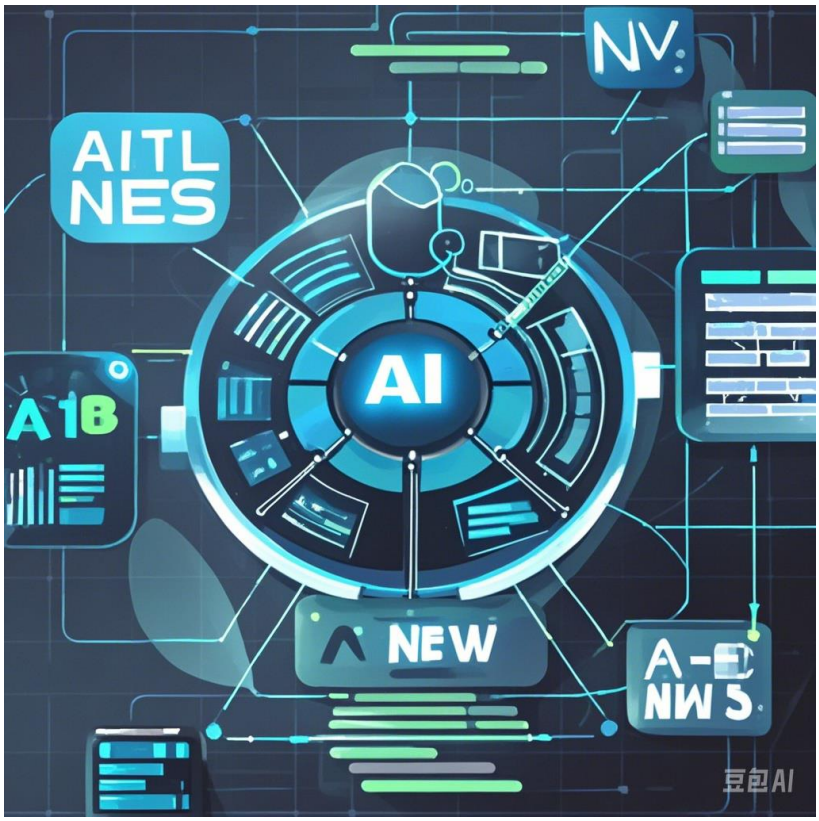
AI 智能体：下一代 GenAI 的前沿

——如何利用 AI 智能体开启新的收入流并实现生产力的飞跃

在人工智能（AI）的世界里，AI 智能体（agentic AI）正成为下一代 GenAI 的前沿。这一领域的潜力已经被众多软件供应商所强调，他们认为智能体 AI 能够开启全新的收入流——这是投资者们急切期待的发展。此外，智能体 AI 也可能是实现 GenAI 所承诺的重大生产力提升的关键一步，尽管这一承诺至今在很大程度上尚未兑现。本文我们将介绍 AI 智能体的核心概念和关键发展领域，以及我们如何将 AI 的使用从被动的信息处理转变为主动的问题解决。

智能体 AI 在企业中的热潮已经开始兴起，因为这些新方法开始成形。模型正在迅速进步，框架、技术和服务的发展已经导致了早期商业化产品的出现。聊天机器人和虚拟助手在我们的调查基础中被一致认为是 GenAI 的顶级用例。然而，随着这些实体变得更加复杂和自主，隐藏的复杂性、治理挑战和额外的开发工作也随之而来。这些问题必须主动解决，否则，这一趋势可能只会延续围绕 GenAI 在短期至中期内出现的幻灭感。



从长远来看，智能体向类人贡献价值的进步也将需要我们对获取智能体实体的思考方式进行转变。目前，运营成本仍然是采用的一个重大障碍，因为更复杂的工作负载增加了模型的负担，以及支持它们的基础设施。可能会出现一种新的模式，它将从传统的消费或订阅模式转变为 ROI 的函数。我们预测这将成为一个巨大的支出类别，因为对价值的重新聚焦，加上随时间推移的运营成本降低，将使支出合理化成为可能。

模块化 AI 系统的兴起

围绕 GenAI 的初始兴奋主要集中在基础模型本身。这些先进的模型产生了高质量的创意内容，并展示了令人印象深刻的理解力，每一次新迭代都在记忆、速度和最关键的推理方面显示出快速进步。然而，随着企业深入构建和扩展 GenAI 应用，许多企业受到它们实际限制的阻碍。

从 LLMs 到智能体 AI 的道路

数据限制

大型语言模型和基础模型受到它们所训练数据的限制，或者直接访问的限制。这些模型通常缺乏直接访问实时信息，这可能阻碍了及时或准确的响应。此外，LLMs 是无状态的，意味着它们不保留关于过去互动的信息，或者对话的特定上下文。

