解析 AI 与加密结合的潜力与现实挑战

作者：@ed\_roman；编译：白话区块链

最近，人工智能成为加密市场上最热门、最有前景的领域之一。 包括：

去中心化的AI训练

GPU去中心化物理基础设施网络

无审查的AI模型 这些是突破性的进展还是仅仅是炒作？

在@hack\_vc，我们正努力拨开迷雾，将承诺与现实区分开来。 本文将深入分析加密与AI的顶尖创意。让我们一同探讨真正的挑战与机遇。

## 一、Web3与AI结合的挑战

### 1、去中心化的AI训练

在链上进行AI训练的问题在于，训练需要GPU之间高速的通信和协调，因为神经网络在训练时需要进行反向传播。Nvidia为此提供了两项创新技术（NVLink和InfiniBand）。这些技术可以极大地加快GPU通信速度，但它们只能在单个数据中心内的GPU集群中使用（速度超过50 Gbps）。

如果引入去中心化网络，由于增加了网络延迟和带宽，速度会显著变慢。这与Nvidia在数据中心内提供的高速互联相比，对于AI训练用例来说根本不可行。此外，去中心化环境中的网络带宽和存储成本相比于本地集群中的固态硬盘也要高得多。

在链上训练AI模型的另一个问题是，与推理相比，这个市场的吸引力较小。目前，大量GPU计算资源用于AI大语言模型（LLM）的训练。但从长远来看，推理将成为GPU的主要应用场景。想想看：为了满足需求，需要训练多少个AI大语言模型？相比之下，将有多少客户使用这些模型？

请注意，这方面已经有一些创新，可能为链上AI训练的未来提供希望：

1）基于InfiniBand的分布式训练正在大规模开展，NVIDIA本身也通过其集体通信库支持非本地分布式训练。不过，这仍处于初期阶段，采用情况还有待观察。物理距离带来的瓶颈依然存在，因此本地InfiniBand训练仍然显著更快。

2）已有一些新研究发表，探讨了减少通信同步次数的去中心化训练，可能在未来使去中心化训练更为实际。

3）智能分片和训练调度可以帮助提升性能。同样，未来可能会有新模型架构专门为分布式基础设施设计（Gensyn正在这些领域进行研究）。

4）创新如Neuromesh尝试通过一种称为预测编码网络（PCN）的新方法，以较低成本实现分布式训练。

### 2、去中心化的AI数据迭代

训练的数据信息部分也是一个难题。任何AI训练过程都涉及处理大量数据。通常，模型是在集中且安全的数据存储系统上进行训练，这些系统具有高可扩展性和高性能。这需要传输和处理数TB的数据，而且这不是一次性的循环。数据通常是嘈杂且含有错误的，所以在训练模型之前，必须对数据进行清洗和转换，使其成为可用格式。这个阶段涉及标准化、过滤和处理缺失值的重复任务。在去中心化环境中，这些都构成了严重挑战。

训练的数据信息部分也是迭代的，这与Web3不太兼容。OpenAI花费了成千上万次迭代才取得他们的成果。训练过程是迭代的：如果当前模型未达到预期效果，专家会返回到数据收集或模型训练阶段以改进结果。现在，想象在去中心化环境中进行这个过程，而现有的最佳框架和工具在Web3中不易获得。

一种有前景的技术是0g.ai（由Hack VC支持），他们提供链上数据存储和数据可用性基础设施。他们拥有更快的架构和在链上存储大量数据的能力。

### 3、利用过度冗余的AI推理计算达成共识

加密与AI结合的一个挑战是验证AI推理的准确性，因为你不能完全信任单一的中心化方来执行推理操作，存在节点行为不端的可能性。在Web2的AI中，这个挑战不存在，因为没有去中心化的共识系统。

