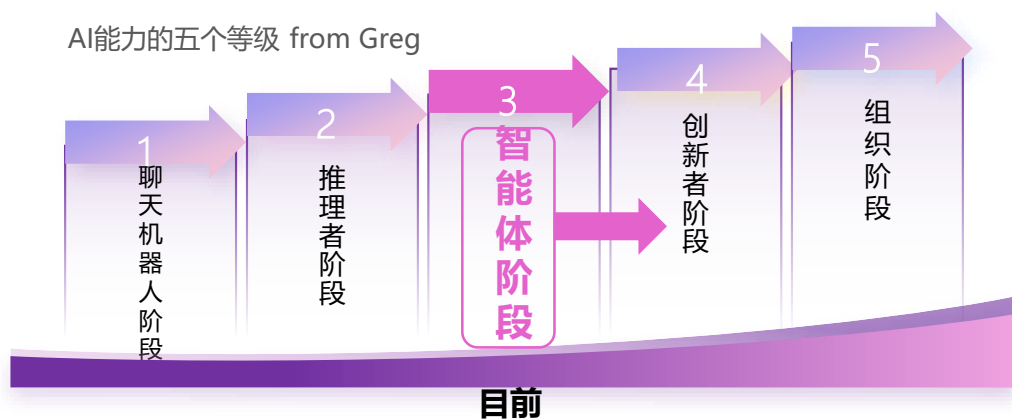
2025年成为“智能体”的关键之年

AI能力的五个等级：

1. 聊天机器人阶段，2. 推理者阶段，3. 智能体阶段，4. 创新者阶段，5. 组织阶段。

目前我们处于第三阶段，即将进入第四阶段。这个阶段的AI将能够进行真正的创新和发现。

——Greg Brockman: OpenAI 联合创始人



竞争核心已从通用大模型转向为特定任务优化的“Agent大脑”，并开始比拼数据生态与独特交互设计的护城河。然而，全行业仍普遍面临着信息准确性（AI幻觉）、版权合规与高昂算力成本的共同考验。未来，真正的胜者将是那些能够在保证可靠性的前提下，将AI Agent无缝融入特定 workflow，实现更高程度自动化的玩家

2025年3月至6月，智能体AI（Agentic AI）生态系统进入了关键的成熟期。行业叙事已从讨论孤立的智能体开发框架，转向构建一个由新兴通信协议定义的、分层的、可互操作的体系架构。这一转变标志着智能体经济从理论走向实践的重大拐点。

“Agent元年”——这个在2024至2025年间被科技和商界频繁提及的词汇，精准地捕捉到了人工智能（AI）发展的新浪潮。它标志着AI正从一个听从指令的“工具”，向一个能够自主理解、规划并执行复杂任务的“智能体”（Agent）演进。

“
甲子智库解读



第二章

大咖观点：
萃取真知灼见，关注前沿风向

大咖观点

——
改变、机遇和希望



“超级智能”会更快地出现

Elon Musk: 我们正处于智能大爆炸的极其、极其早期的阶段。我认为我们相当接近数字超级智能了。如果今年没有发生，明年肯定会发生。

Elon Musk

Tesla、X、SpaceX CEO



- 也许有10%到20%的毁灭可能性。但往好处想，那可是80%到90%的概率能有一个极佳的结果。

AI时代的经济预言：十倍增长

AI将如此深刻地改变未来，很难量化其影响程度。但假设我们不会偏离正轨，AI不会杀死我们所有人，你最终会看到一个规模约为当前经济十倍的经济体。最终，对于我们未来的机器后代，或者主要是机器后代，可能会发展成K1级文明或更高级别。我们谈论的是一个比今天大数千倍、也许数百万倍的经济体。

超级智能与机器人



比任何人类都更聪明的数字超级智能将在今年（2025）或明年内出现，但也伴随着 10% 到 20% 的毁灭性风险。确保 AI 安全最重要的一点是让 AI 严格地“追求真相”。

“超级智能”会更快地出现

AGI 5 - 10 年之约下的技术与社会变局

Demis Hassabis: AGI只有5到10年的时间

Demis Hassabis

Nobel Laureate

01:00:00



Demis Hassabis On The Future of Work in the Age of AI

创立 DeepMind 时有一个宏大的目标：找出智能，然后用它来解决其他所有问题。他相信他们已经走上了正轨，可能距离许多人所说的 AGI 只有 5 到 10 年的时间。这是一个可以做人类可以做的所有思考事情的系统。这是一个伟大的想法，而且越来越近了。

AGI 的试金石是“不出错”，而非“超能力”。

真正的 AGI，其标志不在于解决顶尖难题的巅峰表现，而在于横跨所有简单与复杂任务的持续、稳定、不出错的泛化能力。一个在国际奥数竞赛上能夺金，却在简单数数上会犯错的系统，并未真正理解世界。评估智慧时，稳定性和一致性比峰值性能更为本质；

技术的能力边界，不等于人类社会的需求边界。

即使 AGI 在技术上能做所有工作，我们仍会出于人性的需求，选择将某些角色（如护理）保留给人类，因为这些角色的核心价值是机器无法提供的同理心与人性关怀。效率和能力并非唯一的标尺，人类的情感需求将定义一部分工作的永恒价值；

技术提供的是“可能性”，但无法解决人类的“意愿”问题。

AI 或许能通过解决能源等“根节点问题”，从物理上打破零和博弈的困局（例如，若能源近乎免费，水资源冲突便迎刃而解）。然而，我们早已知道解决方案的气候问题，却因缺乏协作意愿和牺牲精神而停滞不前。技术可以递上解决问题的钥匙，但开门的动作仍需人类社会自己完成；

评估 AI 风险的视角，应从“绝对风险”转向“相对风险”。

与其只忧虑 AI 可能带来的挑战，不如思考一个没有 AI 的未来会怎样。面对气候变迁、流行病等既有的巨大挑战，一个缺乏 AI 这样革命性工具的未来，可能才更令人绝望。AI 不仅是一个需要被管理的新风险，更是应对其它现存巨大风险的必要杠杆；

数字智慧的指数级发展，会被物理世界的线性规则所“制动”。

即使 AGI 诞生，它对现实世界（如改造工厂、部署机器人）的影响也需要时间，因为物理世界有其固有的运行规律和变更时滞。这为社会适应新技术提供了宝贵的缓冲期，意味着转变更可能是渐进的，而非毁灭性的瞬间颠覆；

社会系统的“软件”升级，已严重落后于技术的“硬件”迭代。

当 AGI 可能从根本上改变经济学中关于“价值”和“金钱”的定义时，我们的经济学理论和社会治理模式却尚未开始为此准备。

我们正以惊人的速度构建一个全新的技术世界，却缺乏相应的新社会科学理论来理解和引导它，这形成了潜在的巨大风险。

AGI 5 - 10 年之约下的技术与社会变局

AI时代将进入“软件3.0”

