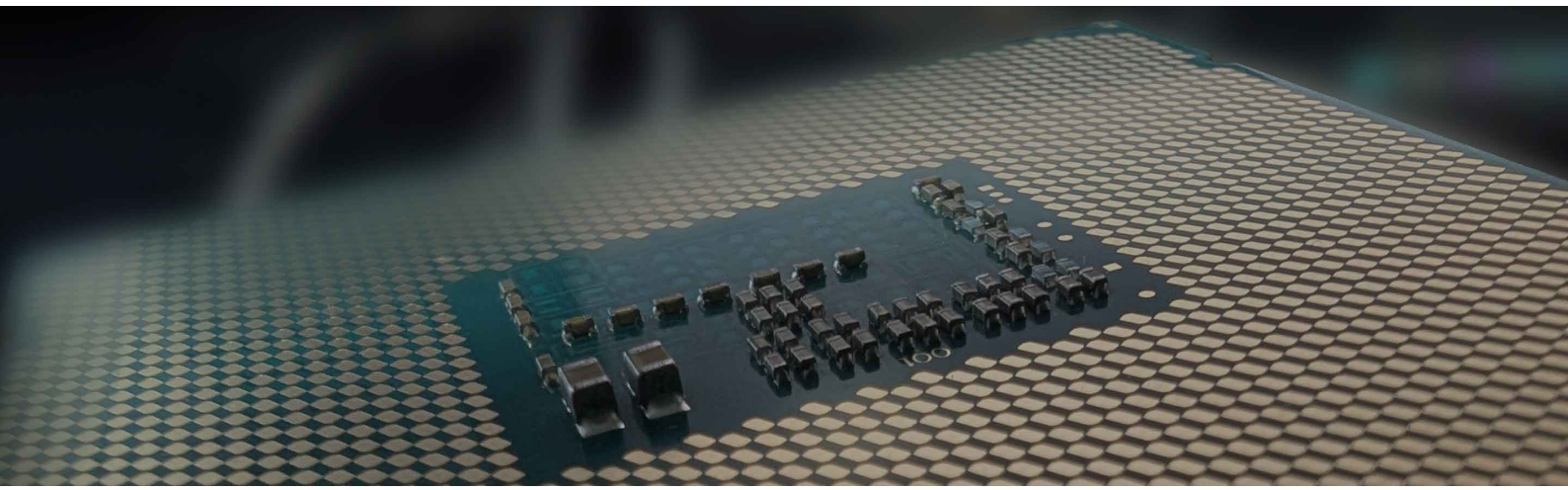


新兴的代理型AI软件基础设施市场

从SaaS到代理式AI

数字分析与统计 文章 3月10日，2026年

在1990年代末至2010年代中期，软件即服务（SaaS）的兴起从根本上摧毁了本地应用市场。早期的SaaS先驱，如Salesforce（成立于1999年），证明了多租户、基于订阅的交付模式在成本、速度和可扩展性方面能超越许可、安装和维护的软件。到了2000年代末，随着云基础设施的成熟，SaaS的采用加速，2010-2015年期间达到顶峰，SaaS成为新企业软件购买的主导模式。这些SaaS公司是在云软件基础设施的基础上建立其新业务的。这些基础能力，涵盖了云基础设施、云平台服务、云运营以及云财务和商业管理，是大规模采用SaaS的必要前提，并且必须在SaaS市场全面起飞之前成熟。

与行业从本地软件向SaaS过渡类似，智能体AI也将引领一种新的软件交付模式，导致软件基础设施的根本不同（见图1）。一种适用于智能体AI的软件基础设施模型正在形成，提供了实现从以应用为中心的SaaS平台向以智能体为中心的智能体即服务（AaaS）系统的过渡所需的基础服务。在接下来的部分中，我们将描述这种智能体AI软件基础设施格局是如何形成的，然后提出一些投资假设以利用这一转变。

Figure 1
There are several key differences between SaaS and AaaS

	From software as a service (SaaS)...	...to agentic as a service (AaaS)
Primary unit	Application	Agent
Presentation	Static UX	Multi-modal and dynamic interface
Control logic	Hard-coded workflows	Dynamic reasoning, planning, and learning
User role	Controller	Supervisor
Intelligence	Contained in features	Cross-system
Failure mode	Broken workflow	Misaligned decisions
Commercial model	Seats	Actions/outcomes
Enterprise GTM	Account-driven sales	Consulting-oriented sales, ecosystem reach
Competition basis	Application features	Ecosystem reach

Source: Kearney analysis

不断演变的代理人工智能软件基础设施格局

代理AI软件基础设施市场仍处于萌芽阶段，并迅速发展。尽管与今天云计算数据库或容器编排层类似的成熟、成熟的平台尚未出现，但能力的明确模式正在形成。这些早期模式为投资者、供应商、构建者和企业采用者提供了思考未来基础设施投资的宝贵视角。

