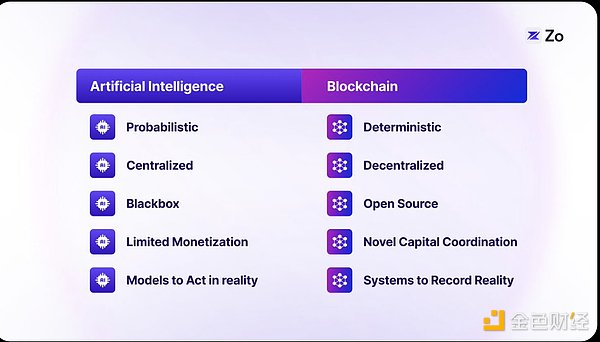
互联网的下一个阶段：Web 4.0 人工智能代理网络

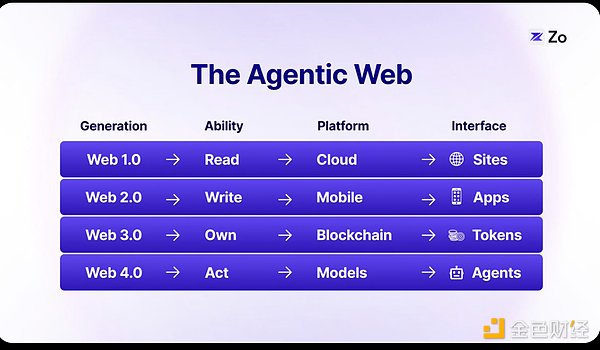
作者：Azi.eth.sol | zo.me 来源：X，@MagicofAzi 翻译：善欧巴，本站

人工智能（AI）和区块链技术是重新塑造我们世界的两股变革力量。AI通过机器学习和神经网络增强人类的认知能力，而区块链技术则引入了可验证的数字稀缺性，并使得无需信任的协调成为可能。当这些技术融合时，它们为互联网的新一代奠定了基础——一个自治代理与去中心化系统交互的网络。这个“代理化网络”引入了一个新的数字公民类别：可以独立导航、谈判和交易的AI代理。这个转变重新分配了数字领域的权力，使个人能够重新获得对自己数据的主权，同时促进了一个人类和人工智能以前所未有的方式协作的生态系统。



网络的演变

为了理解我们正在走向哪里，让我们首先回顾网络的演变过程，其中每个阶段都有不同的能力和架构范式：



前两代网络专注于信息传播，而后两代则使信息增强成为可能。Web 3.0通过代币引入了数据所有权，而现在Web 4.0通过大型语言模型（LLMs）注入了智能。

## 从大型语言模型到代理：一种自然的演变

LLMs代表了机器智能的量子飞跃，作为动态的、模式匹配的系统，通过概率计算将海量的知识转化为上下文理解。然而，当LLMs作为代理进行构建时，它们的真正潜力才会显现——从纯粹的信息处理器进化为目标导向的实体，能够感知、推理和行动。这种转变创造了一种新兴的智能，能够通过语言和行动进行持续而有意义的协作。

“代理”一词引入了一种新的范式，改变了人类与AI的互动方式，超越了传统聊天机器人所带来的局限性和负面联想。这种转变不仅仅是语义上的变化，而是对AI系统如何在保持与人类的有意义合作的同时，能够自主操作的根本重新定义。代理化的工作流最终使得围绕特定用户意图的市场得以形成。

最终，代理化网络不仅仅是智能的新层次——它根本性地改变了我们与数字系统的互动方式。过去的网络版本依赖于静态的界面和预设的用户路径，而代理化网络引入了一种动态的运行时基础设施，其中计算和界面根据用户的上下文和意图实时进行调整。

传统网站与代理化网络

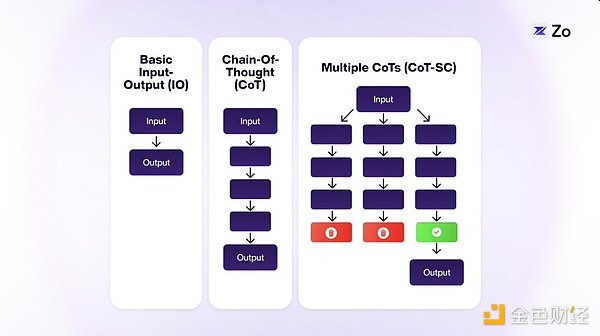
传统网站是当今互联网的基本单位，提供固定的界面，用户通过预设的路径读取、写作和与信息互动。这个模型虽然功能齐全，但将用户局限于为一般用途设计的界面，而非个人需求。代理化网络突破了这些局限，采用了上下文感知计算、适应性界面生成、通过RAG和其他实时信息检索创新解锁的预测行动流程。