Andrej Karpathy:

自然语言在成为新的编程接口，大模型会完成剩下的工作

Andrej Karpathy Eureka Labs

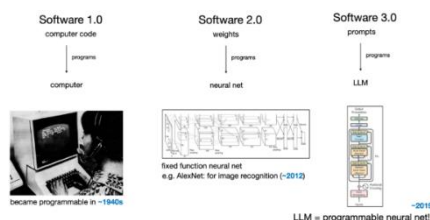


我们正在进入「软件 3.0」时代，自然语言在成为新的编程接口，大模型会完成剩下的工作。

软件开发的一次范式转移：软件将进入3.0 阶段

Karpathy将软件的演进划分为三个阶段：

- **软件1.0 (Software 1.0):** 经典代码。由人类用Python、C++等编程语言编写的明确指令集。计算机完全按照这些指令执行。
- **软件2.0 (Software 2.0):** 权重。随着深度学习的兴起，软件的核心从“代码”变成了神经网络的“权重”。这些权重不是由人编写的，而是通过在大量数据上进行优化（训练）得到的。例如，一个图像分类器就是一个由权重构成的程序。



软件3.0 (Software 3.0): 提示词 (Prompts)。

LLM本身就像一个可编程的计算机，而它的“编程语言”就是自然语言（如英语）。我们通过编写提示词来“编程”LLM，让它完成复杂的任务。这标志着软件开发的又一次范式转移。

Karpathy用一个“软件版图”的比喻来形象说明：如果说软件1.0是GitHub上的所有代码，软件2.0是Hugging Face上的所有模型，那么软件3.0正在以前所未有的速度“吞噬”和重构前两者。

AI时代将进入“软件3.0”

空间智能补充当下人工智能的完整

李飞飞：如果没有空间智能，人工智能将是不完整的。

李飞飞

World Labs



- 语言是思想和信息的令人难以置信的强大编码，但它实际上不是3D物理世界的强大编码。如果审视人类智能，很多都超出了语言的范畴。
- 语言虽然重要，但它是捕捉世界的有损方式，更重要的是，语言是纯粹生成的，语言在自然界中并不存在。

- **空间智能是AI的基石**：人工智能若要完整，就必须具备理解、推理和生成3D世界的能力，因为这是我们所处物理世界的基本属性。
- **构建3D世界模型是下一个前沿**：继语言模型之后，攻克3D生成式基础模型是解锁大量应用（从创意设计到机器人技术）的关键，也是当前AI领域最困难的问题之一。



- **无畏精神驱动创新**：无论是创建ImageNet还是创办新公司，敢于挑战不确定性、大胆假设的“无畏”精神是推动科学和技术突破的核心动力。
- **AI的终极目标是赋能人类**：技术应以人为本，作为协作工具来增强人类的能力，以应对医疗、教育等领域的复杂挑战，促进社会福祉。
- **数据与模拟至关重要**：高质量、多模态的数据（包括被低估的触觉数据）和高保真模拟环境，对于训练更强大的机器人和通用智能体是不可或缺的。

空间智能补充当下人工智能的完整

英伟达：中国市场非常重要

黄仁勋：AI开发者在中国，所以，AI技术体系就会生长在那里

黄仁勋 Jensen Huang

NVIDIA CEO



- AI开发者在中国，所以，AI技术体系就会生长在那里
- 在中国的情况是，他们拥有充足的能源供应。从很多角度来看，中国的技术水平已经足以自给自足了。

- AI技术与中国市场关联：AI开发者集中在中国，AI技术体系会在此生长；中国有充足能源供应，技术水平能自给自足。
- 中国市场重要性：中国市场规模庞大，是全球50%的AI研究人员的家园，对技术平台成功至关重要；英伟达在中国市场份额因产品限制下降，未来可能面临巨额收入损失。
- 中国技术实力：中国有非凡制造能力，芯片制造数量和能力逐年翻倍增长
- AI技术革命与需求：推理AI发展迅猛，计算需求呈指数级增长；英伟达设计Blackwell等产品以适应技术突破。
- 未来就业与经济影响：AI和机器人技术是工作生成器而非替代者，可用于扩展世界GDP，缓解劳动力短缺。
- 中美关系与芯片供应：即便美国选择退出中国市场，华为有能力掌控中国市场，甚至开拓全球其他市场；没有美国技术，中国技术的可用性将填补芯片市场。



英伟达：中国市场非常重要

AI时代的人性坐标与技术进化

Sundar Pichai: AI是超越火与电的技术，因为它可以自我迭代，人性是末日风险的最终解药

Sundar Pichai

Google & Alphabet CEO



- 当一个威胁（如P(Doom)）真大到足以毁灭全人类时它将成为一个前所未有的能统一全人类去解决它的强大动力，这形成一种“风险自我调节机制”。人类文明的历程证我们总能在关键时刻挺身而出集体应对挑战。

- **技术信念的本源：**皮查伊在印度金奈的成长经历（缺水困境、五年等待一部电话）塑造了其技术的终极信仰——技术不仅是工具，更是打破生存枷锁的文明驱动力。这种信念促使他以“解决人类真实需求”为核心使命，带领谷歌通过AI重构信息获取与生活交互的范式。
- **AI的颠覆性本质：**自我迭代的文明加速器。AI是超越火与电的“终极技术”，因其具备“递归自我改进”能力，能像AlphaGo通过自我对弈实现指数级进化，未来将成为创造力的独立引擎。AI将催生“创新包”效应，降低从想法到落地的门槛（如编程民主化），释放80亿人的认知潜能，同时重塑医疗、教育等行业（如癌症早筛、个性化学习）。
- **谷歌的战略实践：在平衡中推进自我颠覆。**合并Google Brain与DeepMind，整合学术多元性与AGI目标导向的文化冲突，形成技术攻坚合力。搜索业务采用“双轨制”——AI Overviews作为现有框架的总结层，AI Mode作为前沿试验田，通过“验证-回迁”机制确保创新与稳定共生。自研TPU芯片、开源TensorFlow生态、布局边缘计算，构建AI时代的“水电煤”底层能力。
- **风险哲学：人性作为终极安全网。**当前AI处于“锯齿状智能”阶段，AGI可能略晚于2030年实现，但2030年前需应对其社会冲击。当AI威胁（P(Doom)）上升到毁灭级时，人类将触发“风险自我调节机制”——共同威胁会成为统一全球智慧的催化剂，历史经验证明人类具备集体应对危机的理性。
- **未来图景：人机共生的文明新形态。**AR将成为下一个范式，安卓系统向“智能体”进化，XR设备通过多模态交互打破语言与空间壁垒。AGI的终极价值不是解答宇宙奥秘，而是帮助人类更深刻地理解自身——技术的终点应回归对人性的探索。皮查伊的核心原则是“信号降噪”。他将外界的批评、股价波动和媒体热议视为“噪音”，而将第一性原理、核心技术数据和不可逆转的长期趋势视为“信号”。在对话中，他透露正是在外界对Waymo自动驾驶项目最悲观时，他基于内部技术遥遥领先的“信号”，决定“加倍投入”。这一思维框架让他能够顶住巨大压力，专注于长期价值，最终带领谷歌在AI竞赛中重回领先地位。