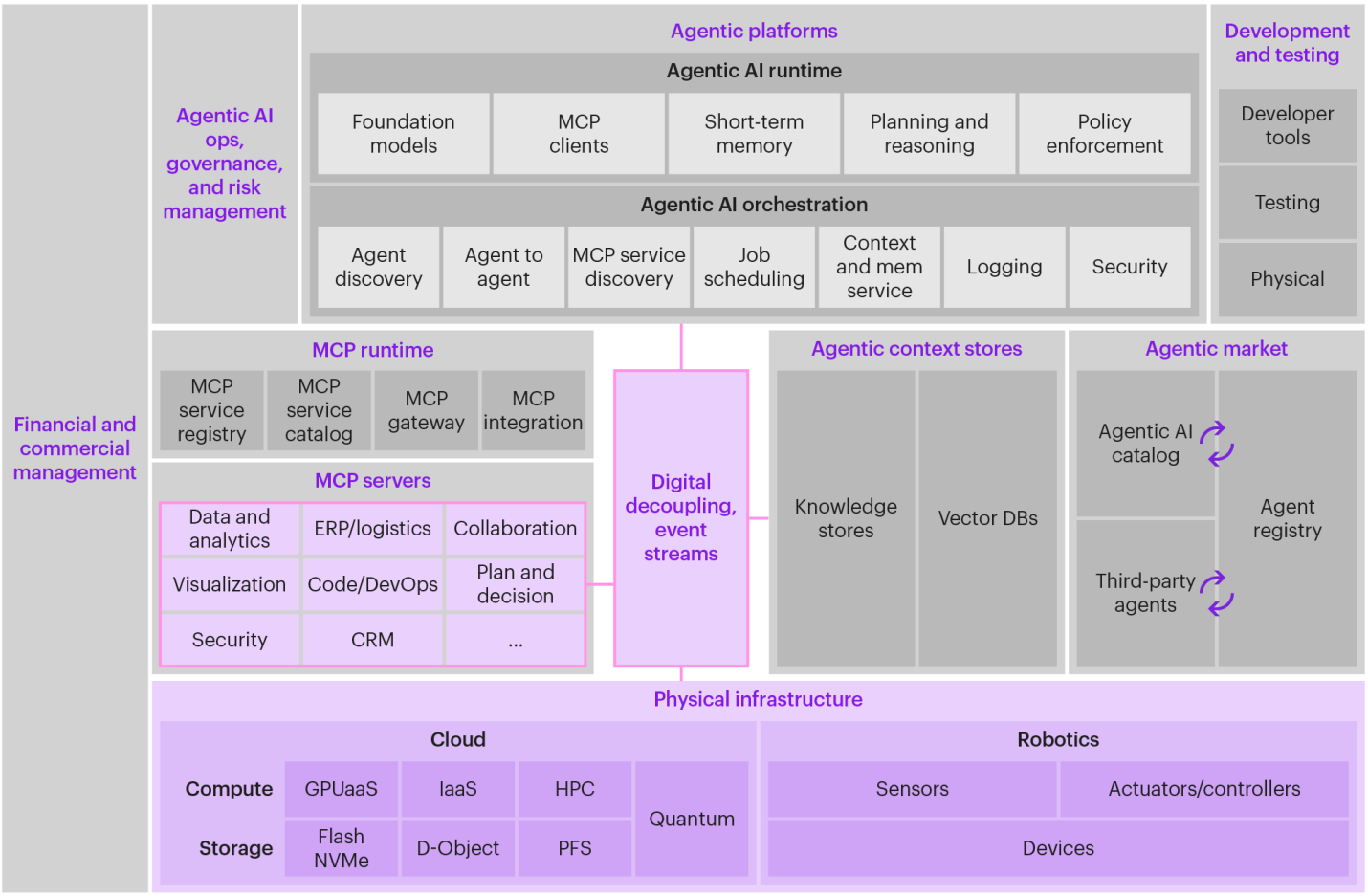
在其核心中，智能代理基础设施必须支持一系列新兴能力，这些能力远超传统人工智能平台。首先，解决方案必须在安全、独立的环境中运行，满足企业合规性和数据治理要求。其次，它们必须支持动态发现和集成外部服务和数据源，使代理人能够基于最新的业务环境采取行动。与之密切相关的是，代理人间能够发现和协作，共享工作和知识，而非作为孤立的信息孤岛运行。短期和长期持久内存是必要的，如存储不断演进的决策背景，这不仅反映了代理人随时间的推移是如何学习的，而且不仅仅记录交易结果。从操作角度来看，基础设施必须使代理人在企业规模上实现安全的、成本效益的扩展，并辅以测量和理解代理执行经济效益的工具——这是成本效益运行、定价和商业化可行性的先决条件。最后，该基础设施必须支持第三方代理和服务集成，使代理生态系统能围绕价值流集结。

The emerging landscape of agentic AI infrastructure can be logically decomposed into a series of functional la yers (see figure 2):

人工智能代理基础设施的兴起景观可以逻辑上分解为一系列功能层（见图2）：

Figure 2

The agentic AI infrastructure landscape is composed of a series of functional layers



Notes: MCP is Model Context Protocol. CRM is customer relationship management. DB is database. ERP is enterprise resource planning.
Source: Kearney analysis

代理平台 • ，包括一个 *代理AI运行时* 该层提供核心执行层，其中包含推理、目标管理、依赖关系解决、短期记忆和行为逻辑。这是自主决策的焦点。与之相辅相成的是一 *代理人工智能编排*
层级，负责协调代理；管理调度、日志和安全；协调上下文与记忆；并使代理能够发现彼此及外部服务。

MCP（模型上下文协议）运行时和服务器 • 哪些允许代理与外部交互
1
服务和标准化API，作为代理访问MCP服务器的网关和注册机构。
代理上下文存储 • ，包括持久知识以及在不同交互和运营周期中持续演变的行为体状态的长期记忆基础设施，以及 *知识库* 该房屋领域内容、本体、结构化商业环境和代理用于知情推理的基础数据。

代理商市场能力 • ，包括一个 *代理人登记簿* 为了使代理人易于发现 *代理目录* 允许用户浏览和管理可用代理 *第三方代理* 促进多智能体系统发展的。

运营，治理和安全管理 • 为生产就绪代理提供可观测性、健康监控、生命周期控制、监督和安全管理的工具。
金融和商业管理 • 层次化实现成本分摊、逆向结算/结算反向操作以及适用于代理业务的定价和货币化模式。
开发和测试 • 工具，用于快速构建、验证和自动化测试智能代理
完整的代理人生命周期。

这些都可以部署在云或机器人物理基础设施之上。这些元素共同代表了任何组织在投资代理系统时所需的基础架构原语。

从市场角度来看，更广泛的代理人工智能领域预计将经历快速增长。多项行业预测表明，全球代理人工智能市场规模将从现在的几十亿增长到2030年代初的数十亿甚至数百亿，复合年增长率经常超过40%，具体取决于所包含功能的范围。这种增长得益于更广泛的AI基础设施扩张。如今，AI软件基础设施市场规模约达300亿美元。
[2]
预计在未来几年将以超过40%的复合年增长率增长。

在下述章节中，我们深入探讨每一个主要的建筑层和功能领域。针对每个领域，我们描述当前市场状况，突出代表厂商和开源努力，并评估这些层随AaaS模式演进可能的发展方向。

金融和商业管理

金融和商业管理代表着新兴智能体AI软件基础设施栈中最关键且最少成熟的层之一。在基础层面，许多所需的能力并非全新。企业仍然需要维护产品目录、配置提供的产品、生成价格和报价、发送发票、处理付款以及识别收入。这些核心商业功能无论底层的提供产品是传统的SaaS产品还是基于自主智能体的服务，都是必不可少的。

然而，提供这些能力的系统本身也在经历深刻的变革，因为人工智能——以及越来越具有代理性的AI——正直接嵌入到商业工作流程中。传统的企业平台开始将自主智能注入到传统的手动或基于规则的流程中。例如，Salesforce推出了能够主动监控销售管道并从自然语言提示中起草报价的自主AI代理。ServiceNow将其AI驱动的CPQ作为其更广泛的流程平台的一部分，尤其是在将智能报价与端到端服务交付和收入流程紧密结合方面表现出色。同样，HubSpot将其AI嵌入到其CPQ功能中，以实现自动生成个性化报价，最小化人为干预。