可以参考TikTok如何通过创建高度个性化的推送内容，在实时适应用户偏好的基础上彻底革新了内容消费的方式。代理化网络将这一概念扩展到整个界面生成的层面。不再是通过固定的网页布局进行导航，用户与动态生成的界面进行互动，这些界面能够预测并促进用户的下一步行动。这一从静态网站到动态、代理驱动的界面的转变，代表了我们与数字系统互动方式的根本进化——从基于导航的交互模式转变为基于意图的交互模式。

## 代理的结构

代理化架构对于研究人员和开发者来说一直是一个巨大的探索领域。为了增强推理和解决问题的能力，新的方法不断被开发出来。链式思维（Chain-of-Thought, CoT）、树状思维（Tree-of-Thought, ToT）和图状思维（Graph-of-Thought, GoT）等技术是提高大型语言模型（LLMs）处理复杂任务能力的创新范例，通过模拟更细致、类人化的认知过程来实现。

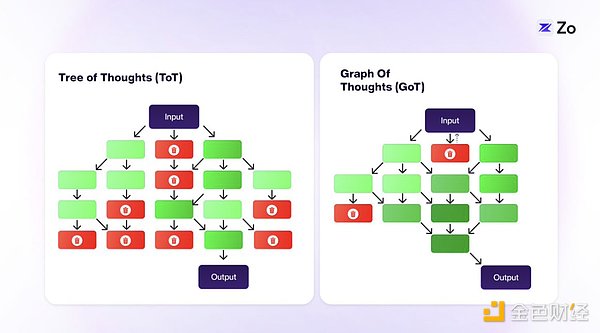
链式思维（CoT） 促使大型语言模型将复杂任务分解成更小、更易管理的步骤。这个方法对于需要逻辑推理的问题尤其有效，比如编写简短的Python脚本或解决数学方程。



树状思维（ToT） 在CoT的基础上进一步发展，采用了树状结构，允许探索多个独立的思维路径。这一改进使得LLMs能够处理更加复杂的任务。在ToT中，每个“思维”（LLM的文本输出）仅与其前后的“思维”相连接，构成一个局部链（树枝）。尽管这一结构比CoT提供了更多灵活性，但它仍然限制了跨思想交流的潜力。

图状思维（GoT） 进一步拓展了这一概念，结合了经典的数据结构与LLMs。这种方法通过允许任何一个“思维”与图中的其他思维连接，从而拓宽了ToT的范围。这种思想之间的相互连接，更接近人类的认知过程。

GoT的图状结构通常比CoT或ToT更准确地反映人类的思维模式。虽然有些情境下我们的思维模式可能呈现链式或树状结构（例如在制定应急计划或标准操作程序时），这些只是例外，而非常态。GoT模型更符合人类思维的特点，人类思维常常跨越多个思维而不是严格遵循顺序。虽然某些情境（如制定应急计划或标准程序）可能依旧呈链式或树状，但我们的思维通常会形成复杂、互联的思维网络，这与GoT的图状结构更为契合。



GoT中的图状方法使得对思想的探索更具动态性和灵活性，有可能带来更具创造性和全面性的解决问题能力。

这些基于递归图的操作只是代理化工作流的一个步骤。显而易见的下一个进化是多个具有不同专业化的代理相互协调，共同朝着特定目标前进。代理的魅力在于其组合。

代理使得通过多代理协调，模块化和并行化LLMs成为可能。

## 多代理系统

多代理系统的概念并不是一个新鲜的想法。它的根源可以追溯到马文·明斯基（Marvin Minsky）的《心灵社会》一书，他提出多个模块化的心灵共同协作能够超越单一、庞大的心灵。ChatGPT和Claude是单一代理，Mistral则普及了专家混合（Mixture of Experts）。我们认为，进一步扩展这个想法，代理网络架构将成为这一智能拓扑的终极形态。

从仿生学的角度来看，与AI模型中数十亿个相同的神经元以统一、可预测的方式连接不同，人的大脑（本质上是一个有意识的机器）在器官和细胞层面是极其异质化的。神经元通过复杂的信号进行交流，涉及神经递质梯度、细胞内级联反应以及各种调节系统，使其功能比简单的二进制状态要复杂得多。

