AI on AO 发布会：AO协议的三大AI技术突破

作者：Kyle\_13，来源：作者推特@kylewmi

感谢今天各位的到来。我们有一系列关于 AO 技术的超级令人兴奋的进展要与大家分享。我们会先进行一个演示，然后 Nick 和我会在这里尝试构建一个 AI 代理，这个代理将在智能合约中使用大型语言模型，基于你们即将听到的系统中的聊天情绪进行买卖。今天我们将现场从头开始构建它，希望一切顺利。

是的，你们将看到如何自己做这一切。

这里的技术进步确实让 AO 远远超越了其他智能合约系统。这在以前已经是事实，现在它越来越像一个去中心化的超级计算机，而不是传统的智能合约网络。但它具备了智能合约网络的所有特性。因此，我们非常激动地与大家分享这一切。事不宜迟，让我们开始演示，然后我们会进行讨论，并且我们将一起现场构建一些东西。

大家好，感谢今天的到来。我们非常激动地宣布 AO 协议的三项重大技术更新。它们共同实现了一个大目标，即支持在去中心化环境中运行的大型语言模型作为智能合约的一部分。这些不仅仅是玩具模型，小模型，或者被编译成自己二进制文件的模型。

这是一个完整的系统，可以让你运行几乎所有当前开源和可用的主要模型。比如 Llama 3 在链上运行在智能合约中，GPT 也是如此，还有苹果的模型等等。这是整个生态系统共同努力的成果，有三项主要的技术进步也构成了这个系统的一部分。因此，我非常激动地向大家介绍这一切。



大体情况是，现在 LLM（大型语言模型）可以在智能合约中运行。你可能多次听说去中心化 AI 和 AI 加密货币。实际上，除了一种我们今天要讨论的系统外，几乎所有这些系统都是作为预言机的 AI，也就是在链下运行 AI，然后将执行结果放到链上，用于下游的某些用途。

我们说的不是这个。我们说的是将大型语言模型推理作为智能合约状态执行的一部分。这一切都得益于我们拥有的 AO 硬盘以及 AO 的超并行处理机制，这意味着你可以运行大量的计算，而这不会影响我正在使用的不同进程。我们认为，这将允许我们创建一个非常丰富的去中心化自主代理金融系统。



到目前为止，在去中心化金融（DeFi）中，我们基本上已经能够使原始交易的执行变得无需信任。在不同经济游戏中的交互，比如借贷和交换，都是无需信任的。这只是问题的一个方面。如果你考虑全球金融市场。

是的，有各种不同的经济原件在不同方式中发挥作用。有债券、股票、商品和衍生品等等。但我们真正谈论市场时，不仅仅是这些，实际上是智能层。是决定买卖、借贷或参与各种金融游戏的人。

到目前为止，在去中心化金融生态系统中，我们成功地将所有这些原件转移到一个无需信任的状态中。因此，你可以在 Uniswap 上进行交换，而不需要信任 Uniswap 的运营者。实际上，从根本上讲，没有运营者。市场的智能层被留在了链下。所以，如果你想参与加密货币投资，而不想自己做所有研究和参与，你必须找到一个基金。

你可以信任他们的资金，然后他们去执行智能决策，并将其下游传递到网络本身的基本原件执行中。我们认为，在 AO 中，我们实际上已经具备了将市场的智能部分，即导致决策的智能，转移到网络本身的能力。因此，一个简单的理解方式可能是想象一下。

一个你可以信任的对冲基金或投资组合管理应用程序，可以在网络内执行一组智能指令，从而将网络的无需信任性转移到决策过程中。这意味着一个匿名账户，例如 Yolo 420 Trader Number One（一个大胆、随意的交易者），可以创建一个新的有趣策略，并将其部署到网络上，而你可以将资本投入其中而无需真正信任。

你现在可以构建与大型统计模型交互的自主代理。而最常见的大型统计模型是大型语言模型，可以处理和生成文本。这意味着你可以将这些模型放入智能合约中，作为由某个有新颖想法的人开发的策略的一部分，并在网络中智能地执行它们。



你可以想象做一些基础的情绪分析。比如你读取新闻，决定这是买卖这个衍生品的好时机。这是执行这种或那种操作的好时机。你可以让类似人类的决策以无需信任的方式执行。这不仅仅是理论。我们创建了一个有趣的迷因币，称为 Llama Fed。基本上这个想法是，它是一个法币模拟器，其中的一群 llama（羊驼）由 Llama 3 模型表示。

它们就像 llama 和美联储主席的结合体，你可以去找它们，请求它们发放一些代币，它们会评估你的请求。大型语言模型本身操作货币政策，完全自主、无需信任。我们构建了它，但我们无法控制它。它们操作货币政策，决定谁应该获得代币，谁不应该。这是这一技术的一个非常有趣的小应用，希望能激发生态系统中所有其他可能的应用。