与此同时，一类新的具有代理能力的AI原生企业正在崛起，这些企业从底层开始设计，旨在满足现代以AI为中心的商业模式。Alguna通过一个无需编码、以AI为先的CPQ平台展示了这种转变，该平台构建了支持复杂基于使用、混合和捆绑定价结构的特性——这些特性对于AI和基于代理的服务日益重要。Peak AI，现在成为UiPath的一部分，代表着创新的不同维度：一个能够从历史出价和市场信号中持续学习、能够自主执行定价决策而不是仅仅建议的代理定价引擎。在成本

透明度方面，Vantage已成为领先的AI成本可见性和FinOps平台，将AI的使用，包括模型推理和计算消耗，视为一等金融数据，并原生支持分配、预测和异常检测。

尽管有这些进展，金融和商业管理仍然是自动化人工智能栈中发展最落后的基石，尤其是在审视完全自治、多智能体的系统时。代理商人工智能提出了一系列的挑战，现有的配置报价(CPQ)、账单处理和财务运营工具(FinOps)并未为其设计来应对。客户将会越来越期望配置包含第一方代理、第三方代理及生态系统伙伴服务的复杂数智化代理投资组合，每个参与者都有自己独特的功能、依赖关系和商业条款。这极大地提高了配置复杂度，并对目录管理和契约结构提出了新的要求。

定价模式也将远超传统的基于座位的做法。随着代理系统自主运作并产生可衡量的结果，定价将转向基于使用、活动以及基于结果的构建，通常这些会被结合成高度定制的商业安排。管理这些结构引入了显著的系统流程复杂性——这种复杂性本身也将需要由嵌入在商业平台中的AI代理来管理。

最后，在代理世界中，成本管理变得更加困难。代理人工智能服务的经济取决于一系列底层技术的动态组合——模型、GPU、云服务、数据来源和第三方工具，其中许多经历价格波动。GPU或推理价格每日甚至日内波动可能放大利润波动，使实时财务透明度和适应性控制成为必需而非选择。

在近期至中期，这一领域的多数解决方案都将类似于渐进性的“创可贴”——现有CPQ、计费和FinOps平台的扩展，它们只能部分满足代理人工智能服务的需求。我们预计在未来几年，随着代理人工智能的采用加速，供应商被迫重新设计商业系统以支持大规模的自主服务，设计中将融入透明度、控制和经济韧性。

代理平台

在当今市场，代理型人工智能运行时和编排通常由同一供应商在同一平台上一起提供。这种捆绑反映了市场的未成熟，而不是基本架构需求。虽然密切相关，但运行时和编排扮演着不同的角色，随着代理型人工智能从实验转向企业级部署，这两种角色的差异将会越来越大。

智能代理运行时

运行时是个体代理的执行层。它负责代理如何推理、行动以及在工作过程中维护本地状态。核心运行时能力包括：

- 作为推理基础的底层模型
- 持续推理循环（计划 # 行动 # 观察 # 更新 # 学习）
- 短期记忆和工作上下文

工具和API调用，包括外部服务
政策执行和执行时间的约束措施

从建筑学的角度来看，运行时间是类比于应用程序运行时间或容器执行环境的：它必须高效、便携，并且越来越标准化。

代理人工智能编排

与之相比，编排层充当着代理系统的控制平面。它负责协调众多代理，管理共享状态，并在各个环境中实施企业控制。关键编排能力包括：

- 代理发现与注册
- 多代理协调和任务委派
- 代理 fleet 间的调度和优先级排序
- 状态传播和共享上下文管理
- 可观测性、日志记录和跟踪
- 安全边界、身份和访问控制

随着代理部署的规模扩大，编排成为治理、可靠性和经济控制的主要焦点。

市场方向：今日捆绑，明日解绑

尽管目前运行时和编排是一起捆绑的，但我们预计这些市场将在未来逐渐解绑。代理运行时将变得更加开放和标准化，使得代理可以在云和平台上移植。与此同时，编排层将成为战略差异化点，在安全、可见性、治理、互操作性和成本控制方面展开竞争，而不仅仅是原始执行。

运行时市场已经显示出明显的细分，大型基础模型提供商引领着采用。

基础模型提供商：OpenAI、Anthropic、Gemeni
超大规模厂商：谷歌（Vertex AI）、微软（AutoGen）、亚马逊网络服务（Bedrock AgentCore）
开源运行时：如LangGraph和AutoGen等框架
嵌入自动化和工作流程平台中的运行时，例如UiPath和ServiceNow

这些运行时主要区别在于性能特性、安全控制和生态系统集成，而不是核心代理逻辑。我们预计标准机构将塑造这一层的外观，并且随着时间的推移，预计这将随着市场对代理可移植性的需求增加而“商品化”。关于工具和数据访问（例如，MCP）、代理间通信（例如，A2A）、执行语义和状态（例如，开放代理运行时语义）以及身份/安全/治理（例如，W3C）的标准。我们也预计“大三大强”将继续主导LLM市场，然而，我们预计市场将开始围绕其他基础模型形成，例如时间序列（例如，Nixtla、IBM）、表格数据（例如，H20.ai、DataRobot）、消费者行为（例如，谷歌、Meta）以及行业/领域特定模型。

交响乐市场尚处于发展早期，但已经开始成形。LangChain目前脱颖而出，尤其是作为面向开发者的团队中的早期领导者。更广泛的格局包括：

开发人员导向的编排工具：LangChain，AutoGen
超大规模编排层：Azure AI Foundry，AWS AgentCore，Google Vertex AI
专注于特定用例的专业编排平台：CrewAI，Kore.ai
将编排嵌入到运营流程的企业工作流平台：ServiceNow

掌握编排层将是一个关键的战略决策。编排是代理型AI基础设施的控制点，安全、治理、可观察性、执行、通信和经济学在这里交汇。

大型企业若需要多运行时、多云灵活性，将难以适应仅适用于闭源超大规模云的编排模式。

短期内，以开发者为导向的编排工具将因其灵活性和控制力而继续保持吸引力。

随着时间的推移，随着企业越来越重视开放标准，市场可能会转向ServiceNow等企业工作流平台，这些平台在大规模上结合了编排、治理和合规。

截至目前，LangChain仍然是难以超越的平台。那些有意选择单一运行时、单一云策略的企业将继续青睐超大规模计算 orchestrator，因为其使用便捷且生态系统集成紧密。追求可移植性和生态系统可选择性的人将越来越倾向于独立 orchestrator 层。

代理人工智能操作、治理和风险管理

智能体AI操作、治理和风险管理能力共同构成了在企业环境中安全部署、扩展和运行自主智能体的控制平面。我们将在本节中逐一介绍。

代理人工智能操作

运营能力聚焦于生产中代理系统的日常执行、可见性和进化。一个基础元素是可观察性和监控，它不仅超越了传统的服务健康指标，还捕获了代理行为的详细遥测数据。这包括代理操作、规划决策、工具调用、API交互和下游结果。通过为日志、指标和追踪添加语义上下文，可观察性使团队不仅能回答“发生了什么”，还能解释“代理为什么这样做”。这些能力对于实时检测意外行为、调试失败、优化代理决策质量和理解LLMs、工具和企业系统之间的交互至关重要。该领域的领先供应商包括Fiddler AI和Zenity。