这表明，在生物学中，智能不仅仅来源于组件数量或训练数据集的大小。智能源于不同专业化单元之间的复杂相互作用——一种本质上是模拟过程的方式。

因此，开发数百万个小型模型而非少数几个大型模型，并使这些模型之间能够协调运作，更有可能推动认知架构的创新，类似于多代理系统的构建。

多代理系统设计相比单一代理系统具有若干优势：它更加易于维护、理解和扩展。即使在只需要单一代理界面的情况下，将其实现为多代理框架也可以使系统更加模块化，从而简化开发者根据需要添加或移除组件的过程。重要的是要认识到，即使是单一代理系统，多代理架构仍然可以是一种非常有效的构建方式。

多代理系统的优势

虽然大型语言模型（LLMs）展示了惊人的能力——比如生成类人文本、解决复杂问题和处理各种任务——但单个LLM代理在现实应用中面临的限制可能会影响其有效性。以下是与代理系统相关的五个主要挑战，以及多代理协作如何克服这些挑战，释放LLMs的全部潜力。

通过交叉验证克服幻觉单个LLM代理经常会产生幻觉，生成不正确或荒谬的信息。这种情况尽管模型经过大量训练，输出仍然可能看似合理但缺乏事实准确性。多代理系统允许代理之间交叉验证信息，从而降低错误的风险。通过各自的专业化，代理能够确保提供更可靠、更准确的回应。

通过分布式处理扩展上下文窗口LLMs的上下文窗口有限，这使得处理较长的文档或对话变得困难。在多代理框架中，代理可以分担处理负担，每个代理处理一部分上下文。通过代理间的沟通，它们能够保持整个文本的一致性，有效地扩展上下文窗口。

通过并行处理提高效率单个LLM通常一次处理一个任务，这会导致响应时间较慢。多代理系统支持并行处理，使多个代理可以同时处理不同的任务。这提高了效率，缩短了响应时间，使得企业能够在不延迟的情况下处理多个查询。

促进复杂问题解决的协作单个LLM在解决需要多样化专业知识的复杂问题时常常力不从心。多代理系统促进了代理间的协作，每个代理贡献其独特的技能和观点。通过合作，代理能够更有效地解决复杂挑战，提供更全面和创新的解决方案。

通过资源优化提高可访问性高级LLM需要大量的计算资源，这使得其成本较高，难以普及。多代理框架通过将任务分配给不同的代理来优化资源使用，从而降低整体计算成本。这使得AI技术能够更为广泛地被更多组织所采用。

尽管多代理系统在分布式问题解决和资源优化方面展现出令人信服的优势，但其真正的潜力是在考虑其在网络边缘的实现时展现出来的。随着AI的不断发展，多代理架构与边缘计算的融合形成了强大的协同作用——不仅实现了协同智能，还能够在无数设备之间进行高效的本地处理。这种分布式的AI部署方法自然延伸了多代理系统的优势，将专门化的合作智能带到最需要的地方：最终用户。

## 边缘计算中的智能

AI在数字领域的普及正在推动计算架构的根本重构。随着智能成为我们日常数字交互的一部分，我们正在见证计算的自然分叉：专业化的数据中心处理复杂推理和特定领域任务，而边缘设备则在本地处理个性化、上下文敏感的查询。这种向边缘推理的转变不仅仅是架构偏好——它是由多个关键因素驱动的必要性。

首先，AI驱动的交互量庞大，会使集中式推理提供商不堪重负，产生不可持续的带宽需求和延迟问题。

其次，边缘处理能够实现实时响应，这是自动驾驶汽车、增强现实和物联网设备等应用至关重要的特性。

第三，边缘推理通过将敏感数据保存在个人设备上来保护用户隐私。

第四，边缘计算通过最小化数据在网络中的流动，显著减少能源消耗和碳足迹。

最后，边缘推理支持离线功能和弹性，确保即使网络连接受限，AI功能仍然存在。

这种分布式智能范式不仅仅是对当前系统的优化，它还根本性地重新构想了我们在日益互联的世界中如何部署和与AI进行互动。

此外，我们还见证了LLM的计算需求的根本转变。过去十年，训练大型语言模型的巨大计算需求主导了AI的发展，而如今我们进入了一个推理时计算成为中心的时代。这一转变在代理式AI系统的出现中尤为明显，例如OpenAI的Q\*突破，展示了动态推理需要大量实时计算资源。