推理和问题解决

LLMs 有时会犯逻辑错误或跳到错误的结论，特别是在处理复杂推理问题时。它们经常在复杂的数学计算或需要对数学概念的深入理解问题上挣扎，除了基本算术之外。

准确性

由于它们的生成性，LLMs 经常产生不准确或误导性信息，被称为“幻觉”。这些可能发生在模型试图填补其知识空白，或者提供听起来合理但实际上不正确的响应时。

效率

基础模型非常庞大且通用性。它们经常缺乏深入的领域特定知识，使得它们在较小、更专业的任务上计算密集（且昂贵）。

业界已经出现了解决这些限制的开发技术。微调仍然是最受欢迎的方法（在我们的 AI 和机器学习、基础设施 2024 年调查中被 60%的组织引用），可能由于其早期突出和优化模型以满足特定需求的直观吸引力。另一种快速增长的方法是检索增强生成（RAG），它涉及将大型语言模型与检索系统（即向量数据库）结合，可以搜索并返回相关、事实和组织特定的信息。这两种方法，连同特定的 AI 防护措施（规则和安全约束），提高了生成内容的准确性、相关性和道德使用，这对于指导更复杂工作流程的内容变得越来越关键。

然而，从单体模型——所有功能都包含在单个大型模型中，如 LLM——转向更模块化 AI 系统的趋势更加重要。单体方法虽然强大，但在可扩展性、性能和效率方面提出了挑战，以及对核心模型本身的复杂性过度依赖。模块化 AI 系统，另一方面，将复杂的 AI 任务分解为更小的组件，可以通过多个交互工具（如其他模型、检索器、数据源）独立处理。这使得这些系统能够更有效地处理更复杂的交互，并促进更复杂的“思考”过程，从而提高输出质量。

早期研究表明，通过查询或请求提供结构来帮助 LLM“思考”也可以显著提高 LLM 性能，即使与大得多的模型相比。进入智能体工作流程——模块化 AI 系统的关键范式转变之一。

进入智能体

智能体是由 LLM 驱动的决策引擎。它被定义为能够摄取信息、推理、计划、行动，并最终记住并从它们随时间的行为中学习的能力。在它们的核心，这些智能体通常有一个 LLM，它被训练来协调不同 AI 组件或服务之间的交互。可以通过微调 LLM 在相关数据上，并向智能体提供特定工具（计算器、搜索引擎、其他模型）和资源（数据库、电子邮件、指令）。在它们最先进的形式中，智能体就像一个数字知识工作者，能够自主执行工作流程、完成任务和管理其他智能体或过程。

今天的智能体工作流程仍然具有预定义的元素，或约束，工作流程和控制主要手动规划，这大约在成熟度曲线的 2 级（见下图）。在这个阶段，这些实体没有独立行动的能力，或者做出证明“智能体”一词的选择。移动到更高的水平将显著需要 LLM 自身的进步。推理能力对于准确分解任务和计划至关重要，这是智能体今天的局限性。

智能体框架

技术供应商和研究小组的生态系统正在积极开发构建和部署 AI 智能体的基础结构和架构——智能体框架。许多都锚定在两种主要架构之一：reACT 智能体或功能调用智能体。选择取决于应用程序的性质和所需的灵活性、效率和可靠性的平衡。混合智能体可以结合两者的优势。

reACT 智能体。根植于 reACT 逻辑，这种方法归结为一个强大的提示，它已经被设计成将 LLMs 转变为结构化推理引擎。LLM 被引导接收一个查询，推理它，选择一个适用的工具，采取行动，并观察自己的响应，通常通过这个序列循环直到产生所需的输出。ReACT 智能体高度可定制化，对于更复杂的应用程序理想。然而，它们确实给组织工程团队带来了显著的开发负担。

功能调用。相比之下，功能调用智能体依赖 LLM 本身来扮演工具选择的角色。当面对查询时，模型确定哪些工具最相关，并从预定义的选项列表中选择适当的行动。许多商业可用的 LLM 支持工具调用，并已微调以确定选择哪些工具的逻辑。这些智能体虽然可定制性较低，但在许多场景中提供了更直接和有效的方法。

生态系统中的研究团队一直在开发更强大的商业和开源智能体框架，为不同的任务和用例优化。这可能涉及添加其他支持功能，如聊天机器人应用程序的对话记忆，或支持医疗诊断用例的并行功能。可能性是巨大的，新方法正在迅速出现，具有不同程度的专业化。

已经观察到，当为智能体编码了更狭窄的任务集和领域时，智能体变得更有表现。围绕多智能体架构的显著开发正在进行，其中多个智能体，每个都有专门的指令和工具基础，被调用来处理特定任务或任务的组成部分。一个简单的例子可能是一个层次化的多智能体设置，其中协调智能体可能提示编码智能体和批评智能体顺序开发和调试代码，产生更高质量的输出。

这解决了单智能体架构固有的挑战，如一般智能体的性能和效率限制。一些流行的多智能体框架的例子包括 OpenAI 的 Swarm 项目、CrewAI 和 Microsoft Corp.的 AutoGen。其他在智能体框架中发挥重要作用的包括 LangChain、LlamaIndex、Phidata 和基础模型公司。

值得一提的是，随着智能体架构的发展，对于强大的治理结构的需求也随之增长，以确保这些实体在规模上合规和安全使用。与 LLM 不同，LLM 主要处理和生成文本，AI 智能体可以与环境互动，做出决策，并可能采取行动。随着这些智能体变得更加自主和有能力，关于责任、偏见和潜在滥用的问题将出现。