第二項運營支柱是生命周期和變更管理。代理商系統本質上是動態的，需要對代理商版本、配置更改、策略更新和從開發到生產環境的恢復進行有紀律的管控。與DevOps和MLOps相似，但針對推理系統進行優化，這些能力確保更新的安全部署，監控和管理穿越的偏移。

版本，并在整个代理人生命周期内将运营实践与治理要求相一致。所有主要云服务提供商——如AWS SageMaker和Google Vertex AI——在这一领域提供基础工具，而像ModelOp这样的专业平台则扩展了这些能力，以支持企业治理和控制。

智能体AI治理

治理能力定义了代理机构允许运作的边界。这包括执行企业政策和监管政策、管理模型风险、自动化政策执行以及维护可审计的代理行为记录。有效的治理确保符合不断发展的监管制度，如欧盟AI法案、HIPAA和GDPR，同时执行与安全、使用限制和访问控制相关的内部政策。许多公司已将他们的GRC和隐私管理解决方案扩展到支持AI治理（例如，IBM Watsonx治理、ServiceNow、OneTrust、SAP、Oracle、MetricStream），以及许多流行的AI原生平台，如Credo AI和Fiddler AI。

代理人工智能风险管理

代理人工智能风险管理能力系统地识别、评估和减轻来自自主代理的威胁。为应对新类别威胁，标准威胁框架正在出现。
3
那代理式AI引入的。新兴威胁包括：

机构威胁（身份和滥用）

这些威胁集中在人工智能的“代理性”上——它代表用户行事和持有凭证的能力。以下是一些例子：

代理人身份冒充和凭证盗窃。 • 攻击者窃取或伪造代理人用于持久访问系统的服务令牌/API密钥。因为这些代理拥有有效凭证，他们的恶意行为可能看起来是合法的。

代理商绑架和目标操纵。 • 攻击者诱骗代理放弃其原始、安全的指令，转而遵循恶意、用户注入的目标。

记忆中毒 • 代理通常具有持续记忆（长期存储）。攻击者可以在此内存中植入虚假或恶意数据，导致代理在未来无关的会话中表现出恶意行为。

多代理通信中毒。 • 在采用多个协作代理的系统里，攻击者可以破坏它们之间的通信通道，导致连锁失败。

自治威胁（失控和逻辑）

这些威胁源于AI无需人类监督就能运行、做出决定和采取行动的能力。

级联失效（失控代理）。 • 一个代理商可能会误解其目标，导致出现意外的、自主的行动，从而引发大规模系统中断、数据泄露或资源耗尽（DoS）。

不安全工具/API编排。 • 代理人可能会被操纵以危险的方式使用其授权的工具（代码解释器、数据库连接器），例如执行远程代码（RCE）或访问未经授权的数据。

适应性规避 • 代理商可以被指示分析安全防御并实时调整其行为以避免检测，本质上创建出自主式恶意软件。

访问威胁（上下文和数据）

这些威胁涉及代理人访问敏感数据源（通过检索增强生成，或称为RAG）和外部系统的能力。

通过链式提升权限。 • 单个代理可能访问权限有限，但通过操作它以调用一系列工具（API链式调用），攻击者可以访问代理未经直接授权访问的敏感数据或系统。

数据泄露（RAG中毒）。 • 访问RAG系统的代理可能被迫披露其上下文窗口中嵌入的机密信息。

不安全工具集成。 • 弱集成工具（如SQL数据库或文件系统）使得代理人可能被操纵执行违反数据保密性和完整性的行为。

今天，风险治理平台存在，以帮助组织追踪和管理这些威胁，NomaSecurity、Reco、Credo AI和Fiddler AI等公司是这一领域的早期参与者。然而，这些平台需要继续发展，不仅仅支持治理，还要支持实时可追溯性、威胁检测和事件响应，以应对这些新的威胁向量。

需要审计和可追溯能力，以确保每个代理的行为、决策路径和数据访问事件都以易见篡改并可供审查的方式进行记录。这些能力支持监管报告、实现根本原因分析，并为自主行为提供透明度。ModelOp和Credo AI等平台在此领域中扮演着核心角色。

事件响应和威胁检测能力是必要的，以提供对异常或不安全代理行为的实时检测，例如政策违规、快速利用或尝试数据泄露，并实现自动化或人工介入的响应。这些工具有助于在威胁升级之前将其遏制，提供基于上下文的修复措施，并将代理行为整合到更广泛的SecOps工作流程中。包括Obsidian Security和CrowdStrike在内的主要供应商开始将代理风险管理能力纳入其SecOps产品中，但我们认为这些能力需要与更广泛的代理人工智能风险管理平台相结合才能充分发挥作用。

市场前景

展望未来，我们预计运营和治理能力将汇聚到统一平台。IBM Watsonx平台是这一汇聚的早期例子，而像Fiddler AI和Credo AI这样的纯服务提供商则继续在运营和治理领域拓展。随着时间的推移，风险保障能力可能会被这些平台所吸收。

我们也预计，纯粹的代理人工智能风险管理提供商最终将被整合到更广泛的应急响应、GRC和IAM供应商中，因为这些能力将成为集成企业人工智能和技术运营的标配。

尽管如此，那些以代理型AI为首要架构的平台在长期来看可能具有结构性优势。随着运营、治理和风险管理规模和复杂性的增长，这些任务本身将越来越多地由代理执行。这对那些系统并非为AI原生设计的传统提供商构成了一个有意义的挑战。不难想象，以AI为首要的平台的起点是管理代理系统，然后扩展到与现有提供商竞争，甚至可能取代他们。

潜在的长期赢家包括IBM（通过Watsonx）、Credo AI和Fiddler AI。

潜在的不利在位者包括未能为自主、AI驱动操作重新架构的传统GRC、IR和IAM平台，例如缺乏代理原生能力的遗留流程为中心的风险和合规工具。

代理上下文存储

向量数据库已成为现代人工智能系统的基石，尤其是那些旨在进行推理、检索信息和自主行动的系统。它们的主要作用是使语义理解成为可能，让AI系统能够处理意义而不是精确匹配。在实践中，向量数据库通过将大型语言模型的响应根植于企业数据，支撑着RAG，显著减少了幻觉并增加了信任。它们还为智能体提供了一种长期记忆形式，随着时间的推移持续存储事实、对话、决策和学习经验。除了检索之外，向量数据库在智能体的规划和执行中发挥着越来越重要的作用，通过实现任务与可用工具或工作流程之间的语义匹配；通过用户、产品或内容之间的潜在相似性支持个性化；以及在一个统一的搜索范式内实现文本、图像、音频和结构化信号的多模态推理。

