一文读懂Crypto x AI新星Hyperbolic

作者：Josh Ho & Teng Yan，Chain of Thought；翻译：本站xiaozou

本文，我们将深入研究Hyperbolic，一款当前大热的开放获取AI云服务。Hyperbolic的宏大使命是通过提供负担得起的推理计算能力，让AI变得更加普及。

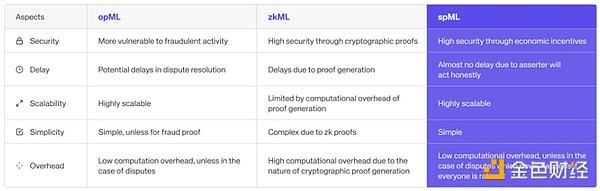
但在此之前，让我们先来看看我们所认为的关于Hyperbolic最有趣的一些事……

## 1、Hyperbolic的秘笈——抽样证明（Proof of Sampling）

Hyperbolic正在通过解决人工智能领域最棘手的挑战之一来开疆扩土：验证一个输出是否真正来自特定的AI模型。

这个问题对于像OpenAI这样的集中式闭源提供商来说尤其棘手。在你请求GPT-4提供输出的时候，你怎么能确定你没有被欺骗——比如，OpenAI运行更便宜的GPT-3.5模型（每代币价格的1/20）？

目前，这样的保证还依赖于声誉，但Hyperbolic认为应该以一种无需信任的去中心化方式来处理这个问题。



目前有几种方法可以做到这一点：

·Optimistic机器学习（OpML）：假设所有事务都是有效的，除非有验证者（validator）提出质疑。

·零知识机器学习(zkML)：使用ZK电路来验证计算是否被正确执行。

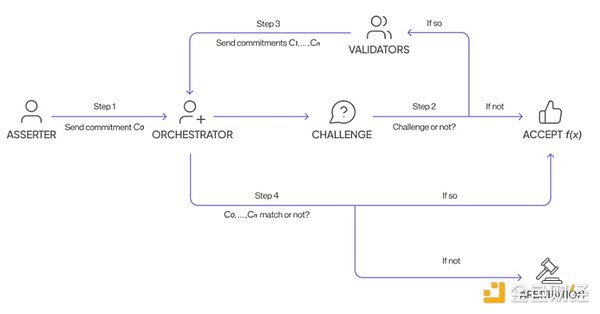
然而，两者都具有局限性：

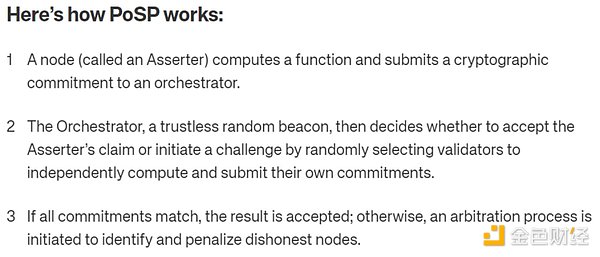
·OpML依赖于validators来查验结果，这就会因争议期而延迟最终确定性。另外，缺少内在激励来确保validators的诚信行为。

·zkML的计算量非常大，有时甚至需要几天的时间才能生成具有70B+参数的大模型证明。

Hyperbolic的目标是通过其抽样证明（PoSP）协议和抽样机器学习（SpML）来克服这些缺点。SpML利用抽样和博弈论来鼓励诚信行为，而不需要进行持续不断的监督。

它基于一种被称为Nash Equilibrium（纳什均衡）的纯策略型博弈论概念，即所有参与者都有明确的动机诚信行事，因为作弊的成本超过了潜在的收益。





最简单的思维模式是把它想象成一个公共汽车售票系统。

检票员只进行随机检查，所以你可能会认为乘客经常会冒险逃票。但令人惊讶的是，他们并没有这样做，因为逃票的惩罚力度足以阻止乘客作弊。只要罚款远远超过购票成本，诚信就会占上风。

