IOSG｜同质化 AI 基础设施的出路在哪儿？

本研究旨在探讨对开发者而言哪些人工智能领域最为重要，以及在 Web3 和人工智能领域哪些可能是爆发的下一个机遇。

在分享新的研究观点之前，首先很高兴我们参与了RedPill总计500万美元的第一轮融资，也非常激动，期待接下来能够和RedPill共同成长！



TL;DR

随着Web3 与 AI 的结合成为加密货币界的瞩目话题，加密世界的AI 基础设施构建兴旺起来，但实际利用 AI 或为 AI 构建的应用程序并不多，AI基础设施的同质化问题逐渐显现。近期我们参与的RedPill的第一轮融资，引发了一些更深入的理解。

构建 AI Dapp 的主要工具包括去中心化 OpenAI 访问、GPU 网络、推理网络和代理网络。

之所以说 GPU 网络比 “比特币挖矿时期” 还要热门，是因为：AI 市场更大，并且增长快速且稳定；AI 每天支持数以百万计的应用程序；AI 需要多样化的 GPU 型号和服务器位置；技术比过往更成熟；面向的客户群也更广。

推理网络和代理网络有相似的基础设施，但关注点不同。推理网络主要供有经验的开发者部署自己的模型，而运行非 LLM 模型并不一定需要 GPU。代理网络则更专注于 LLM，开发者无需自带模型，而是更注重提示工程和如何将不同的代理联结起来。代理网络总是需要高性能的 GPU。

AI 基础设施项目承诺巨大，仍在不断推出新功能。

多数原生加密项目仍处于测试网阶段，稳定性差，配置复杂，功能受限，还需要时间来证明其安全性和隐私性。

假设 AI Dapp 成为大趋势，还有许多未开发的领域，如监控、与 RAG 相关的基础设施、Web3 原生模型、内置加密原生 API 和数据的去中心化代理、评估网络等。

垂直整合是一个显著趋势。基础设施项目试图提供一站式服务，简化 AI Dapp 开发者的工作。

未来将是混合型的。部分推理在前端进行，而部分在链上计算，这样做可以考虑到成本和可验证性因素。

Source:IOSG



引言

Web3 与 AI 的结合是当前加密领域中最受瞩目的话题之一。才华横溢的开发者正在为加密世界构建 AI 基础设施，致力于将智能带入智能合约。构建 AI dApp 是极其复杂的任务，开发者需要处理的范围包括数据、模型、计算力、操作、部署和与区块链的整合。针对这些需求，Web3 创始人已经开发出许多初步的解决方案，如 GPU 网络、社区数据标注、社区训练的模型、可验证的 AI 推理与训练以及代理商店。

而在这兴旺的基础设施背景下，实际利用 AI 或为 AI 构建的应用程序并不多。开发者在查找 AI dApp 开发教程时，发现这些与原生加密 AI 基础设施相关的教程并不多，大多数教程仅涉及在前端调用 OpenAI API。

Source:IOSGVentures



当前的应用未能充分发挥区块链的去中心化和可验证功能，但这种状况很快将会改变。现在，大多数专注于加密领域的人工智能基础设施已经启动了测试网络，并计划在未来 6 个月内正式运行。

本研究将详细介绍加密领域人工智能基础设施中可用的主要工具。让我们准备迎接加密世界的 GPT-3.5 时刻吧！



## 1. RedPill：为 OpenAI 提供去中心化授权

前文所提到的我们参投的RedPill是一个很好的引入点。

OpenAI 拥有几种世界级强大的模型，如 GPT-4-vision、GPT-4-turbo 和 GPT-4o，是构建先进人工智能 Dapp 的优选。

开发者可以通过预言机或前端接口调用 OpenAI API 以将其集成到 dApp 中。

RedPill 将不同开发者的 OpenAI API 整合在一个接口下，为全球用户提供快速、经济且可验证的人工智能服务，从而实现了对顶尖人工智能模型资源的民主化。RedPill 的路由算法会将开发者的请求定向到单一贡献者处。API 请求将通过其分发网络执行，从而绕过任何来自 OpenAI 的可能限制，解决了加密开发者面临的一些常见问题，如：