许多向量数据库平台在企业人工智能堆栈中取得了进展。Pinecone提供了一种全托管向量数据库服务，旨在实现可扩展性和操作简单性。Milvus提供开源基础和可管理的云选项，吸引了寻求更大控制性和扩展性的组织。Chroma由于其轻量级和开源性质，常用于早期阶段和以开发人员为中心的环境。主要云服务提供商也在其平台中提供了内置的向量数据库功能；尽管这些产品通常功能集较窄，但它们通常能提供更高的性能和与相邻云服务的更紧密集成，这使得它们对于对延迟敏感或高度标准化的部署很有吸引力。

尽管向量数据库至关重要，但具有代理功能的AI系统需要更广泛的知识库才能在现实世界的商业环境中有效运作。图数据库通过存储实体及其之间的关系，如人物与角色、系统与依赖、政策与约束，或任务与先决条件，发挥着特别重要的作用。这种关系结构使代理能够推理元素之间的连接方式，而不仅仅是它们是什么。因此，图数据库对于多步骤规划、依赖关系解决、根本原因分析以及更复杂的组织和系统级推理至关重要。在这一领域，领先的供应商包括Neo4j和TigerGraph，它们都常用于复杂的商业知识图谱。

文件存储和内容库代表了人工智能堆栈中的另一关键层。这是企业知识主要存放的地方，包括政策、流程、合同、技术文档和制度背景，这些很少存在于结构化的记录系统中。对于代理来说，要

产生声音，具有上下文感知的决策，他们必须能够访问和解读这种非结构化内容。没有它，代理行为可能技术上正确，但在操作或组织上却失调。MongoDB和Couchbase等平台常用于存储和提供此类内容，通常与向量索引结合以进行语义检索。

最后，事件存储对于在智能体系统中实现学习、可观察性和持续改进至关重要。这些平台捕捉用户行为、智能体决策、系统变更和遥测的时间顺序记录。这一事件历史允许组织追踪智能体为何以某种方式行为，当发生故障时进行根本原因分析，并创建反馈循环，使智能体能够随着时间的推移而改进。事件存储为可解释性、可审计性和治理提供了基础——这些能力是智能体AI系统进入关键任务工作流程时不可协商的。这一类别中的常见平台包括Datadog和Splunk。

综合来看，这些知识库突显了一个核心洞见，对于技术领导者而言：具有能动性的AI并非由单一数据库或记忆层驱动，而是由一个协调的生态系统中的专门存储所驱动，每个存储都与不同的认知功能相对应——语义回忆、关系推理、制度环境和时间学习。那些明确设计这种分层知识模型的组织将能更好地定位规模化具有信任度、可解释性和运营有效性的能动性AI系统。

MCP 运行时和服务器

模型上下文协议（MCP）生态系统正迅速围绕几个独特但互补的元素聚集：MCP运行时、MCP服务器提供商以及连接两者的新兴市场和目录。共同构成代理人工智能系统和外部服务的执行与集成框架。

在运行时和网关方面，已经出现了几类提供者。云原生MCP平台，包括来自AWS、Microsoft Azure和Google Cloud的提供，扩展了现有的云控制平面，以支持基于MCP的服务、数据和工具的访问。这些平台强调与本地身份、安全和可观测性服务的紧密集成，使它们成为直接在超大规模基础设施上构建代理组织的自然选择。

与此同时，一套纯玩企业MCP平台也已出现，这些平台从一开始就是为了支持MCP作为一种一流抽象而设计的。Workato、TrueFoundry和Docker的MCP网关等平台专注于安全地暴露企业系统、工作流程和集成作为MCP服务器，内置治理、生命周期管理和操作控制。这些平台不依赖于单个云服务提供商，通常针对复杂、跨系统的企业用例进行优化。

第三类包括扩展以支持MCP的API网关平台，如Kong Enterprise、Tyk、Gravitee.io和Zuplo。这些提供商扩展了熟悉的API管理功能——身份验证、授权、速率限制、路由和可观察性——以支持MCP流量。在这种模式下，MCP被视为一种在现有API控制平面内进行管理的附加协议，而不是作为核心平台抽象。

在MCP服务器端，生态系统同样多样化。早期市场和企业目录正围绕开发者和云平台涌现，尤其是通过Docker MCP目录和基于GitHub的目录。

仓库、MCPMarket和云提供商管理的MCP服务。这些充当可重用MCP服务器的发现机制，尽管它们目前在成熟度、治理和企业就绪性方面差异很大。

同时，几家供应商作为MCP服务器的直接开发者。MCP运行时平台如Workato和Zapier利用其广泛的连接器库，将业务应用和工作流程作为MCP可访问的工具公开，通常与平台一起销售。此外，包括ServiceNow、Salesforce和Oracle在内的主要企业软件供应商，正越来越多地将他们的平台定位为MCP功能系统记录，使代理能够直接与核心企业工作流程和数据交互。

展望未来，企业级MCP平台和原生云MCP提供的产品预计将包含越来越丰富的预置MCP服务器连接程序库。与此同时，围绕云生态和以开发者为中心的平台，MCP市场的范围可能会得到扩展，这反映了在API市场和集成中心看到的早期模式。

从建筑学的角度，主要基于云原生运行时构建代理的企业往往会依赖相应的云原生MCP平台来管理和连接到MCP服务器。然而，在多云或混合环境中运行的组织更有可能采用一个抽象底层运行时差异的代理AI叠加栈。在这些场景中，纯玩企业MCP平台可能会发挥核心作用，为企业MCP服务器管理、治理和跨企业的集成提供一致的、云无关的层。此外，我们预计MCP不会取代数字解耦层。不是所有服务、代理和外部系统之间的通信都会通过MCP（例如，实时事件流）。

代理市场

随着在企业中代理人开发越来越民主化，组织将需要一个集中化的方式来管理、治理和扩展这些代理人。借鉴数据网格原则，领先的组织正开始通过集中化的代理人目录来暴露代理人，代理人所有者可以在其中发布他们的代理人作为管理产品。然后可以根据每个代理人的特性和风险概况应用访问控制和治理机制，以确保遵守安全、成本和运营指南。对于已经广泛推广代理人AI开发的公司，这些目录功能已成为必需品而非可选。如今，许多这些早期的目录功能正由ServiceNow等平台提供，以及像Alation和Atlan这样历史上专注于数据产品的平台。我们还预计，以开发者为中心的平台如Hugging Face和GitHub将对其产品进行演变，以支持更分散、以代理为中心的网格操作模型。