AI时代的人性坐标与技术进化

聚焦场景、融合应用与商业价值挖掘

朱啸虎：所有的AI应用都是套壳，创业公司不要浪费一分钱去训练底层模型

朱啸虎

金沙江创投董事总经理



3年做到

2年做到

5个月做到

- 不拥抱AI的企业肯定会被淘汰!
- 令不要迷信A1，聚焦尖刀场景尽快落地!
- 令不要投入底层技术，AI不够人工补;

- “套壳论”的延续与强化：许多企业仅依赖现有技术“套壳”商业化，而非底层突破。例如，美国AI绘图公司多基于开源工具整合功能，缺乏深度融合创新，而国内部分AI写作平台仅通过简单模板生成内容，夸大实际效果。某些企业仅通过优化客服系统等“套壳”操作，即可年增数亿利润。

所有的应用都是套壳应用，关键是如何建立长期壁垒

- 深度整合工作流/编辑工具;
- 专有硬件/IP内容/数据;
- 必须人工交付的苦活累活;

AIGC应用重要的
不是生成能力，
而是编辑能力，
交付结果的能力!

GSR Venture

AI不够人工补，保证交付质量，更适合中国的商业模式



- 软件的毛利!
- 软件的粘性和复购!
- 逐步提高AI的比例!

- 技术壁垒的争议：大模型 口 技术已足够成熟，创业公司无需在底层技术上过度投入，应聚焦垂直场景的价值创造。例如，通过AI赋能企业降本增效(如优化客服、营销)，而非盲目追求通用大模型研发。部分企业以“技术壁垒”为噱头，实则是资本驱动的泡沫。

聚焦场景、融合应用与商业价值挖掘

大咖观点

危机、困境和博弈



人类必须拥有不同于AI的“独有的能力”

Geoffrey Hinton: 人工智能将在多个领域「取代所有人」，只有顶尖技能人才能够找到 AI 无法处理的工作。

Geoffrey Hinton Nobel Laureate



- 在个性化算法时代，人类的共同体验消失了。数字智能必然超越生物智能，但20%的灭绝概率正被忽视。
- 当AI能做所有脑力劳动时，人类需要重新定义自己的存在意义。这不仅是经济转型，更是文明转型——我们必须找到除了生产力之外的其他方式来体现人类价值。这个挑战将影响每一个行业、每一个职业，甚至每一个人。

AI风险的两大类别：人类滥用与机器觉醒

第一类风险涵盖大部分风险和所有短期风险，包括网络攻击、生物武器、选举腐败、回音室效应和自主杀伤性武器。这些威胁源于人类如何使用AI技术。如，2023年到2024年间，网络攻击增加了约12倍，增长率达到1200%。这种爆炸性增长很大程度上归因于大型语言模型使网络钓鱼攻击变得更容易。

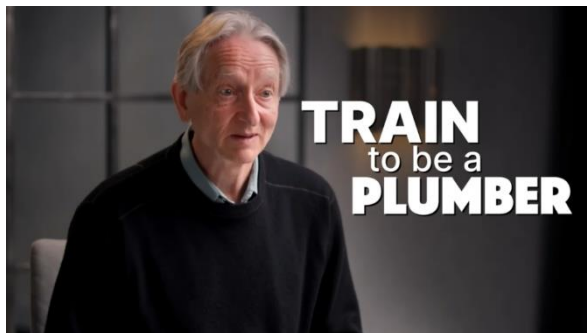
第二类风险是存在性威胁：AI系统本身变得比人类更聪明，可能决定人类是多余的。“是的，这是真实的风险”，辛顿断言。但***“我们从来没有遇到过比我们更聪明的东西”***，任何声称知道如何应对的人都是胡说八道。

关于概率评估，学术界存在极端分化。扬·勒昆认为风险微乎其微，坚信人类总能控制AI系统。另一极端是埃利泽·尤德科夫斯基，他确信超级智能AI必然消灭人类。辛顿认为这两种立场都过于极端，**估计AI消灭人类的概率在10%到20%之间**。为说明威胁程度，辛顿比喻：“如果你想知道当你不是顶级智能时生活是什么样的，问问鸡就知道了。”

我们得重新定义什么叫‘人类独有的能力’。

AI 正在打碎的，是人类三道最关键的认知护城河——

- **理解边界**：语言模型不再复述，而是“读懂”；
- **推理边界**：AI 不靠数据记忆，也能构造逻辑链；
- **学习边界**：它们已不再依赖训练集，而是边做边学。



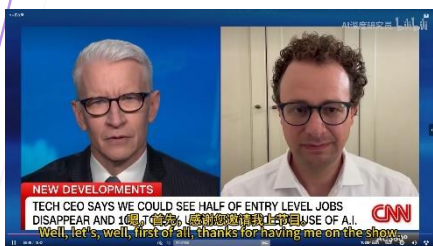
人类必须拥有不同于AI的“独有的能力”

AI可能让“就业”问题更为严峻

阿莫代：五年内一半白领工作消失，整体失业率上升20%，比90%的编码由AI完成吓人多了。

阿莫代

Anthropic CEO



- 普通民众要学会使用AI，学会理解这项技术的发展方向。如果你不会被突然冲击，你适应的机会就大得多。与以往的技术变革相比，我更担心这次的劳动力市场冲击，因为变化发生得如此之快，人们可能无法及时适应。人们最终会适应，可能会出现一个调整期。
- 在未来一到五年内，可能会有有一半的入门级白领工作消失，整体失业率可能上升到10%到20%。

技术进步速度前所未有，人类适应期可能极其痛苦

AI 技术进步飞速，能力从相当于聪明高中生跃至聪明大学生水平且不断提升。未来一到五年内，约一半入门级白领工作可能消失，整体失业率或升至 10% - 20%。与以往技术变革相比，此次 AI 引发的劳动力市场冲击速度快，人类适应期可能缩短，或出现痛苦的“调整期”。