限制 TPM（每分钟代币）：新账户对代币的使用有限，无法满足热门且依赖 AI 的 dApp 的需求。

访问限制：一些模型对新账户或某些国家的访问设置了限制。

通过使用相同的请求代码但更换主机名，开发者能以低廉的成本、高扩展性和无限制的方式访问 OpenAI 模型。







## 2. GPU 网络

除了使用 OpenAI 的 API，许多开发人员还会选择自行在家中托管模型。他们可以依托去中心化 GPU 网络，如 io.net、Aethir、Akash 等流行的网络，自行建立 GPU 集群并部署及运行各种强大的内部或开源模型。

这样的去中心化 GPU 网络，能够借助个人或小型数据中心的计算力，提供灵活的配置、更多的服务器位置选择以及更低的成本，让开发人员可以在有限的预算内轻松进行 AI 相关的试验。然而，由于去中心化的性质，此类 GPU 网络在功能性、可用性和数据隐私方面还存在一定的局限。



过去几个月，GPU 的需求火爆，超过了之前的比特币挖矿热潮。此现象的原因包括：

目标客户增多，GPU 网络现在服务于 AI 开发者，他们的数量不仅庞大而且更为忠实，不会受到加密货币价格波动的影响。

相比挖矿专用设备，去中心化 GPU 提供了更多的型号和规格，更能满趡iez要求。尤其是大型模型处理需要更高的 VRAM，而小型任务则有更适合的 GPU 可选。同时，去中心化 GPU 能够近距离服务终端用户，降低延迟。

技术日趋成熟，GPU 网络依赖高速区块链如 Solana 结算、Docker 虚拟化技术和 Ray 计算集群等。

在投资回报方面，AI 市场正在扩张，新应用和模型的开发机会多，H100 模型的预期回报率为 60-70%，而比特币挖矿则更为复杂，赢家通吃，产量有限。

比特币挖矿企业如 Iris Energy、Core Scientific 和 Bitdeer 也开始支持 GPU 网络，提供 AI 服务，并积极购买专为 AI 设计的 GPU，如 H100。

推荐：对于不太重视 SLA 的 Web2 开发者，io.net 提供了简洁易用的体验，是个性价比很高的选择。



## 3. 推理网络

这是加密原生 AI 基础设施的核心。它将在未来支持数十亿次 AI 推理操作。许多 AI layer1 或 layer2 为开发者提供了在链上原生调用 AI 推理的能力。市场领导者包括 Ritual、Valence 和 Fetch.ai。

这些网络在以下方面存在差异：

性能（延迟、计算时间）

支持的模型

可验证性

价格（链上消耗成本、推理成本）

开发体验

### 3.1 目标

理想的情况是，开发者可以在任何地方，通过任何形式的证明，轻松地访问自定义的 AI 推理服务，整合过程中几乎没有任何阻碍。

推理网络提供了开发者所需的全部基础支持，包括按需生成和验证证明、进行推理计算、推理数据的中继和验证、提供 Web2 和 Web3 的接口、一键式模型部署、系统监测、跨链操作、同步集成及定时执行等功能。



Source:IOSGVentures

借助这些功能，开发者可以将推理服务无缝集成到他们现有的智能合约中。例如，在构建 DeFi 交易机器人时，这些机器人会利用机器学习模型寻找特定交易对的买卖时机，并在基础交易平台上执行相应的交易策略。

在完全理想的状态下，所有的基础结构都是云托管的。开发者只需将他们的交易策略模型以通用格式如 torch 上传，推理网络就会存储并为 Web2 和 Web3 查询提供模型。