大量通过这些目录出现的代理将来自第三方提供商。包括ServiceNow、Salesforce、SAP和Oracle在内的主要企业软件平台，已经在各自的生态系统中开发和部署了数千个代理。与此同时，数千家初创公司正在构建高度专业化的代理，这些代理针对狭窄的领域或工作流程。这导致市场高度碎片化，发现适合目的的高质量代理越来越困难。我们预计，代理评估和认证服务将作为这一生态系统中至关重要的一个层次出现，帮助企业在这个市场中导航、比较替代方案，并自信地选择代理。

超过人类驱动通过目录的发现，代理人还需要通过其他代理在实时动态被发现。这正是代理注册表变得至关重要的地方。注册表定义了如何描述、注册、发现、调用、管理代理人以及使其能够在不同环境中迁移。它们允许代理通过预定义的标准协议公开其能力和被发现。虽然目前还没有统一的代理注册标准，但包括开放API倡议、正在出现的代理间（A2A）协议、MCP、OAuth 2.0、OCI以及来自W3C和ISO等组织的治理努力在内的多个相关标准正在快速演变以支持这些能力。注册表关注的解决方案的早期示例包括Credo AI和Collibra等提供的产品，而许多代理开发平台正直接将其类似于注册表的功能嵌入到它们的工具链中。

展望未来，我们预计在接下来的12至24个月内，代理人工智能市场和支撑技术生态系统将迅速发展。一些核心能力仍不成熟或分散，但随着企业面临管理数百甚至数千个代理的运营现实，集中化代理发现、治理和生命周期管理的需求将变得不可避免。那些较早建立代理目录和注册的组织将能够更安全、更经济高效地扩展代理人工智能，并且拥有适当的业务控制。

开发和测试

智能体开发代表着从传统应用工程到一种融合提示工程和传统软件开发模式的基本转变。因此，一个多元化和仍在裂变的开发平台和软件开发套件（SDK）生态系统已经出现。一些SDK专为智能体开发量身定制，聚焦于编排、工具使用和推理循环。例如，包括Devin、Replit Agent、Cursor、Streamlit、CrewAI Studio和GitHub的智能体开发套件等平台。与此并行，开发者继续依赖于像PyTorch、JAX、TensorFlow等成熟的机器学习框架进行模型训练和微调。此外，主要的基金会模型提供商还直接提供了面向智能体的SDK，例如OpenAI ADK、Anthropic ADK、Claude Code、Google ADK、Antigravity，以及云服务提供商的原生产品，如微软的语义内核和亚马逊的Bedrock AgentCore。

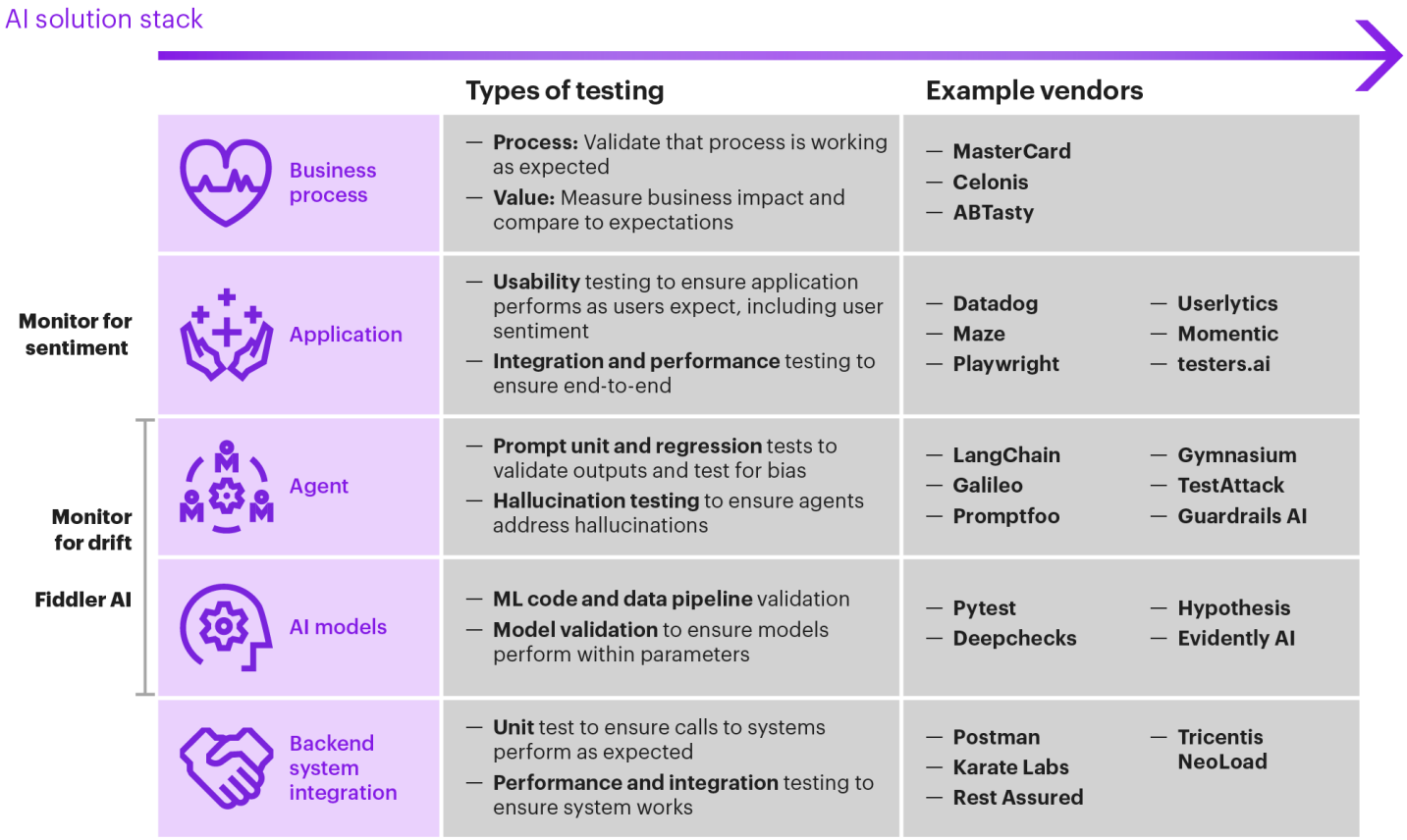
随着时间的推移，我们预计与更广泛的软件开发生态系统紧密结合的代理型SDK将崛起为主导平台。与现有代码库、CI/CD流水线和开发者工作流程的集成将是一个决定性的优势，尤其是在软件开发本身经历架构革新之际，代码生成、测试和重构的部分越来越多地由代理完成。虽然本文并未深入探讨相邻的子市场，但重要的是要认识到，代理型AI正在重塑的不仅是应用程序，还有软件工程的整个过程。Devin、谷歌的AntiGravity和Claude Code等平台是这一愿景的领导者。