为了实现这一点，我们必须为 AO 创建三种新的基础能力，其中一些在基础协议层，另一些在应用层。这不仅对大型语言模型的执行有用，对 AO 开发者来说更为广泛和激动人心。所以我很高兴今天能向大家介绍这些。

这些新技术中的第一个是 Web Assembly 64 位支持。听起来有点像技术术语，但我有办法让每个人都理解它的意义。从根本上讲，Web Assembly 64 支持允许开发人员创建使用超过 4GB 内存的应用程序。稍后我们会介绍新的限制，它们相当惊人。

如果你不是开发人员，可以这样理解：有人让你写一本书，你对这个想法很兴奋，但他们说你只能写 100 页。不多不少。你可以表达书中的想法，但无法以自然和正常的方式进行，因为有外部限制，你必须迎合它，改变你的写作方式以适应它。



在智能合约生态系统中，这不仅仅是 100 页的限制。我会说这有点像在早期版本的 AO 中构建。以太坊有 48KB 的内存限制，就像有人让你写一本只有一句话长的书，而且你只能使用前 200 个最流行的英语单词。在这个系统中构建真正令人兴奋的应用程序极其困难。



然后是 Solana，你可以访问 10MB 的工作内存。这显然是一个改进，但基本上我们说的是一页纸。ICP，互联网计算机协议，允许支持 3GB 的内存。理论上是完整的，但他们不得不降低到 3GB。现在有了 3GB 的内存，你可以运行很多不同的应用程序，但你肯定不能运行大型 AI 应用程序。它们需要加载大量数据到主内存，以便快速访问。这在 3GB 内存中无法有效地实现。

当我们在今年二月份发布 AO 时，我们也有 4GB 内存限制，这实际上源自 Web Assembly 32。现在，这个内存限制在协议层面完全消失了。相反，协议层面的内存限制是 18EB（艾字节）。这是一个巨大的存储量。

直到这在内存中用于计算而不是长期存储介质之前，会需要相当长的时间。在实现层面，AO 网络中的计算单元，现在能够访问 16GB 内存，但在未来无需更改协议就可以将其替换为更大容量的内存，这将相对容易。16GB 已经足够运行大型语言模型计算，这意味着你今天可以在 AO 上下载和执行 16GB 的模型。比如 Llama 3 未量化版本的 Falcon 3 以及许多其他模型。

这是构建智能语言基础计算系统所必需的核心组件。现在它作为智能合约的一部分在链上完全支持，我们认为这是非常非常令人兴奋的。

这消除了 AO 以及随后智能合约系统的一大主要计算限制。当我们在今年二月发布 AO 时，你可能会注意到视频中我们多次提到你有无限的计算能力，但有一个限制，就是不能超过 4GB 内存。这就是那个限制的解除。我们认为这是一个非常令人兴奋的进步，16GB 已经足够运行当前 AI 领域几乎所有你想运行的模型。

我们能够在不更改协议的情况下提升 16GB 限制，未来这将相对容易，与最初运行 Web Assembly 64 相比，这是一大进步。因此，这本身就是系统能力的巨大进步。第二个使大型语言模型能够在 AO 上运行的主要技术是 WeaveDrive。



WeaveDrive 让你可以像本地硬盘一样访问 AO 内的 Arweave 数据。这意味着你可以打开 AO 中任何经过调度单元认证的事务 ID 并上传到网络。当然，你可以访问这些数据并将其读入你的程序中，就像本地硬盘上的文件一样。



我们都知道，目前 Arweave 上存储了大约 60 亿条事务数据，所以这是一个巨大的数据集起点。这也意味着在未来构建应用程序时，上传数据到 Arweave 的动机会增加，因为这些数据也可以在 AO 程序中使用。例如，当我们使大型语言模型在 Arweave 上运行时，我们大约上传了价值 1000 美元的模型到网络上。但这只是开始。

有了本地文件系统的智能合约网络，你可以构建的应用程序数量是巨大的。因此，这非常令人兴奋。更好的是，我们构建的系统允许你将数据流式传输到执行环境中。这是一个技术上的细微差别，但你可以想象回到书的类比中。

有人对你说，我想访问你这本书中的一个数据。我想获取这本书中的一个图表。在一个简单的系统中，甚至在目前的智能合约网络中，这都会是一个巨大的进步，你会给出整本书。然而这显然效率不高，特别是如果那本书是一个包含数千页的大型统计模型。