现代社会制度面临根本性威胁，社会契约亟需重构

AI 冲击触及现代民主制度根基，民众因对经济有贡献而拥有的话语权可能因 AI 发展丧失，社会契约面临重构。需确保普通人有经济话语权和谋生方式，否则社会契约难以为继。同时，AI 时代下人类价值和目标的定义也需重新审视。

应对AI革命的全方位行动指南

普通民众要学会使用 AI 并理解其发展方向，以更好适应变化。面对 AI 创造的财富分配问题，需采取行动，避免财富过度流向 AI 公司，确保普通人能从中获益。

AI可能让“就业”问题更为严峻

不懂Agent，就是不懂AI时代的管理

OpenAI董事长：理解 AI Agent，才有资格参与下一轮组织竞争

Bret Taylor OpenAI 董事长



“Agent，不是助手，不是插件，也不是你旁边那个听话的小助手。”

“它是一个真正的‘角色’——能推理、能判断、能执行任务，甚至可以不经你同意就完成一整条流程。”

新一代Agent已成为具有自主决策权的“组织角色单元”——它能独立完成客户服务、合规审批等完整业务流程，甚至被赋予OKR目标和绩效评估体系。这种角色进化呈现三级跃迁：从个人事务助手（被动响应），到专业领域代理，最终升级为品牌前台角色（直接代表企业与客户交互）。

这种转变重构了组织的最小作战单元：过去由“岗位”构成的结构，正在被“人-Agent协同网络”取代。企业若仅将Agent视为效率工具，实则是用旧框架束缚新生产力；唯有将其作为组织设计的核心变量，通过权限配置、目标管理和责任归属的系统性重构，才能在竞争中掌握战略主动权

Agent不是工具，而是角色



Agent 与“助手”在角色上的差异

助手	Agent
辅助你工作	独立完成工作
等你发指令	自己判断行动
提高效率	直接产出成果
是插件	是员工

让 Agent 成为组织成员，而不是外挂，这些问题值得深思

问题	说明
它是否拥有明确目标？	有具体任务边界
它是否有执行权限？	能直接调用系统、触发操作
它归属于哪个角色/部门？	有组织结构中的位置
谁为它的结果负责？	明确管理与问责机制
它的绩效如何衡量？	能够通过数据追踪结果表现

不懂Agent，就是不懂AI时代的管理

SaaS 软件的AI时代求生之路

Satya Nadella: SaaS软件没有活路，除非把自己和MCP结合，融入企业智能体网络。

萨提亚·纳德拉

微软 CEO



- SaaS软件没有活路，除非把自己和MCP结合，融入企业智能体网络。
- 以后企业招聘员工，不只是招聘一个人，可能还在招聘员工所携带的智能体（所以智能体的知识产权如何界定）

- **技术栈革新：**AI时代需从基础设施到数据层全面重构技术栈。Azure区域转型为“AI工厂”，整合GPU、存储等资源；数据层嵌入LLM智能生成查询，既利用15年技术积累又实现智能升级，以适配AI工作负载的资源需求。
- **产品变革模式：**Microsoft 365通过三种模式进化。全新UI集成AI智能体（如研究员角色）与Copilot Studio；Teams支撑多人协作中的AI介入；沉浸式工作模式让AI如数据科学家般嵌入Excel、文档等场景，使Office画布成为带聊天功能的IDE。
- **SaaS未来转型：**垂直SaaS需融入AI智能体网络，不能仅作记录系统。Dynamics 365通过MCP服务器参与多智能体业务编排，SaaS应支持MCP等平台，成为AI智能体网络的后端，而非固守传统 workflow 构建模式。
- **智能体归属管理：**企业AI智能体属公司IP，微软用Entra ID实现身份管理与数据隔离，如区分工作与个人Copilot账户。员工个人智能体进入工作场景需确保数据不泄漏，类似个人与公司邮箱的隔离机制。
- **零成本智能价值：**智能成本趋近零将推动生产力增长，如斯坦福医学用多智能体框架提升肿瘤会议效率，医疗领域20%的GDP占比中，AI可优化工作流程、改善护理。技术需创造实际价值以获社会许可，应对能源消耗等挑战。
- **可持续发展路径：**以“每美元每瓦特的令牌数”为标准，微软作为新能源大买家，推动高效用能。科技行业2%-3%的电力消耗若翻倍，需在医疗、材料科学等领域创造价值，实现“可持续的富足”。
- **未来系统趋势：**计算架构中确定性与非确定性界限模糊，如Muse模型生成游戏场景。未来系统虽随机，但需在关键环节保持可解释性，通过理解“智能的物理学”实现约束与管理，避免完全脱离人类可控范围。

SaaS 软件的AI时代求生之路

AI时代跳出概念陷阱，聚焦可落地的实践

吴恩达：营销人员把Agent这个词搞坏了。Vibe Coding也是一个糟糕的术语，让人有了AI编码无所不能的幻觉

吴恩达

DeepLearning.AI CEO



- 将智能体视为连续光谱，不同系统具有不同程度的自主性。没必要花时间争论这是否是真正的智能体。让我们把所有这些都称为具有不同自主程度的智能化系统。
- 大多数成功的智能体应用都是相对简单的线性工作流，而非复杂的自主决策系统。

- **重新定义智能体：**营销人员滥用词汇致智能体概念模糊，陷入「是否自主」的哲学争论无意义。应将智能体视为「自主性连续光谱」，涵盖不同自主程度，引导开发者摆脱概念纠结，专注提升系统实用性。
- **线性工作流占据主导地位：**大多数成功的智能体应用都是相对简单的线性工作流，而非复杂的自主决策系统。当前商业场景中，线性工作流（如表单处理、合规检查等顺序任务）反而是智能体落地的主流。企业需突破流程分解、评估体系等技术瓶颈，挖掘简单工作流的智能化改造空间。
- **MCP协议现状与多智能体系统局限：**MCP 协议旨在简化数据集成，获行业认可，但目前服务器不稳定、认证系统不完善。多智能体系统中，团队内协作相对可行，跨团队协作更为超前，发展程度更低。开发者需掌握数据集成（如 MCP 协议）、动态评估框架搭建，更需培养「战术直觉」（快速定位系统瓶颈）。AI 工具如「乐高积木」，但需紧跟 LLM 进化（如上下文长度突破致 RAG 策略迭代）。
- **"Vibe Coding"误解与AI时代编程本质：**"Vibe Coding" 误导大众对 AI 辅助编程的认知，制造 AI 万能幻觉，实则 AI 辅助编程是高强度智力活动。从历史发展看，编程技术进步吸引更多参与，AI 时代理解编程原理（告诉计算机需求）愈发关键。
- **AI创业的核心成功要素：**创业成功首要因素是速度，熟练团队执行速度优势明显。技术知识（如判断 MCP 协议适用性）稀缺且比商业策略更稀缺重要，AI Fund 倾向与深度技术人员合作，利用其技术直觉提升速度，商业知识相对易获取。