一种解决方案是冗余计算，即多个节点重复相同的AI推理操作，以便在无信任的环境中操作，并避免单点故障。

这种方法的问题在于，我们生活在一个高端AI芯片严重短缺的世界。高端NVIDIA芯片的等待期长达数年，导致价格上涨。如果你还要求AI推理在多个节点上多次重复执行，这将大幅增加这些昂贵的成本。对于许多项目来说，这是行不通的。

### 4、Web3特定的AI用例（短期内）

有人建议，Web3应该有其独特的AI用例，专门针对Web3客户。

目前，这仍是一个新兴市场，用例尚在发现中。一些挑战包括：

Web3原生用例需要的AI交易量较少，因为市场需求还处于起步阶段。

客户较少，因为Web3客户相比Web2客户少了好几个数量级，所以市场不那么分散。

客户本身不够稳定，因为他们是资金较少的初创公司，因此这些初创公司可能会随着时间的推移倒闭。针对Web3客户的AI服务提供商可能需要随着时间的推移重新获取部分客户，以替代那些倒闭的客户，使其业务扩展更加困难。

从长远来看，我们对Web3原生的AI用例非常看好，特别是随着AI代理的普及。我们设想未来每个Web3用户都会有多个AI代理为他们提供帮助。这个领域的早期领先者是Theoriq.ai，他们正在构建一个可组合AI代理的平台，能够服务于Web2和Web3客户（由Hack VC支持）。

### 5、消费级GPU去中心化物理基础设施网络（DePIN）

有许多去中心化的AI计算网络依赖消费级GPU，而不是数据中心的GPU。消费级GPU适用于低端AI推理任务或延迟、吞吐量和可靠性要求较为灵活的消费用例。但对于严肃的企业用例（即占据主要市场份额的用例），客户希望网络比家庭机器更可靠，且复杂推理任务通常需要更高端的GPU。对于这些更有价值的客户用例，数据中心更为适合。

需要注意的是，我们认为消费级GPU适合于演示用途或那些能容忍较低可靠性的个人和初创公司。但这些客户的价值基本较低，因此我们认为，面向Web2企业的去中心化物理基础设施网络（DePIN）从长远来看会更有价值。因此，知名的GPU DePIN项目通常已经从早期主要使用消费级硬件发展到现在具备A100/H100和集群级别的可用性。

## 二、加密 x AI 的实际且可行的用例

现在，让我们讨论加密 x AI 能显著提升价值的用例。

### 实际收益1：服务于Web2客户

麦肯锡估计，生成式AI每年可为他们分析的63个用例带来2.6万亿至4.4万亿美元的附加价值——相比之下，英国2021年的GDP总量为3.1万亿美元。这将使所有人工智能的影响增加15%到40%。如果我们将生成式AI嵌入目前用于其他任务的软件中，这一估计的价值将大致翻倍。

有趣的是：

根据上述估算，这意味着全球AI（不仅仅是生成式AI）的总市场价值可能达到数十万亿美元。

相比之下，所有加密货币（包括比特币和所有山寨币）加起来的总价值今天仅约为2.7万亿美元。

所以，让我们现实一点：短期内需要AI的客户绝大多数将是Web2客户，因为实际需要AI的Web3客户只是这2.7万亿市场中的一小部分（考虑到BTC占据了一半的市场份额，而BTC本身并不需要/使用AI）。

Web3的AI用例才刚刚起步，目前尚不清楚其市场规模会有多大。但有一点直观上可以确定——在可预见的未来，它只会是Web2市场的一部分。我们相信Web3 AI依然有光明的前景，但这意味着目前Web3 AI最普遍的应用还是服务于Web2客户。

可以从Web3 AI中受益的Web2客户的示例包括：

从头开始构建并以AI为核心的垂直行业软件公司（例如Cedar.ai或Observe.ai）

为自身目的微调模型的大型企业（例如Netflix）

快速增长的AI提供商（例如Anthropic）

在现有产品中加入AI功能的软件公司（例如Canva）

这是一个相对稳定的客户群体，因为这些客户通常规模大且价值高。他们不太可能在短期内倒闭，并且代表了AI服务的非常大的潜在客户群。服务于Web2客户的Web3 AI服务将受益于这一稳定的客户基础。