这极其低效。相反，我们在 AO 中所做的是让你可以直接读取字节。你直接到书中的图表位置，只复制图表到你的应用程序中并执行。这使得系统的效率得到了极大的提升。这不仅是一个最小可行产品（MVP），它是一个功能齐全、构建良好的数据访问机制。因此你有一个无限计算系统和一个无限硬盘，将它们结合在一起，你就有了一台超级计算机。

这在以前从未被构建过，现在它以最低成本提供给每个人使用。这就是 AO 的现状，我们对此非常兴奋。该系统的实现也在操作系统层面。因此，我们将 WeaveDrive 变成 AO 的一个子协议，它是一个计算单元扩展，任何人都可以加载。这很有趣，因为这是首个此类扩展。

AO 一直具备让你为执行环境添加扩展的能力。就像你有一台计算机，你想插入更多的内存，或者插入一张显卡，你物理上将一个单元放入系统中。你可以对 AO 的计算单元执行此操作，这就是我们在这里做的。因此，在操作系统层面，你现在有一个硬盘，它只是代表数据存储的文件系统。

这意味着你不仅可以在 AO 中访问这些数据，以通常的方式构建应用程序，但实际上你可以从任何带到网络上的应用程序中访问它。因此，这是一个广泛适用的能力，所有在系统中构建的人员都可以访问，不论他们使用何种语言编写，Rust、C、Lure、Solidity 等都可以访问，像它是系统的原生功能一样。在构建这一系统的过程中，它还迫使我们创建了子协议协议，创建其他计算单元扩展的方法，以便将来其他人也能构建激动人心的东西。

现在我们有能力在任意大小的内存集中运行计算，并能将网络中的数据加载到 AO 内的进程中，接下来要问的是，如何进行推理本身。

由于我们选择在 Web Assembly 上构建 AO 作为其主要虚拟机，因此将现有代码编译并运行在该环境中相对容易。由于我们构建了 WeaveDrive 使其暴露为操作系统层面的文件系统，将 Llama.cpp（一个开源的大型语言模型推理引擎）运行在系统上实际上是相对容易的。



这非常令人兴奋，因为这意味着你不仅可以运行此推理引擎，还可以轻松运行许多其他引擎。因此，使大型语言模型在 AO 内运行的最后一个组件是大型语言模型推理引擎本身。我们移植了一个名为 Llama.cpp 的系统，听起来有点神秘，但实际上它是当前领先的开源模型执行环境。

直接在 AO 智能合约内运行，一旦我们有能力在系统中拥有任意数量的数据，然后从 Arweave 中加载任意数量的数据，这实际上是相对容易的。

为了实现这一点，我们还与称为 SIMD（单指令多数据）计算扩展合作，这使你可以更快地运行这些模型。因此我们也启用了这一功能。这意味着目前这些模型在 CPU 上运行，但速度相当快。如果你有异步计算，它应该适合你的使用场景。像读取新闻信号然后决定执行哪些交易，在目前的系统下运行良好。但我们也有一些令人兴奋的升级将在不久后谈到，关于其他加速机制，如使用 GPU 加速大型语言模型推理。



Llama.cpp 允许你加载不仅是 Meta 的领先模型 Llama 3，还可以加载许多其他模型，实际上大约 90% 以上你可以从开源模型网站 Hugging Face 下载的模型都可以在系统内运行，从 GPT-2 如果你愿意，到 253 和 Monet，苹果自己的大型语言模型系统以及许多其他模型。因此我们现在有了框架，可以将任何模型从 Arweave 上传，使用硬盘上传我想在系统中运行的模型。你上传它们，它们只是正常的数据，然后你可以加载它们到 AO 的进程中并执行，获取结果并以你喜欢的方式工作。我们认为这是一个包，它使在以前的智能合约生态系统中无法实现的应用程序成为可能，即使现在可能实现，在现有系统如 Solana 中进行架构更改的数量也只是难以预料，不在其路线图上。因此为了向你们展示这一点并使其变得真实和易于理解，我们创建了一个模拟器 Llama Fed。基本思想是我们得到一个美联储成员委员会，它们是 llama，无论在作为元 llama 3 模型还是在作为美联储主席这一点上都是 llama。

我们也告诉它们它们是 llama，像 Alan Greenspan 或美联储主席一样。你可以进入这个小环境。

一些人会熟悉这个环境，实际上它就像我们今天工作的 Gather，你可以和 llama 交谈，请求它们给予你一些代币用于一个非常有趣的项目，并且它们会根据你的请求决定是否给予你代币。因此你燃烧一些 Arweave 代币，wAR 代币（由 AOX 团队提供），它们会根据认为你的提案是否好给予你代币。因此这是一个迷因币，货币政策完全自主和智能化。虽然这是一种简单形式的智能，但仍然有趣。它将评估你的提案和其他人的提案，并运行货币政策。通过分析新闻标题并做出智能决策或与客户支持交互并返回价值，所有这些现在都可以在智能合约内实现。Elliot 现在将为大家展示。