AI时代跳出概念陷阱，聚焦可落地的实践

AI的浪潮，淘汰的不是人，而是岗位

刘凯：大航海时代，别再跟AI比划桨了，争取成为领航员和轮机长

刘凯

五源资本合伙人



- AI带来的最深刻的变革，不是一场你死我活的竞争，而是一次岗位升级的邀约。它用一种近乎强制的方式，淘汰了划桨这个工种，从而逼迫我们所有人离开那个低矮、拥挤、视野受限的划桨手座位。
- AI的浪潮，淘汰的不是人，而是岗位。它淘汰了划桨手，但同时创造了对领航员和轮机长的大量需求。令不要迷信AI，聚焦尖刀场景尽快落地！

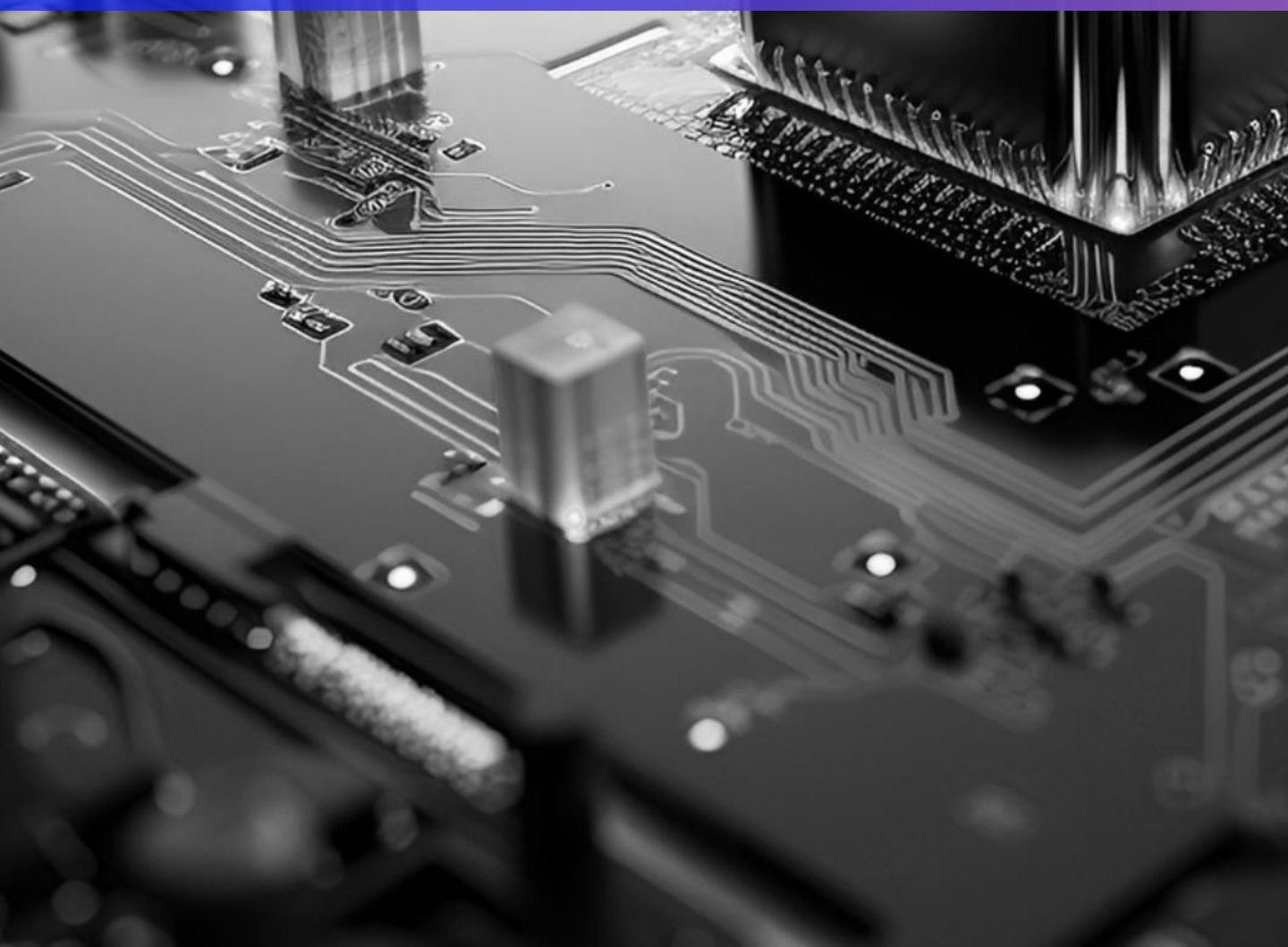
- **AI创业还处于早期：**刘凯认为当前AI创业才刚刚开始，和2010年移动互联网的早期参与者一样，才刚刚走到起点。AI领域除了大模型本身之外，还没有看到像移动互联网时代那样级别的产品落地，所以现在AI创业阶段，还比2010年更早。
- **AI技术发展从集中走向开放：**过去AI世界像“旧教”时代，知识和能力集中在大厂手中。但现在进入“新教”式时代，技术更加开放、工具更容易上手，创业者都有机会构建自己的模型、跑出自己的应用。
- **AI技术范式变革：**这一代的生成式AI对数据的理解与使用环节与上一代machine learning的模式能力非常不同，具备了以前所有的软件公司都不具备的能力，将孕育很多新的工具和新的公司。
- **AI投资估值相对论：**刘凯认为贵跟便宜是相对的，只要选对了赛道，而且这个赛道真的在变大，团队也在成长，那这个项目从长期看就不贵。
- **AI投资判断标准：**对于技术类项目，更多靠判断团队和人；对于产品类应用，老指标体系正在崩塌，很多时候也还是在“看感觉”。
- **AI应用门槛降低：**过去想做出影响力大的AI产品，通常认为必须“出海去美国”，或者依赖大厂和知名研究员背景。现在门槛在快速降低，越来越多的年轻人，包括高校学生自己就可以做非常好的应用。
- **AI商业转化优势：**与移动互联网时期不同，大模型时代大家都意识到技术背后有高昂的成本，一定程度上更利于商业转化。垂直行业模型框架的价值会更加凸显，其既能结合传统软件开源模式，也能进行针对性的定制设计。

AI的浪潮，淘汰的不是人，而是岗位



第三章

产品解读：
透视设计精髓，评估市场潜力



多模态AI模型快速走向消费段应用，多类产品聚焦Agent方向

多模态AI和Agent技术取得了显著进展，正在深刻影响各行各业。多模态AI不再局限于单一数据类型，而是能同时处理和理解文本、图像、音频、视频等多种模态信息，这极大地拓展了AI的感知和推理能力。