但为什么Web2客户会想要使用Web3技术栈？本文的其余部分将解释这个理由。

### 实际收益2：通过GPU去中心化物理基础设施网络（GPU DePIN）降低GPU使用成本

GPU DePINs汇集了未充分利用的GPU计算能力（其中最可靠的来自数据中心），并使这些资源可用于AI推理。可以简单地将其视为“GPU的Airbnb”（即协作消费未充分利用的资产）。

我们对GPU DePINs感到兴奋的原因如上所述，主要是因为NVIDIA芯片短缺，目前有许多GPU周期被浪费了，这些资源可以用于AI推理。这些硬件所有者已经承担了沉没成本，当前没有充分利用他们的设备，因此可以以比现状更低的成本提供这些部分GPU周期，因为对硬件所有者来说，这实际上是“意外之财”。

具体例子包括：

1）AWS机器：如果你今天从AWS租用一台H100，你需要承诺至少租用一年，因为市场供应紧张。这会导致浪费，因为你不太可能全年365天、每周7天都使用你的GPU。

2）Filecoin挖矿硬件：Filecoin网络有大量的补贴供应，但实际需求并不大。不幸的是，Filecoin从未找到真正的产品市场契合点，因此Filecoin矿工面临破产的危险。这些机器配备了GPU，可以重新用于低端AI推理任务。

3）ETH挖矿硬件：当ETH从工作量证明（PoW）转向权益证明（PoS）时，大量硬件立即变得可用，这些硬件可以重新用于AI推理。

GPU DePIN市场竞争激烈，有多个参与者提供产品。例如Aethir、Exabits和Akash。Hack VC选择支持io.net，后者还通过与其他GPU DePINs的合作来汇集供应，因此他们目前支持市场上最大的GPU供应。

需要注意的是，并非所有GPU硬件都适用于AI推理。一个明显的原因是较旧的GPU没有足够的GPU内存来处理大语言模型（LLMs），尽管在这方面已经有一些有趣的创新。例如，Exabits开发了技术，将活跃的神经元加载到GPU内存中，而将不活跃的神经元加载到CPU内存中。他们预测哪些神经元需要活跃/不活跃。这使得即使在GPU内存有限的情况下，也能使用低端GPU处理AI工作负载。这实际上提高了低端GPU在AI推理中的实用性。

此外，Web3 AI DePINs需要随着时间的推移强化他们的产品，提供企业级服务，如单点登录（SSO）、SOC 2合规、服务级别协议（SLAs）等。这将与当前Web2客户享受的云服务相媲美。

### 真正的优势 #3：避免OpenAI自我审查的非审查模型

关于AI审查的问题已经有很多讨论。例如，土耳其曾一度暂时禁止OpenAI（后来他们在OpenAI改进其合规性后取消了禁令）。我们认为这种国家级别的审查从根本上说并不值得关注，因为各国需要拥抱AI才能保持竞争力。

更有趣的是，OpenAI会自我审查。例如，OpenAI不会处理NSFW（不适合在工作场合观看）内容，也不会预测下届总统选举的结果。我们认为在OpenAI因政治原因不愿涉及的AI应用领域，存在一个有趣且巨大的市场。

开源是解决这一问题的一个好办法，因为一个Github仓库不受制于股东或董事会。一个例子是Venice.ai，它承诺保护用户隐私并以非审查的方式运作。当然，关键在于其开源性，这使得这一切成为可能。Web3 AI可以有效提升这一点，通过在低成本的GPU集群上运行这些开源软件（OSS）模型以进行推理。正因为如此，我们相信OSS + Web3是铺平非审查AI道路的理想组合。