所有模型部署步骤完成后，开发者可以直接通过 Web3 API 或智能合约调用模型推理。推理网络将持续执行这些交易策略，并将结果反馈给基础智能合约。如果开发者管理的社区资金量很大，还需要提供推理结果的验证。一旦收到推理结果，智能合约就会根据这些结果进行交易。



Source:IOSGVentures

 3.1.1异步与同步

从理论上讲，异步执行的推理操作可以带来更好的性能表现；然而，这种方式在开发体验上可能让人感到不便。

在采用异步方式时，开发者需要先将任务提交到推理网络的智在合约中。当推理任务完成后，推理网络的智能合约会将结果返回。在这种编程模式下，逻辑被分为推理调用和推理结果处理两个部分。



Source:IOSGVentures

如果开发者有嵌套的推理调用和大量的控制逻辑，情况会变得更糟。



Source:IOSGVentures

异步编程模式使得它难以与现有的智能合约集成。这需要开发者编写大量额外的代码，并进行错误处理和管理依赖关系。

相对地，同步编程对于开发者来说更加直观，但它在响应时间和区块链设计上引入了问题。例如，如果输入数据是区块时间或者价格这种快速变动的数据，那么在推理完成后数据已不再新鲜，这可能会导致在特定情况下智能合约的执行需要回滚。想象一下，你用一个过时的价格来做交易。



Source:IOSGVentures

大部分 AI 基础架构采用异步处理，但 Valence 正在尝试解决这些问题。

### 3.2现实情况

实际上，许多新的推理网络还在测试阶段，如 Ritual 网络。根据他们的公开文件，这些网络目前的功能较为有限（诸如验证、证明等功能还未上线）。他们目前没有提供一个云基础设施以支持链上 AI 计算，而是提供了一个框架，用于自我托管 AI 计算并将结果传递至链上。

这是一个运行 AIGC NFT 的体系结构。扩散模型生成 NFT 并上传至 Arweave。推理网络会用这个 Arweave 地址在链上铸造该 NFT。



Source:IOSGVentures

这个过程非常复杂，开发者需要自己部署和维护大多数基础设施，如配有定制服务逻辑的 Ritual 节点、Stable Diffusion 节点及 NFT 智能合约。

推荐：目前的推理网络在整合和部署自定义模型方面相当复杂，且在这一阶段大多数网络还不支持验证功能。将 AI 技术应用到前端会为开发者提供一个相对简单的选择。如果你非常需要验证功能，ZKML 提供商 Giza 是个不错的选择。



## 4. 代理网络

代理网络让用户能轻松自定义代理。这样的网络由能自主执行任务、相互交云以及与区块链网络交互的实体或智能合约组成，这一切无需人工直接干预。它主要针对 LLM 技术。例如，它可以提供一个深入了解以太坊的 GPT 聊天机器人。这种聊天机器人目前的工具较为有限，开发者还不能在此基础上开发复杂的应用。



Source:IOSGVentures

但是将来，代理网络将提供更多的工具给代理使用，不仅仅是知识，还包括调用外部 API、执行特定任务的能力等。开发者将能够将多个代理连接起来构建工作流。例如，编写 Solidity 智能合约会涉及多个专门的代理，包括协议设计代理、Solidity 开发代理、代码安全审查代理以及 Solidity 部署代理。

Source:IOSGVentures

我们通过使用提示和场景来协调这些代理的合作。

一些代理网络的例子包括 Flock.ai、Myshell、Theoriq。

推荐：当今大部分代理的功能都相对有限。对于特定用例，Web2 代理能够更好的服务，并且拥有成熟的编排工具，例如 Langchain、Llamaindex。



## 5.代理网络与推理网络的差异

代理网络更侧重于 LLM，提供了如 Langchain 这样的工具来整合多个代理。通常情况下，开发者无需亲自开发机器学习模型，代理网络已经将模型开发和部署的过程简化。他们只需要链接必要的代理和工具即可。大多数情况下，最终用户将直接使用这些代理。