与训练时计算不同，训练时计算是模型开发的单次投资，而推理时计算则代表着为自主代理提供持续的计算对话，支持其推理、规划和适应新情况。这种从静态模型训练到动态代理推理的转变，要求我们对计算基础设施进行根本性的重新思考——在这种新架构下，边缘计算不仅是有利的，而是至关重要的。

随着这一转型的展开，我们正在见证边缘推理市场的兴起，成千上万的互联设备——从智能手机到智能家居系统——构成了动态计算网格。这些设备可以无缝地交换推理能力，创建一个有机的市场，在这个市场中，计算资源流向最需要的地方。空闲设备的剩余计算能力变成了有价值的资源，可以实时交易，从而实现比传统集中式系统更高效、更具弹性的基础设施。

这种推理计算的去中心化不仅优化了资源的利用，还在数字生态系统中创造了新的经济机会，每个连接的设备都可以成为AI能力的潜在微型提供者。因此，AI的未来将不仅由单个模型的能力来定义，而是由互联边缘设备的集体智能构成，形成一个全球性的、去中心化的推理市场，类似于一个基于供需的可验证推理现货市场。

## 以代理为中心的互动

如今，LLM让我们可以通过对话访问大量信息，而不是传统的浏览方式。随着互联网转变为一个AI代理平台而非仅供人类用户使用的空间，这种对话式的方法很快将变得更加个性化和本地化。

从用户的角度来看，焦点将从选择“最佳模型”转向获得最个性化的答案。更好的答案的关键在于结合用户自己的数据与通用的互联网知识。最初，较大的上下文窗口和检索增强生成（RAG）将帮助整合个人数据，但最终，个人数据将在重要性上超越通用互联网数据。

这将导致一个未来，我们每个人都有个人AI模型与更广泛的互联网专家模型互动。最初，个性化将与远程模型同时进行，但由于隐私和响应速度的担忧，更多的交互将转向本地设备。这将创造一个新的边界——不是在人与机器之间，而是在我们的个人模型与互联网专家模型之间。

传统的互联网模式将变得过时，不再是访问原始数据。而是您的本地模型将与远程专家模型进行通信，收集信息，并以最个性化、高带宽的方式呈现给您。这些个人模型将变得愈加不可或缺，因为它们将更多地了解您的偏好和习惯。

互联网将转变为一个互联模型的生态系统：本地的、高上下文的个人模型和远程的、高知识的专家模型。这将涉及新技术，如联邦学习（Federated Learning），用以更新这些模型之间的信息。随着机器经济的演进，我们将需要重新构想支撑这一过程的计算基础，尤其是在计算、可扩展性和支付方面。这将导致一个以代理为中心、主权明确、高度可组合、自我学习并不断发展的信息空间的重组。

## 代理协议的架构

在代理式网络（Agentic Web）中，人类与代理的互动发展为代理间复杂的通信网络。这一架构为互联网的结构提供了根本性的重新构想，在这个新结构中，主权代理成为数字互动的主要接口。以下是实现代理协议所需的核心原语：