AI Agent（智能体）作为能够自主理解指令、规划、决策并执行任务的智能系统，正成为推动AI技术革命的关键力量。基于多模态大模型的Agent，其能力边界被无限扩展，能够调用各种工具和API，甚至与外部硬件交互，从而解决更复杂、更现实的问题。Gartner预测，到2028年，企业应用中代理型AI的使用比例将大幅增长。然而，数据质量和Agent的可解释性仍是亟待解决的问题。

Agent概念泛化：应用场景的链接与服务生态的建设

“万物皆可Agent”

AI

数字生态的建设

AI Agent正处在协议与平台建设的初期阶段，各类协议如Agent-Manager和Agent-Proxy等，旨在为Agent间的通信、任务协作和能力扩展提供标准化框架。通过统一的协议，Agent能够更有效地调用工具、整合资源、协同工作，从而突破单一模型的局限，处理更复杂的任务。这种标准化的努力，结合日益强大的底层模型和丰富的开发平台，正推动Agent生态走向成熟，使其能力呈指数级增长。

AI

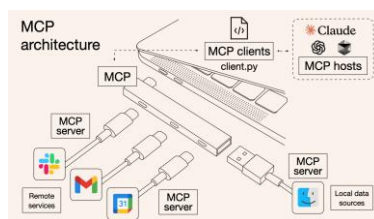
应用场景的链接

过往的数字化产品，无论是操作系统、搜索引擎还是各类应用软件，其核心都在于“人机交互”。如今，这些产品都显现出向Agent模式演进的清晰趋势。操作系统正集“Copilot”类的智能体，变被动响应为主动服务；搜索引擎则从提供链接转向直接生成答案和执行任务。这标志着一个重要的范式转移：产品不再仅仅是工具，而正在成为能够理解用户意图、主动完成目标的智能伙伴。

生态建设：Agent的“协议战争”是个伪命题，真正的故事是定义技术栈的层次

分层架构取代“协议战争”：市场并非陷入一场“赢者通吃”的协议战争，而是在协作构建一个分层的“智能体技术栈”。Anthropic的模型上下文协议（Model Context Protocol, MCP）已成为智能体与工具（Agent-to-Tool）通信（即“垂直整合”）事实上的标准。与此同时，谷歌的Agent2Agent（A2A）协议与IBM的**智能体通信协议（Agent Communication Protocol, ACP）**正在引领智能体间（Agent-to-Agent）通信（即“水平整合”）的标准化。

这些协议并非直接竞争，而是共同构成了一个新兴的、功能互补的“智能体技术栈”的不同层次。这种分层架构类似于传统网络中的OSI模型，每个协议都旨在解决一个独特的互操作性挑战，共同为复杂的多智能体系统奠定基础。Anthropic的模型上下文协议（Model Context Protocol, MCP）已牢固确立其作为基础性“垂直整合”层的地位，其核心目标是标准化单个AI智能体如何安全、可靠地连接到外部工具、数据和服务。它被广泛比喻为AI智能体的“USB-C端口”，为不同系统间的即插即用提供了可能



图灵社区

特性	模型上下文协议 (MCP)	Agent2Agent (A2A) 协议	智能体通信协议 (ACP)
主要焦点	智能体-工具 (Agent-to-Tool)	智能体-智能体 (Agent-to-Agent)	智能体-智能体 (Agent-to-Agent)
核心目标	标准化单个智能体访问外部工具、数据和知识的方式	实现跨平台、跨供应商的企业级智能体协作与互操作	提供一个简单、轻量、对开发者友好的智能体间通信方式
发起方/主要支持者	Anthropic, OpenAI, 微软, 谷歌	谷歌发起，后捐赠给Linux基金会，获微软、AWS、Salesforce等100多家公司支持	IBM, Linux基金会 (BeeAI项目)
治理模式	开放规范，社区驱动	开放标准，Linux基金会治理	开放标准，Linux基金会治理
通信风格	JSON-RPC 2.0 over HTTP	JSON-RPC 2.0 over HTTP(S)，支持SSE流式传输	RESTful over HTTP，MIME类型可扩展
关键特性	客户端-服务器架构，三阶段生命周期，工具/资源发现	“智能体名片” (Agent Card) 用于发现，结构化任务对象，多模态消息	异步优先，支持同步；无需SDK；离线发现；支持任何架构模式
主要用例	为智能体提供数据库访问、API调用、文件系统操作等能力	跨企业应用的复杂 workflows，如客户支持、供应链协同	需要快速原型设计和灵活集成的多智能体系统，如研究与开发场景

各协议的比较，资料来源于网络

请注意：大型科技公司以“开放”为名，行“平台统治”之实

开放标准成为战略武器：大型平台公司，特别是微软和谷歌，战略性地拥抱A2A和MCP等开放标准。此举并非纯粹的利他主义，而是作为一种竞争工具，旨在建立行业联盟、获取企业信任、对抗OpenAI等竞争对手可能构建的“围墙花园”，并最终推动其底层云服务和AI平台的采用。

应用场景的链接：从“人机交互”（HCI）到“模型-系统交互”（MSI）的权力交接。

过去，软件和App设计的终极目标是“易用性”（Ease of Use），即如何让用户通过界面（GUI）更简单、更高效地完成操作。而AI Agent的目标是“零用性”（No Use），即用户根本无需关心操作过程，只需表达意图，由模型驱动的智能体直接与系统后端交互，完成所有任务。

AI Agent带来的真正革命，不是对现有“人机交互”（Human-Computer Interaction, HCI）模式的优化或改良，而是用一种全新的“模型-系统交互”（Model-System Interaction）模式取而代之。

这里的“模型”特指驱动Agent的AI大模型（如LLM），而“系统端”则指代一切可被程序调用的软件、App、API接口、数据库等。

交互方式：人不再是“驾驶员”，而是“乘客”。人只需要告诉Agent最终的目的地（Goal）。

交互过程：Agent是“自动驾驶的司机”。在接收到用户的目标后，Agent（模型）会自己去操作“汽车”（调用各种系统/API）。它会自己规划路线（制定执行计划）、自己踩油门刹车（调用API）、自己处理各种突发状况（处理错误和异常）。整个过程，机器（模型）和机器（系统）在直接对话。

Agent商业元年的另一种理解：所有产品都可以用Agent重新做一遍

用业务知识复杂度和流程复杂度可以衡量Agent产品交付的特性



图片：悦点科技任鑫琦，云启资本公众号

近期产品解读——昆仑万维天工Agent

关键能力分析

“AI Office智能体”：相关的讨论在社交媒体和行业社群中迅速升温

天工Agent明确将其定位为面向办公场景的AI智能体，致力于革新传统Office工作模式。其核心理念是大幅提升工作效率，其**“8分钟完成8小时工作”**，旨在将用户从繁琐重复的事务中解放出来，**专注于更高价值的创造性工作。**