### 真正的好处 #4：避免向OpenAI发送个人可识别信息

许多大型企业对其内部企业数据存在隐私顾虑。对于这些客户来说，很难信任像OpenAI这样的集中式第三方来处理这些数据。

对于这些企业来说，使用web3可能会显得更加可怕，因为他们的内部数据突然出现在一个去中心化网络上。然而，对于AI而言，隐私增强技术方面已经有一些创新：

诸如Super协议之类的可信执行环境（TEE）

诸如Fhenix.io（由HackVC管理的基金组合公司）或Inco Network（均由Zama.ai提供支持）和Bagel的PPML之类的完全同态加密（FHE）

这些技术仍在不断发展，通过即将推出的零知识（ZK）和FHE ASICs，性能也在不断改善。但长期目标是在微调模型时保护企业数据。随着这些协议的出现，web3可能会成为更具吸引力的隐私保护AI计算场所。

### 真正的好处 #5：利用开源模型的最新创新

在过去的几十年里，开源软件（OSS）一直在侵蚀专有软件的市场份额。我们将LLM视为一种高级专有软件，正逐渐成为开源软件的颠覆对象。一些值得注意的挑战者包括Llama、RWKV和Mistral.ai。随着时间的推移，这个列表无疑会不断增长（在Openrouter.ai上提供了更全面的列表）。通过利用由开源模型提供支持的web3AI，人们可以充分利用这些新创新。

我们相信，随着时间的推移，一个开源的全球开发工作力量，结合加密激励，可以推动开源模型以及构建在其之上的代理和框架的快速创新。一个AI代理协议的例子是Theoriq。Theoriq利用开源模型创建了一个可组合互联的AI代理网络，可以组装在一起创建更高级的AI解决方案。

我们对此深信不疑的原因在于过去的经验：大多数“开发者软件”在经过时间的推移后逐渐被开源软件所超越。微软过去是一家专有软件公司，现在成为了贡献最多的Github公司，这是有原因的。如果你看看Databricks、PostGresSQL、MongoDB等是如何颠覆专有数据库的，就会发现整个行业就是一个被开源软件颠覆的例子，所以先例在这里是相当强大的。

然而，这也有一个小陷阱。OSS LLMs存在一个棘手的问题，就是OpenAI已经开始与组织签订付费数据许可协议，比如Reddit和纽约时报。如果这种趋势持续下去，由于获取数据的经济壁垒，OSS LLMs可能会越来越难以竞争。英伟达可能会将保密计算作为安全数据共享的加强工具。时间会告诉我们这将如何发展。

### 真正的好处 #6：通过高成本的随机抽样或零知识证明实现共识

在web3AI推理中，验证是一个挑战。验证者有可能通过欺骗结果来获取费用，因此验证推理是一项重要的措施。需要注意的是，尽管AI推理还处于初级阶段，但除非采取措施来削弱这种行为的动机，否则这种欺骗是不可避免的。

标准的web3方法是让多个验证者重复相同的操作并进行结果比较。然而，正如前面提到的，由于当前高端Nvidia芯片短缺，AI推理非常昂贵。考虑到web3可以通过未充分利用的GPU DePINs提供更低成本的推理，冗余计算将严重削弱web3的价值主张。

更有希望的解决方案是对离链AI推理计算进行零知识证明。在这种情况下，可以验证简明的零知识证明以确定模型是否经过正确训练，或者推理是否正确运行（称为zkML）。其中的示例包括Modulus Labs和ZKonduit。由于零知识操作需要相当大的计算资源，这些解决方案的性能仍处于初级阶段。然而，随着零知识硬件ASIC在不久的将来推出，这一情况可能会得到改善。

更有希望的想法是一种“乐观”抽样为基础的AI推理方法。在这种模型中，您只需验证验证者生成结果的一小部分，但设置足够高的经济成本来惩罚被抓到作弊的验证者，从而产生强大的经济禁止效应。这样一来，您可以节省冗余计算（例如，参见Hyperbolic的"Proof of Sampling"论文）。

另一个有希望的想法是使用水印和指纹技术的解决方案，例如Bagel Network提出的解决方案。这类似于亚马逊Alexa为其数百万设备上的AI模型质量保证提供的机制。