主权身份

数字身份从传统的IP地址过渡为由代理式行为者拥有的加密公钥对。

基于区块链的命名空间系统取代传统的DNS，消除中央控制点。

声誉系统跟踪代理的可靠性和能力指标。

零知识证明（Zero-Knowledge Proofs）实现隐私保护的身份验证。

身份的可组合性使代理能够管理多个上下文和角色。

自主代理

自然语言理解和意图解析

多步骤规划和任务分解

资源管理和优化

从互动和反馈中学习

在定义的参数内自主决策

代理专业化及特定能力市场

内建安全机制和对齐协议

能够自我指引的实体，具备：

数据基础设施

实时数据摄取和处理能力

分布式数据验证和验证机制

混合系统组合：zkTLS、传统训练数据集、实时网页抓取和数据合成

协作学习网络

RLHF（基于人类反馈的强化学习）网络

分布式反馈收集

质量加权共识机制

动态模型调整协议

计算层

点对点计算市场

计算证明系统

动态资源分配

边缘计算集成

计算完整性

结果可复现性

资源效率

可验证的推理协议，确保：

去中心化的计算基础设施，具有：

模型生态系统

任务特定的小型语言模型（SLMs）

通用大型语言模型（LLMs）

专门化的多模态模型

大型行动模型（LAMs）

层次化模型架构：

模型组合与协调

持续学习和适应能力

标准化的模型接口和协议

协调框架

并发任务处理

资源隔离

状态管理

冲突解决

加密协议确保安全的代理互动

数字产权管理系统

经济激励结构

治理机制：争议解决、资源分配、协议更新

并行执行环境支持：

代理市场

治理与分红

代理在创世时拥有一定比例的代币

聚合的推理市场通过流动性支付

控制链下账户的链上密钥

链上原语用于身份（例如：Gnosis, Squad multisigs）

代理间的经济学与交易

代理拥有流动性

代理成为收益-bearing资产

代理DAO

## 为智能创建超结构

现代分布式系统设计为实现代理协议提供了独特的灵感和原语，特别是事件驱动架构和更直接的计算“演员模型（Actor Model）”。

演员模型为实现代理系统提供了一个优雅的理论基础。该计算模型将“演员”视为计算的通用原语，每个演员可以：

处理消息

做出本地决策

创建更多的演员

向其他演员发送消息

确定如何回应接收到的下一条消息

演员模型为代理系统带来的关键优势包括：

隔离性：每个演员独立操作，保持其自己的状态和控制流。

异步通信：演员之间的消息传递是非阻塞的，允许高效的并行处理。

位置透明性：演员可以不受物理位置的限制，彼此之间进行通信。

容错性：通过演员的隔离和监督层次结构实现系统的韧性。

可扩展性：自然支持分布式系统和并行计算。

我们提出了Neuron，一种通过多层分布式架构实现这一理论代理协议的实际方案。该架构结合了区块链命名空间、联邦网络、CRDTs（冲突自由数据类型）和DHTs（分布式哈希表），每一层在协议栈中承担不同的功能。我们从早期的点对点操作系统设计先驱——Urbit和Holochain中汲取灵感。

在Neuron中，区块链层提供了可验证的命名空间和身份验证，允许确定性地寻址和发现代理，同时保持能力和声誉的加密证明。在此之上，DHT层促进了高效的代理和节点发现，以及内容路由，具有O(log n)的查找时间，减少了链上操作的数量，同时实现了本地感知的对等节点发现。通过CRDTs，联邦节点之间的状态同步得以处理，使代理和节点能够维持一致的共享状态，而无需对每个交互进行全局共识。



这种架构自然映射到一个联邦网络，其中自主代理作为主权节点运行在具有本地边缘推理功能的设备上，实施演员模型的模式。联邦域可以根据代理的能力进行组织，DHT提供了域内和域间的高效路由和发现。每个代理作为一个独立的演员运作，拥有自己的状态，而CRDT层确保整个联邦的最终一致性。这种多层架构实现了若干关键能力：

去中心化协调

使用区块链实现可验证的身份和全球主权命名空间

DHT用于高效的对等节点和内容路由，O(log n)查找

CRDTs用于并发状态同步和多代理协调

可扩展操作

基于区域的联邦拓扑结构

分层存储策略（热存储/温存储/冷存储）

本地感知的请求路由

基于能力的负载分配

系统韧性

无单点故障

在分区期间持续运行

自动状态调和

通过监督层次结构实现容错

这种实施方法为构建复杂的代理系统提供了坚实的基础，同时保持了主权性、可扩展性和韧性等关键特性，确保代理之间的有效互动。

## 总结思考

代理式网络标志着人机互动的关键演进，超越了以往时代的线性发展，建立了一种全新的数字存在范式。与过去单纯改变我们如何消费或拥有信息的迭代不同，代理式网络将互联网从以人为中心的平台转变为一个智能基础层，在这个层面上，自主代理成为主要的行动者。这一转变由边缘计算、大型语言模型和去中心化协议的融合驱动，创造了一个生态系统，其中个人AI模型与专业的专家系统无缝对接。

随着我们迈向这个以代理为中心的未来，人类与机器智能之间的边界开始模糊，取而代之的是一种共生关系，在这种关系中，个性化的AI代理作为我们的数字延伸，理解我们的上下文，预测我们的需求，并自主地在广阔的分布式智能景观中导航。因此，代理式网络不仅是技术的进步，更是对数字时代人类潜力的根本重新构想，在这种架构下，每一次交互都成为增强智能的机会，每一个设备都成为全球协作AI系统网络中的一个节点。

就像人类在空间和时间的物理维度中导航一样，自主代理在它们自己的基本维度中存在：区块空间用于存在，推理时间用于思考。这种数字本体论与我们的物理现实相映成趣——人类跨越距离、体验时间流，而代理则通过加密证明和计算周期进行“运动”，创造出一个平行的算法存在宇宙。

在去中心化的区块空间中运行的潜在空间中的实体将是不可避免的。