推理网络则是代理网络的基础设施支撑。它提供给开发者较低层次的接入权限。正常情况下，终端用户不直接使用推理网络。开发者需要部署自己的模型，这不仅限于 LLM，并且他们可以通过链下或链上接入点使用它们。

代理网络和推理网络并非完全独立的产品。我们已经开始看到一些竖向整合的产品。他们因为这两种功能依赖相似的基础设施，所以同时提供代理和推理能力。





## 6.新的机会之地

除了模型推理、训练和代理网络外，web3 领域还有很多值得探索的新领域：

数据集：如何将区块链数据转变为机器学习可用的数据集？机器学习开发者需要的是更为具体和专题化的数据。例如，Giza 提供了一些高品质的、关于 DeFi 的数据集，专门用于机器学习训练。理想的数据应不仅仅是简单的表格数据，还应该包括能够描述区块链世界交互的图形数据。目前，我们在这方面还有不足。目前有些项目正通过奖励个人创建新数据集来解决这一问题，比如 Bagel 和 Sahara，它们承诺保护个人数据的隐私。

模型存储：一些模型体积庞大，如何存储、分发及版本控制这些模型是关键，这关系到链上机器学习的性能和成本。在这一领域，Filecoin, AR 和 0g 等先锋性项目已经取得了进展。

模型训练：分布式且可验的模型训练是个难题。Gensyn, Bittensor, Flock 和 Allora 等已有显著进展。

监控：由于模型推理在链上与链下都有发生，我们需要新的基础设施来帮助 web3 开发者追踪模型的使用状况，及时发现可能存在的问题和偏差。有了合适的监控工具，web3 的机器学习开发者可以及时调整，不断优化模型精确度。

RAG 基础设施：分布式 RAG 需要全新的基础设施环境，对存储、嵌入计算和向量数据库有较高的需求，同时要确保数据的隐私安全。这与目前的 Web3 AI 基础设施大不相同，后者大多依赖第三方来完成 RAG，例如 Firstbatch 和 Bagel。

专为 Web3 定制的模型：并不是所有模型都适合 Web3 情景。大多数情况下，需要对模型进行重新训练，以适应价格预测、推荐等具体应用。随着 AI 基础设施的繁荣发展，未来我们期望有更多 web3 本地模型来服务于 AI 应用。比如 Pond 正在开发区块链 GNN，用于价格预测、推荐、欺诈侦测和反洗钱等多种场景。

评估网络：在缺乏人类反馈的情况下评估代理是不容易的。随着代理创建工具的普及，市场上将会出现无数的代理。这就需要有一个系统来展示这些代理的能力，并帮助用户判断在特定情况下哪个代理的表现最佳。例如，Neuronets 是这个领域的一个参与者。

共识机制：对于 AI 任务，PoS 不一定是最佳选择。计算复杂性、验证的困难和确定性的缺失是 PoS 面临的主要挑战。Bittensor 创造了一种新的智慧型共识机制，奖励网络中为机器学习模型和输出做出贡献的节点。



## 7.未来展望

我们目前观察到了垂直整合的发展趋势。通过构建一个基础的计算层，网络能够为多种机器学习任务提供支持，包括训练、推理及代理网络服务。这种模式意在为 Web3 的机器学习开发者们提供全方位的一站式解决方案。

目前，链上推理尽管成本高昂且速度较慢，但它提供了出色的可验证性及与后端系统（例如智能合约）的无缝集成。我认为未来将走向混合应用的道路。一部分推理处理将在前端或链下进行，而那些关键的、决策性的推理则会在链上完成。这种模式已经在移动设备上得到了应用。通过利用移动设备的本质特点，它能够在本地快速运行小型模型，并将更复杂的任务迁移到云端，利用较大的 LLM 处理。