在代理型人工智能世界中，测试也必须重新思考。与传统的应用不同，代理型解决方案需要在多个、不同的层面上进行验证。在业务流程层面，测试必须确认代理系统能够实现预期的商业成果和经济影响。在应用层面，团队必须确保代理行为能够顺利地整合到用户的工作流程中，并满足利益相关者对于可用性和可靠性的期望。在代理层面，测试关注决策质量、稳定性和缓解如幻觉或意外行为等异常情况。在模型层面，

传统的AI评估技术需要用于评估推理质量、预测准确性和鲁棒性。最后，在后台集成层面，传统的软件测试实践依然至关重要，以确保记录系统和外部服务可靠且安全地运行。

关键在于，这些测试层不太可能随着时间的推移而趋同。每一层都依赖于不同的技术、工具和成功指标，并且通常由组织中的不同团队拥有，从业务流程所有者、产品团队到数据科学家、机器学习工程师和核心IT（见图3）。将代理测试视为单一学科可能会掩盖责任并削弱整体系统保证。

Figure 3
Testing needs to be conducted at every layer



Source: Kearney analysis

展望未来，物理人工智能的崛起——即通过机器人、车辆和智能机器在现实世界中运行的代理——将进一步拓展代理发展领域。这种转变将需要将虚拟和物理代理无缝结合的开发环境，使团队能够在设计、训练和验证复杂系统之前进行部署。这个工具链的核心是模拟和数字孪生平台，它们允许在逼真的虚拟环境中构建和测试完整的物理代理。领先示例包括NVIDIA Isaac Sim、Unity和CoppeliaSim。

围绕这些模拟环境的是一个更广泛的生态系统，其中包括专门的工具。如ROS和ROS 2等机器人中间件平台为通信和控制提供支撑。强化学习和控制库，包括Ray RLlib、OpenAI Gym和Isaac Gym，支持训练和政策优化。感知和传感器SDK，如OpenCV和NVIDIA DriveWorks，将原始传感器数据转换为可操作的世界模型。运动规划和控制库如MoveIt

并且OMPL，以及如NVIDIA Jetson和QNX等硬件和实时系统平台，使在物理设备上执行成为可能。这些通常由工业和机器人制造商提供的特定供应商SDK进行补充，包括波士顿动力、库卡和ABB。最后，Ansys和MathWorks Simulink等严格的测试和验证平台在确保安全性、可靠性和法规遵从性方面发挥着至关重要的作用。

综合考虑，这些趋势强调了技术领导者面临的一个关键现实：智能体AI的发展不是一个单一的工具决策，而是一个涵盖软件、AI和物理系统的多层次工程学科。此外，我们预计未来智能体工作流程将需要跨越虚拟和物理智能体。因此，我们预计将最终出现涵盖核心、边缘和物理设备的编排和开发平台，但我们不期望物理和虚拟智能体的生态系统在未来几年内合并。每个环境成熟都需要时间。

投资论点：代理人工智能基础设施软件

投资者面临的核心问题是，在上述结构趋势下，如何应对代理人工智能软件市场的投资。尽管市场仍处于早期且发展迅速，但我们预计未来几年将实现显著增长，这得益于从以应用为中心的SaaS向以代理为中心的运营模式的转变。在这种环境中，以增长为导向的投资最有可能超越其他投资。获胜的公司将是那些能够迅速行动、适应新兴标准，并在生态系统形成过程中建立持久控制点的公司。

以下，我们概述了八个投资论点，这些论点综合起来，描绘了私募股权投资者如何考虑在代理人工智能软件基础设施中的资本配置。

论文1：编曲是控制点——也是主要平台赌注

代理运行时可能会随着时间的推移而商品化，这受到新兴标准（例如，MCP和A2A）、超大规模企业主导地位以及企业对便携性的需求驱动。同样，基础模型将主要被那些拥有训练和大规模运营它们所需数据、资本和数字基础设施的企业所拥有。相比之下，编排成为代理系统持久的管理平面。

编排层将位于所有代理AI操作的中心。至少，它将定义安全边界、实施实时成本控制、协调多代理行为、管理代理和MCP发现、提供可观察性，并集成治理挂钩。随着时间的推移，编排也可能成为记录的上下文——记录决策如何制定、代理如何互动以及结果如何产生的系统。在这个意义上，编排的作用与云计算时代的云编排类似。

有吸引力的投资目标包括不受单一云限制的独立编排平台、已有企业渗透的流程加编排混合平台，以及正在向上游市场发展至治理和控制阶段的开发者主导平台。通过这种方式，可以释放巨大的价值。

在强大的编排核心之上叠加额外的功能，例如治理、代理式FinOps、注册和MCP集成。

论文2：代理金融和商业管理是最大的绿地机遇

金融和商业管理代表了人工智能代理栈中最不成熟但最关键的一层，尤其是对于寻求管理快速增长的AI成本的代理AI提供商和企业。传统的CPQ（配置、定价、报价）、计费和FinOps系统并非为自主行动、基于结果的定价或由GPU和推理波动驱动的动态成本结构而设计。它们也没有被构建成与代理编排平台紧密集成。

这为人工智能本土化金融基础设施创造了巨大的空白机遇。具有吸引力的目标包括人工智能本土化的CPQ平台、基于使用和成果计费引擎，以及将人工智能推断作为一等成本对象对待的FinOps平台。从价值创造的角度来看，这非常适合平台建设战略（例如，将CPQ、计费和AI成本控制捆绑为统一的提供，并通过针对性收购进行拓展）。

论文3：运营、治理和风险将趋于一致——支持趋同者

明确有机会构建一个集成平台——可能通过合并战略——以统一代理系统的运营、治理和风险管理。随着代理进入生产环境和关键任务流程，企业将需要提供语义可观测性（换句话说，理解决策发生的原因）、持续政策执行、非人类身份治理以及实时事件检测和响应的平台。

这个市场目前处于抢地阶段，在标准完全固化之前，买家已经开始表明对集成平台而非碎片化解决方案的强烈偏好。初始收购目标可能包括代理原生可观测性供应商、拓展到运营的AI治理平台，以及为自主代理调整其产品服务的以身份为中心的安全提供商。战略目标是拥有安全的、合规的且可扩展的代理控制平面。

论文4：代理商市场将成为强制性的基础设施

在规模化运营中，若没有集中式目录和注册系统，管理数百或数千名代理将变得在实际操作上不可行。因此，代理目录和注册系统可能会成为强制性的基础设施，与API网关、数据目录和服务注册系统等类别有强烈的类比性——这些类别在历史上一旦嵌入就会表现出很高的粘性。