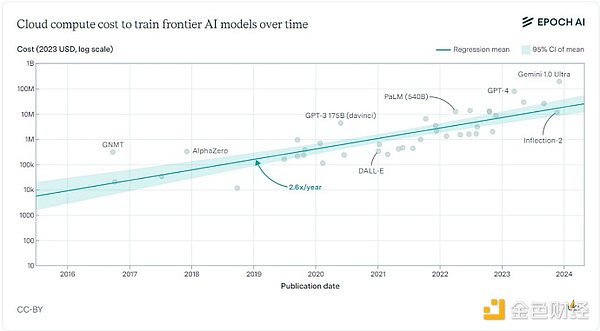
Hyperbolic的SpML使用经济激励来解决当前的验证机制（如OpML和zkML）的局限性。它既提供了速度又提供了安全性，在二者之间达到了很好的平衡，并且没有沉重的计算负担。

需要注意的是什么？那就是它假设每个人的行为都是理性的，但事实并非总是如此。

如果SpML在实践中运行良好，它将改变去中心化AI应用的游戏规则，使无需信任的验证推理成为现实。

## 2、可扩展的低成本计算

训练AI是很昂贵的。电力和计算接入是企业和初创公司面临的最大成本。训练模型所需的算力成本几乎每9个月翻一番。

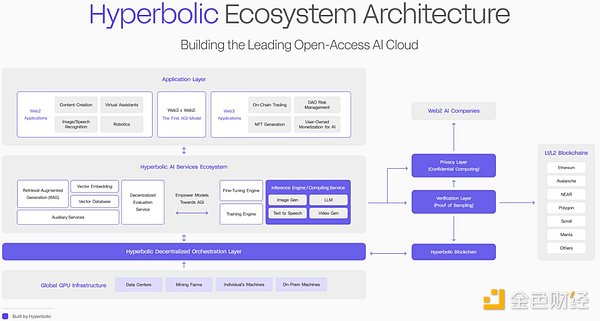


2020年GPT-3的成本约为400万美元，而2023年GPT-4的计算训练成本高达1.9亿美元，让人惊掉下巴。

只有资源充足的机构组织才能活下去。规模较小的参与者和爱好者被过高的成本挤出市场。斯坦福大学的一名博士后不得不停止他的研究，原因是他负担不起所需的数千个GPU。

去中心化计算网络的一个主要挑战是管理异构硬件——不仅仅是顶级的英伟达芯片，还有各种各样的GPU。

Hyperbolic的去中心化操作系统是其计算网络的核心。它将通过内置的自动扩展和容错性无缝地汇集资源。



Hyperbolic的突破在于它对这种复杂性的处理方式。

·它通过优化不同硬件（从Nvidia到AMD GPU）的张量运算提供灵活性。

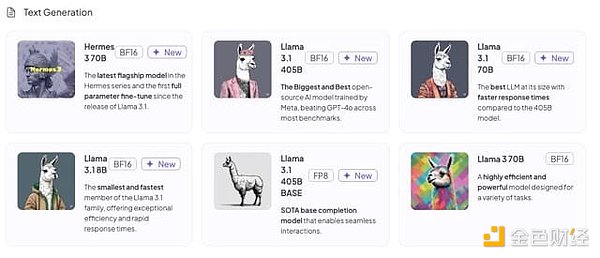
·Hyperbolic的编译栈抽象了复杂性，使开发人员能够在不同的GPU设置中实现高性能，而不会陷入部署和配置方面的困境。