三大专家级智能体与一个通用智能体，覆盖了高频的专业写作需求，包括但不限于深度行业研究报告、竞争对手调研分析、产品规划文档、学术论文辅助撰写、商业计划书、市场推广文案、各类报告总结以及创意性写作等。

- 文档智能体：一键生成行业研究、学术论文、商业计划等专业报告，自动嵌入条形图、雷达图等数据可视化图表，支持导出Word/PDF。
- PPT智能体：输入关键词即可生成动态演示文稿，支持在线编辑与PPTX/PDF导出，数据溯源确保权威性。
- 表格智能体：上传数据即可完成统计分析并生成图表，Excel导出功能无缝衔接传统工作流。
- 通用智能体：集成数十个MCP工具，生成图片、音乐、宣传片等多模态内容，满足新媒体创作需求。



OpenAI推出的agent评测任务SimpleQA上，Skywork也取得了94.5分

甲子智库解读

协作

交付

交互

天工Agent的使用感受：让每一个都应该学会如何和Agent一起进行工作。
天工Agent的行业启示：每一个Agent研发团队都应该让Agent学会和人一起工作

Deep Research: AI智能体（Agent）的特殊形态

Deep Research本质上是AI智能体（Agent）在知识探索领域的一种高度专业化的应用。

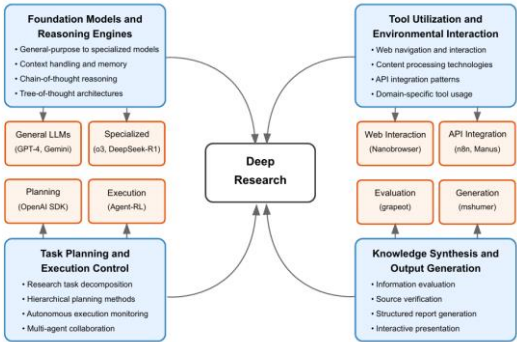
如果说一个通用的AI助手（如基础版ChatGPT）是一个可以回答各种问题的“通用机器人”，那么Deep Research系统就是一个专注于科研任务的“高级专家机器人”。它完美地体现了智能体的核心能力：

- 感知：通过高级检索，感知并“阅读”全网的论文、报告和数据。
- 规划：将一个模糊的研究课题（“分析市场趋势”）分解为一系列具体、可执行的步骤。
- 行动：主动使用工具（如浏览网页、调用API、分析数据）来执行这些步骤，最终生成一份结构化的研究报告。

《A Comprehensive Survey of Deep Research: Systems, Methodologies, and Applications》创新性地提出了一个四维技术框架，用于解构和分析Deep Research系统的核心能力。便于更为系统性地看待Deep Research的特点和使用方式。

- 维度一：基础模型与推理引擎 (Foundation Models & Reasoning Engines), 从通用LLMs（如GPT-4）向专为研究任务优化的模型（如OpenAI的o3、Google的Gemini 2.5 Pro）演进。这些专用模型在多步推理、知识整合和复杂指令遵循方面表现更优。
- 维度二：工具使用与环境交互 (Tool Utilization & Environmental Interaction), 从简单的API调用向复杂的环境交互与工具链编排演进。
- 维度三：任务规划与执行控制 (Task Planning & Execution Control), 从线性任务流向动态、自适应的执行控制演进。
- 维度四：知识整合与输出生成 (Knowledge Synthesis & Output Generation), 从简单的文本总结向可验证、结构化的知识产出演进。

架构模式	特点	优势	劣势
单体式架构 (Monolithic)	所有功能模块紧密耦合，围绕一个核心推理引擎。控制流集中，状态统一管理。	实现相对简单，能保证高度的推理一致性。	扩展性差，难以并行处理复杂任务。
流水线式架构 (Pipeline-Based)	将研究工作流分解为一系列独立的、顺序连接的处理阶段（如查询处理->信息检索->内容分析->输出）。	模块化程度高，易于定制和复用组件。	难以处理需要跨阶段反复迭代的复杂推理任务。
多智能体架构 (Multi-Agent)	由多个自主、专门化的智能体通过标准化的消息协议协同工作。	善于处理需要多种专业技能的复杂任务，易于并行化和扩展。	保持全局推理一致性和智能体间的有效协调是主要挑战。
混合式架构 (Hybrid)	结合上述多种架构的优点，例如用单体式核心处理复杂推理，用多智能体并行搜集信息。	灵活性和优化空间最大，能根据任务特性动态调整策略。	实现复杂度最高



多个产品将Research能力融入到Agent产品当中

Manus

Manus 作为一款智能体应用，具备一定的深度搜索功能，能够在搜索的基础上对信息进行整理和分析，为用户提供个性化的搜索结果和内容呈现。

核心亮点在于将传统的搜索引擎与个性化内容整理深度融合。它所宣传的“深度搜索”并非简单地罗列链接，而是试图理解用户的真实意图，通过对全网信息进行抓取、筛选和重组，最终生成一份高度定制化的“研究报告”。

智谱清言

智谱清言的“沉思”（rumination）功能，其背后的模型是智谱全新推出的 agent 大脑——沉思模型（glm-z1-rumination），通过强化学习，让模型学会自我批评、反思、甚至沉思，实现长程推理和任务执行。

该功能核心理念是“慢思考”，通过模拟人类的自我反思与批判过程，来提升复杂任务的解决质量。当用户提出一个深度问题时，“沉思”功能会启动一个多步骤的内部对话与推理链，主动进行假设、验证、反驳和修正，这个过程对用户是部分可见的，增加了AI决策的透明度。

夸克

2025 年 5 月 20 日，阿里旗下的夸克启动“深度研究”邀测，并开放了第一轮邀请码申领。其“深度研究”功能基于阿里自研的通义大模型，能够为用户提供更深入、更精准的搜索结果和知识解答。

其核心优势在于能够整合阿里生态内的海量数据和知识图谱，为用户提供兼具广度和深度的研究结果。该功能在处理商业分析、行业趋势、产品评测等与消费和产业紧密相关的话题时，有望展现出独特的优势。

月之暗面（KIMI）

2025 年 6 月 20 日晚，月之暗面正式推出其首款 agent 产品 kimi-researcher，并启动小范围灰度测试。该产品基于端到端自主强化学习技术训练的新一代 agent 模型，专为深度研究任务而设计，可帮助用户高效完成复杂研究，提供详实、可溯源的分析结论。