### 真正的好处 #7：通过可组合的开源软件堆栈节省费用（OpenAI的利润）

web3为AI带来的下一个机会是降低成本的民主化。到目前为止，我们已经讨论了通过像io.net这样的DePINs节省GPU成本的方法。但是，web3还提供了节省中心化web2 AI服务的利润率（例如OpenAI，根据本文撰写时的信息，其年收入超过10亿美元）的机会。这些成本节约来自于使用开源软件（OSS）模型而不是专有模型，从而实现了额外的成本节约，因为模型创建者并不试图盈利。

许多开源软件模型将始终完全免费，这为客户提供了最佳的经济效益。但是，也可能有一些开源软件模型尝试这些变现方法。请考虑，Hugging Face上仅有4%的模型由有预算的公司进行训练以帮助补贴这些模型（参见此处）。剩下的96%的模型是由社区进行训练的。这个96%的Hugging Face模型群体面临着实际的成本（包括计算成本和数据成本）。所以这些模型需要以某种方式实现变现。

有许多关于实现这种开源软件模型变现的提议。其中最有趣的之一是“初始模型发行”（IMO）的概念，即将模型本身进行Token化，留下一部分Token给团队，并将模型的一些未来收入流向Token持有人，尽管这其中显然存在一些法律和监管障碍。

其他开源软件模型将尝试基于使用量进行变现。需要注意的是，如果这种情况变为现实，开源软件模型可能开始越来越像它们的web2利润生成对应物。但是，从现实角度来看，市场将会二分，其中一些模型将完全免费。

一旦选择了开源软件模型，您可以在其上进行可组合的层次操作。例如，您可以使用Ritual.net进行AI推理，以及Theoriq.ai作为可组合和自治的链上AI代理的早期领导者（两者都得到了Hack VC的支持）。

### 真正的好处 #8：去中心化的数据采集

AI面临的最大挑战之一是获取适合训练模型的正确数据。我们之前提到过，去中心化AI训练存在一些挑战。但是利用去中心化网络来获取数据（然后可以在其他地方，甚至是传统的web2平台上用于训练）又如何呢？

这正是像Grass这样的初创公司正在做的事情（得到了Hack VC的支持）。Grass是一个去中心化的“数据爬取”网络，由个人贡献他们机器的闲置处理能力来获取数据，以供AI模型的训练。理论上，在大规模应用中，这种数据采集可能比任何一家公司的内部努力更优越，因为庞大的激励节点网络具有强大的计算能力。这不仅包括获取更多的数据，还包括更频繁地获取数据，以使数据更具相关性和最新性。由于这些数据爬取节点本质上是分散的，不属于单个IP地址，因此几乎不可能阻止这个去中心化的数据爬取军团。此外，他们还有一支人力网络，可以清理和规范数据，使其在被爬取后变得有用。

一旦获取了数据，您还需要一个链上的存储位置，以及使用该数据生成的LLM（大型语言模型）。在这方面，0g.AI是早期的领导者。它是一个针对AI进行优化的高性能web3存储解决方案，比AWS便宜得多（这对于Web3 AI来说是另一个经济上的成功），同时也可以作为第二层、AI等的数据可用性基础设施。

需要注意的是，在未来，数据在web3AI中的作用可能会发生变化。目前，对于LLM来说，现状是使用数据对模型进行预训练，并随着时间的推移使用更多的数据进行改进。然而，由于互联网上的数据实时变化，这些模型始终略微过时，因此LLM推理的响应略有不准确。

未来可能发展的一个新范式是“实时”数据。这个概念是当LLM被要求进行推理时，LLM可以通过向其注入实时从互联网上收集的数据来使用数据。这样，LLM将使用最新的数据。Grass也正在研究这一点。

## 三、结论

我们希望这篇分析对您在思考web3 AI的承诺与现实时有所帮助。这只是一个讨论的起点，而且这个领域正在迅速变化，所以请随时加入并表达您的观点，因为我们愿意继续共同学习和建设。