其他市场可能会提供去中心化GPU，但它们通常缺乏Hyperbolic所能提供的精密优化，从而将性能调优的负担放在用户身上。

Hyperbolic通过一个API简化了这一点，该API提供了对针对各种硬件优化的AI模型的访问权，使全球计算资源更加易于访问。

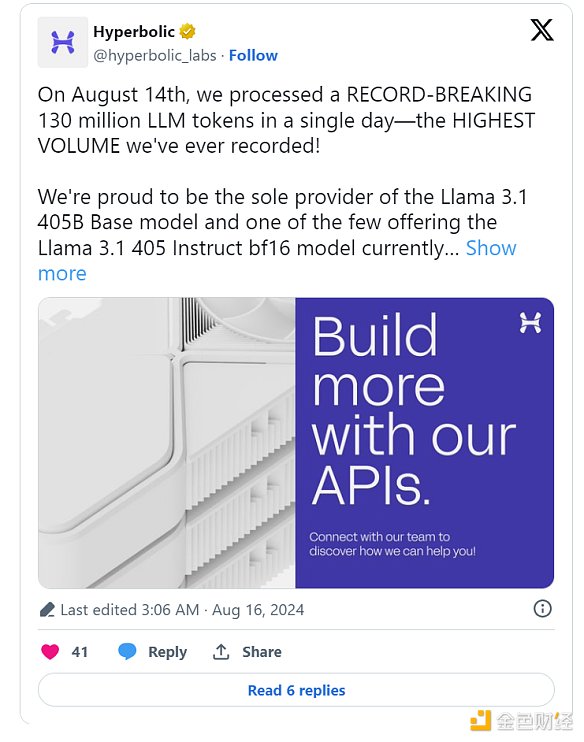
8月15日，Hyperbolic发布了其GPU市场的alpha受限版本，允许100名等候成员试用GPU租赁功能。

## 3、AI服务层



Hyperbolic人工智能生态系统的下一个组件就是AI服务层，它提供推理、模型训练、模型评估和检索增强生成（RAG）等功能。

在Hyperbolic app中，你可以轻松运行顶级开源模型，如Llama 3.1 405B和Hermes 370B。要想微调输出，你可以调整超参数，如max tokens、temperature和top P。



Hyperbolic平台为创新的人工智能应用打开了大门，包括：

·AI智能体收入共享：代币化AI智能体的所有权以重新分配收入。

·人工智能DAO：利用人工智能进行治理决策。

·分割GPU所有权：允许用户拥有和交易部分GPU。

## 4、Crypto扮演什么角色？

Hyperbolic基础设施的核心是其区块链，它支撑着编排、服务和验证层。区块链为Hyperbolic的开源AI云处理结算和治理。它还支持PoSP技术的仲裁和验证机制。

虽然关于该区块链的具体内容还很少，但你大可期待Hyperbolic很快就会透露更多关于这方面的信息。

## 5、研究级Alpha

Hyperbolic仍处于测试网阶段。他们在由Polychain Capital和Lightspeed Faction领投的种子轮融资中筹集了700万美元资金。

有意思的是，Hyperbolic是Llama 3.1 405B Base模型的独家提供商。

Base模型是LLM的初始预训练版本，没有经过微调或通过人类反馈(RLHF)进行强化学习。它具有如下优势：

·全方位支持对特定任务的微调

·是高级人工智能技术的起点，例如合成数据生成或模型蒸馏（model distillation）。



## 6、关于团队

Jasper (Yue) Zhang博士是Hyperbolic Labs的联合创始人兼首席执行官。他曾是Ava Labs的高级区块链研究员和Citadel Securities的量化研究员。他在两年内读取了加州大学伯克利分校的数学博士学位，并在阿里巴巴全球数学竞赛和中国数学奥林匹克竞赛中均获金牌。

Yuchen Jin博士是Hyperbolic Labs的联合创始人兼首席技术官。他拥有华盛顿大学计算机系统和网络博士学位。他之前曾在OctoML工作，这是一家为运行、调整和扩展生成式AI应用程序提供基础设施的公司。

## 7、我们的几点想法

总的来说，Hyperbolic让我们非常兴奋。他们绝对是Crypto AI领域里最值得关注的团队之一。

Hyperbolic不仅仅是一个算力提供商，像PoSP和SpML这样的创新也为去中心化AI增加了新的信任和验证层。

在Hyperbolic上试验base模型是非常有趣的，特别是因为他们是目前为数不多的能够实现这一功能的提供商之一。我们绝对可以相信支持他们的开源AI承诺。

几周前，我们写过关于Prime Intellect的文章。Hyperbolic是否会像Prime Intellect那样专注于分布式AI训练，这一点还有待观察。

虽然我们注意到对算力的需求通常是稀疏的，但对于Hyperbolic来说似乎并非如此。他们已经在研究市场上显示出了早期吸引力，吸引了研究人员和开发人员的极大兴趣。