作为以长文本处理能力闻名的KIMI的“进阶版”，KIMI-Researcher的定位非常明确：**专攻“深度研究”这一硬核场景。其最大的技术亮点是基于端到端自主强化学习训练的新一代Agent模型。**使其在在任务规划、信息源的自主选择与评估、以及最终报告的逻辑链构建上，具备了更高的自主性和智能性。

附录

参考文献/链接

1. 朱啸虎：所有的AI应用都是套壳，创业公司不要浪费一分钱去训练底层模型https://www.sohu.com/a/877838534_115565
2. 红杉 AI 峰会闭门 6 小时，150 位创始人共识浮现：AI 不再卖工具，而是卖收益
<https://mp.weixin.qq.com/s/7atu0CxodoiCipRcppkgoQ>
3. AI教父辛顿最新访谈万字实录：AI将如何颠覆所有脑力劳动？<https://tech.ifeng.com/c/8kEyL0zGWjE>
4. “AI教父”辛顿最新专访：没有什么人类的能力是AI不能复制的 <https://mp.weixin.qq.com/s/N22lzubbcaboExjT7EG-DQ>
5. “AI教父”辛顿最新专访：AI已拥有情感和意识，还学会了欺骗 <https://news.qq.com/rain/a/20250530A09VIB00>
6. Andrej Karpathy on Software 3.0: Software in the Age of AI <https://www.latent.space/p/s3>
7. Andrej Karpathy YC 孵化器万字洞察：软件 3.0 时代创业机会，钢铁侠战衣是人与 agent 协作终极形态
<https://news.qq.com/rain/a/20250624A02PKZ00>
8. AI大神安德烈·卡帕西（Andrej Karpathy）6月18号在YC AI Startup School的主题演讲：软件的革命性变化
https://www.bilibili.com/video/BV1RsNsz5ENb?vd_source=de900c1f230ec4f63c0caa4d3ab03de5
9. 马斯克在YC AI School实录全文：硬核创业、第一性原理与星际征途
<https://cj.sina.com.cn/articles/view/2833534593/a8e44e81020019crm>
10. 两天、14 位大咖、2000 人：YC AI Startup School圈定AI创业3条跑道 <https://news.qq.com/rain/a/20250618A01LAE00>
11. OpenAI联合创始人Greg Brockman：从游戏AI到通用智能，我们的创业一路意外，ChatGPT模式都是不得已的选择
<https://maker.zhiding.cn/2025/0624/3168077.shtml>
12. 对话 OpenAI 联合创始人 Greg Brockman | 6月18日
https://www.bilibili.com/video/BV1bJNsZ5ENb?vd_source=de900c1f230ec4f63c0caa4d3ab03de5
13. 李飞飞：创办World Labs的初衷，就是想无所畏惧地解决空间智能问题 <https://news.qq.com/rain/a/20250615A03SHI00>
14. 李飞飞：在物理三维世界进行推理的人工智能系统
https://www.bilibili.com/video/BV1GhT4zaEJW?vd_source=de900c1f230ec4f63c0caa4d3ab03de5
15. NVIDIA CEO 黄仁勋主题演讲 | GTC 巴黎 2025 <https://www.bilibili.com/video/BV1aDMFzoESy/>
16. 黄仁勋最新警告：如果美国不参与，华为将“覆盖中国” <https://news.qq.com/rain/a/20250613A038XN00>
17. 20250612 | CNBC对英伟达CEO黄仁勋的完整专访
https://www.bilibili.com/video/BV1CCMMzeE1m?vd_source=de900c1f230ec4f63c0caa4d3ab03de5
18. 黄仁勋在英伟达财报公布后接受了Bloomberg和CNBC的专访 <https://weibo.com/1233486457/Pu2LgbRqz>
19. Demis Hassabis On The Future of Work in the Age of AI <https://www.wired.com/video/watch/big-interview-demis-hassabis-on-the-future-of-work-in-the-age-of-ai>
20. Demis Hassabis 畅谈人工智能时代的未来工作
https://www.bilibili.com/video/BV1hvTLzFE76?vd_source=de900c1f230ec4f63c0caa4d3ab03de5
21. 谷歌CEO皮查伊两小时访谈：AI是人类所见过最深远的技术，意义将超越火与电，因为它可以自我迭代
<https://news.qq.com/rain/a/20250609A07RRB00>
22. Sundar Pichai: Google Beam、Google XR、Gemini、AI 末日论的讨论
https://www.bilibili.com/video/BV1f2TxzjEHp?vd_source=de900c1f230ec4f63c0caa4d3ab03de5
23. https://www.linkedin.com/posts/yann-lecun_how-to-attract-scientists-to-your-country-activity-7334694439308697600-V-yp
24. 真正的AGI需达人类大脑理论极限 | DeepMind CEO最新对话实录 <https://blog.csdn.net/cf2SudS8x8F0v/article/details/148321417>
25. DeepMind CEO哈萨比斯：AGI将在2030后到来，年轻人要更重视怎么学习而非学什么专业，梵高画作震撼人心是因为他的人生
<https://news.qq.com/rain/a/20250530A05DKP00>
26. 专访DeepMind CEO 德米斯·哈萨比斯谈人工智能如何重塑谷歌 <https://www.bilibili.com/video/BV1AnjizYEq7/>
27. Anthropic CEO预警“AI致20%失业率”，黄仁勋反对该观点 <https://news.qq.com/rain/a/20250617A04N3700>
28. Anthropic CEO：AI 或致千万白领失业，呼吁对AI公司征税 <https://news.qq.com/rain/a/20250601A05K9N00>
29. Anthropic CEO警告AI或导致全美半数入门级白领工作消失，失业率暴升至20%
https://www.bilibili.com/video/BV1V27czQEX3?vd_source=de900c1f230ec4f63c0caa4d3ab03de5
30. 【Agent智能体】吴恩达：AI智能体发展现状 | LangChain访谈
https://www.bilibili.com/video/BV1T4jyzHERC/?share_source=copy_web
31. 吴恩达LangChain对话：别纠结Agent定义，成功的智能体往往从线性工作流开始，Vibe Coding这个概念充满误导
<https://mp.weixin.qq.com/s/QfozXP29vEjpSRPLXfDMQq>
32. 微软 CEO吹爆「智能体」：AI不仅要“杀死”SaaS，操作系统也会“格式化”重来 <https://news.qq.com/rain/a/20250528A014K300>
33. 微软 CEO Satya：“传统的软件应用层正在瓦解，将逐渐被AI智能体（Agent）所取代”
https://www.bilibili.com/video/BV1JzJQziEcb?vd_source=de900c1f230ec4f63c0caa4d3ab03de5

以上资料排名不分先后