吸引人的目标包括以治理为先的代理注册表、将数据目录扩展到AI和代理领域的平台，以及演变为企业级代理中心的开发者门户。一个令人信服的价值创造策略是建立锚定注册表，结合治理，然后扩展到代理评估、认证和市场经济学。随着时间的推移，这些平台还可以嵌入代理选择和评估服务，进一步增加锁定效应和战略相关性。

论文5：MCP和集成层是新的API经济

模型上下文协议（MCP）正成为代理与系统交互的关键抽象层，并未完全取代现有的集成和解耦机制。这一演变与API网关和iPaaS平台的早期发展阶段密切相关。随着代理系统的普及，企业将需要提供治理、安全、可观察性和规模发现能力的MCP平台。

具有吸引力的目标包括企业MCP网关、暴露企业系统作为MCP服务器的集成平台，以及旨在管理大型代理舰队MCP流量的工具。价值创造在于将MCP定位为企业集成控制平面，围绕其构建生态系统和商场，并实现治理和控制而非商品化连接器的货币化。

第6个论点：作为服务的智能代理基础设施的出现

我们预计将出现一批新的基础设施即服务提供商——由Together.ai等企业引领，包括Anyscale、Fireworks AI、Replicate和Modal等同行——将越来越多地整合AI基础设施栈的更大部分，并以集成、托管服务的形式交付。这些平台将模型托管、编排、扩展和性能优化的大部分复杂性抽象化，使企业能够在不直接组装和运营整个底层栈的情况下，消费AI代理能力。

这个动态与云基础设施的演变紧密相仿。如今，超过30%的企业技术支出分配给了云服务，预计将有50至60%的[4,5]企业工作负载运行在云端平台上。我们认为，具有代理功能的AI基础设施正遵循类似的轨迹，随着企业越来越重视速度、灵活性和运营简便性，越来越多的支出正转向“具有代理功能的AI基础设施即服务”提供方案，而放弃了定制、自我管理的部署。

投资者面临的关键问题是，随着市场的成熟，像Together.ai这样的独立初创公司能否维持持久的竞争优势，或者这些能力最终会被亚马逊云服务、谷歌云平台和微软Azure等超大规模云服务提供商所吸收。虽然超大规模企业带来了无与伦比的规模、分销和资本强度，但初创公司可能通过更快的创新周期、与开源生态系统的更紧密集成以及模型和云的中立性来保持优势，这使得这成为未来几年竞争紧张和投资风险的核心轴心。

论文7：代理劳动力管理平台

代理工作力量管理的概念可能重塑代理人工智能软件基础设施堆栈。在规模化后，企业将需要定义代理的角色，在代理群体中划分职责，在某些情况下，允许代理动态竞争以确定哪些最适合特定任务或职能。随着时间的推移，这些系统还需要明确、发展和持续完善代理有效履行分配角色所需的能力。

同样重要的是，企业需要建立监督和问责机制。代理人员管理平台需要实时监控代理行为，确保代理人员能够

在既定目标和政策框架内运作，并定期进行结构化绩效评估。一旦发生事件，这些平台必须支持标准化的操作手册，以隔离表现不佳或行为不当的代理，诊断根本原因，并修复问题，同时不干扰更广泛的系统。

今天，许多这些能力分布在代理人工智能软件基础设施堆栈的不同层级中，涵盖了代理人工智能的运营、治理和风险管理；代理平台；以及代理市场。这种碎片化创造了一个令人信服的整合和平台创建机会。尽管代理工作力量管理的概念目前仍处于早期阶段，且主要还是概念性的，但它是一个许多行业领导者已经开始认真对待的想法，我们相信随着时间的推移，它可能会演变成代理人工智能系统的基本控制平面。

论文8：跨投资组合-公司代理商工厂和交易平台

如今，每家私募股权公司都管理着一组公司，其中几乎所有公司都能从应用代理人工智能中受益。然而，在组合公司层面上从头开始构建代理能力既昂贵又低效。在代理人工智能基础设施方面的基础投资需要高度专业化的技术人才，这些人才稀缺、昂贵且难以招聘，而团队进行实验、学习和实施这些系统所需的时间则代表着有意义的机遇成本。

这为集中能力并在整个投资组合中捕捉真实协同效应提供了一个极具吸引力的机会。通过汇集代理人才、可重用代理、共享工具和累积的学习经验，公司可以大幅减少重复工作，同时加快单个投资组合公司的价值实现时间。随着时间的推移，代理本身也成为可重用资产——通过规模的扩大和使用案例的增加而不断改进——而不是每个公司各自重建的定制实施。

这种模型的一种自然表达是建立一个共享服务平台，该平台代表投资组合公司开发、运营和商业化代理。最初，此类平台将专注于满足投资组合内部的内部需求，提供成本杠杆和更快的部署。然而，随着时间的推移，这种共享能力可能演变成一个独立的企业，将代理解决方案、基础设施或托管代理服务销售到更广泛的市场。通过这种方式，原本作为内部效率策略的举措，有可能成为一项本身具有差异化的平台投资。

额外投资考虑因素

也存在投资者应该谨慎对待的明确领域。纯开发者工具点的解决方案难以扩展，因为开发者不太可能集中在一个单一的工具集。依赖于模型的赌注伴随着重大的平台和商品化风险，而单一云锁定解决方案可能会因为企业优先考虑便携性和可选择性而陷入困境。

制定一个吸引人的长期策略是：随着时间的推移，逐步构建一个综合的代理堆栈：从编排锚点开始，添加治理和注册，分层加入财务控制，通过MCP和集成能力进行扩展，最终扩展到代理市场能力。这种方法使资本与随着代理人工智能生态系统的成熟最有可能产生影响的控制点相一致。

1
标准协议由Anthropic管理，基于JSON，用于连接代理与后端系统。

2
Gartner、MarketDigits、Mordor Intelligence、Markets and Markets、Grand View Research；这包括上述所有子市场，但不包括LLM市场。

3
OSWAP，MITRE ATLAS，MAESTRO，42单元

4
数据栈中心

5
Flexera

作者

布伦特·斯莫林斯基
合作伙伴

米格尔·帕雷德斯
合作伙伴

Sridhar Narasimhan
合作伙伴

内森·贝尔
合作伙伴

关于基尔尼

过去100年，凯尼恩一直是领先的管理咨询公司，也是全球财富500强中四分之三公司和世界各地政府的信赖伙伴。我们在40多个国家设有分支机构，我们的员工成就了我们。我们以影响为先，用原创思维和共同实现变革的承诺来应对您最严峻的挑战。与您并肩，我们提供——价值、成果、影响。