大家好，我是 Elliot，今天我要向你展示 Llama Land，这是一个在 AO 内运行的链上自主世界，由 Meta 的开源 Llama 3 模型驱动。

我们在这里看到的对话不仅仅是玩家之间的对话，还有完全自主的数字 llama。





例如这个 llama 是人类。



但这个 llama 是链上 AI。

这栋建筑包含了 Llama fed。它就像美联储，但为 llama 服务。

Llama fed 运行世界上第一个 AI 驱动的货币政策，并铸造 Llama 代币。



这个家伙是 Llama 国王。你可以向他提供包装的 Arweave 代币（wAR），并写一个请求以获取一些 Llama 代币。



Llama 国王 AI 会评估并决定是否授予 Llama 代币。Llamafed 的货币政策完全自主，没有人为监督。世界上的每个代理和每个房间本身都是 AO 上的链上进程。

看起来 Llama 国王授予了我们一些代币，如果我查看我的 ArConnect 钱包，我可以看到它们已经在那里了。不错。Llama Land 只是第一个在 AO 上实现的 AI 驱动的世界。这是一个新协议的框架，允许任何人构建自己的自主世界，唯一的限制是你的想象力。所有这些都是 100% 链上实现的，只有在 AO 上才有可能。



谢谢 Elliot。你刚刚看到的不仅是一个大型语言模型参与金融决策并运行一个自主的货币政策系统。没有后门，我们无法控制它，所有这些都是由 AI 本身运行的。你还看到一个小宇宙，一个你可以在物理空间中行走的地方，你可以去那个地方与金融基础设施互动。我们认为，这不仅是一个有趣的小演示。

实际上这里有一些非常有趣的东西，这些地方将不同使用金融产品的人聚集在一起。我们在 DeFi 生态系统中看到，如果有人想参与一个项目，他们首先在 Twitter 上查看，然后访问网站，参与游戏中的基本原件。



然后他们加入 Telegram 群组或 Discord 频道或在 Twitter 上与其他用户交谈。这种体验非常去中心化，我们都在不同应用程序之间跳转。我们正在尝试的一个有趣想法是，如果你有这些 DeFi 应用程序的用户界面，让他们的社区可以聚集在一起并共同管理这个他们集体访问的自主空间，因为这是一个永久的网络应用，可以加入体验。

想象一下你可以去一个看起来像拍卖行的地方，和其他喜欢该协议的用户在一起聊天。当在 AO 上发生的金融机制进程有活动时，基本上你可以和其他用户聊天。社区和社交方面与产品的金融部分结合在一起。

我们认为这非常有趣，甚至有更广泛的影响。你可以在这里构建一个自主 AI 代理，它在这个 Arweave 世界中四处游荡，与它发现的不同应用程序和用户互动。所以如果你正在构建一个元宇宙，当你创建一个在线游戏时，第一件事就是创建 NPC（非玩家角色）。在这里，NPC 可以是通用的。

你有一个智能系统，它四处游荡，与环境互动，所以你没有用户冷启动问题。你可以有一些自主代理，试图为自己赚钱，试图交朋友，像正常的 DeFi 用户一样，与环境互动。我们认为这非常有趣，虽然有点怪异。我们将拭目以待。

展望未来，我们还看到了在 AO 中加速大型语言模型执行的机会。早些时候我谈到了计算单元扩展的概念。这就是我们用来构建 WeaveDrive 的方式。



不仅仅停留在 WeaveDrive，你可以为 AO 的计算环境构建任何类型的扩展。有一个非常令人兴奋的生态项目正在为 GPU 加速大型语言模型执行解决这个问题，就是 Apus Network。我让他们来解释。



嗨，我是 Mateo。今天我很兴奋地介绍 Apus Network。Apus Network 致力于构建去中心化的、无需信任的 GPU 网络。

我们通过利用 Arweave 永久的链上存储，提供一个开源 AO 扩展，提供 GPU 的确定性执行环境，并为去中心化 AI 提供经济激励模型，使用 AO 和 APUS 代币。Apus Network 将使用 GPU 挖矿节点竞争性地执行最优的、无需信任的模型训练在 Arweave 和 AO 上运行。这确保了用户可以以最具成本效益的价格使用最佳 AI 模型。你可以在 X（Twitter）@apus\_network 上关注我们的进展。谢谢。



这就是今天 AO 上的 AI 现状。你可以去试试 Llama Fed，尝试自己构建基于大型语言模型的智能合约应用。我们认为这是引入市场智能到去中心化执行环境的开始。我们对此非常激动，期待看到接下来会发生什么。谢谢大家今天的参与，期待再次与大家交流。