

Poseidon

通过子网将数字经济引入链中

LitePaper v1

2025年10月

Poseidon AI

Story Foundation

总结

我们介绍了Poseidon，这是一个去中心化的平台，旨在促进协作创建、收集和管理对下一代人工智能开发至关重要的专业、长尾、权限明确的训练数据。Poseidon是一个特定于应用程序、高效且可扩展的数据管道集合，任何人都可以在其中贡献AI训练数据，并参与数据质量的处理和评估。在引擎盖下，Poseidon在专门的子网基础上运行，每个子网都针对特定的人工智能领域及其独特的数据处理要求进行定制。这些子网建立在Story现有的L1基础设施之上，以跟踪数据来源和知识产权（IP）谱系，共同形成一个开放的市场，旨在解决人工智能训练数据的供需挑战。

介绍

随着基础模型的初步成功，人工智能行业正在经历价值增长的重大转变。在这个堆栈中，人工智能有三个具有累积价值的核心部分：计算、模型和数据。计算几乎被少数几家规模较大的公司垄断，如英伟达（Nvidia）和AMD，它们优化了芯片生产和分销。虽然成本高昂，但这在很大程度上是资金充足的组织可以解决的资本支出问题。随着半衰期的迅速减少，模型架构已变得极具竞争力。

OpenAI的GPT突破很快被Anthropic的Claude赶上，然后被DeepSeek和Mistral开源。曾经提供多年竞争优势的东西现在只提供几个月的时间，因为架构创新几乎立即在研究社区中传播。然而，数据是制约人工智能进步的关键瓶颈。与计算（随资本而扩展）或模型（快速扩散）不同，独特的训练数据集不容易复制或购买。竞争优势在于访问现有参与者无法简单获取或生成的专业、长尾、多模态数据，从罕见的边缘案例场景到特定领域的多模态记录。这类数据可能千差万别，从用于自动驾驶的建筑工地上不常见的街景，到呼叫中心为转录任务收集的嘈杂语音。

数据可用性已从根本上从丰富转变为稀缺。对于第一代人工智能模型来说，互联网提供了看似无限的文本和图像，可以自由地

刮伤了。那个时代已经结束了。大多数可公开访问的互联网资源，如CommonCrawl，已经耗尽用于人工智能训练。

训练先进的物理人工智能模型需要特定的、真实世界的数据，而这些数据很难获得。这些非常有价值的数据要么驻留在特定领域的组织中，缺乏基础设施来有效地将其货币化，要么是需要生成的不存在的数据。基础模型公司既需要规模（来自YouTube等平台），也需要专一性（就像人类组装PCB或用罕见的方言进行呼叫中心对话）。然而，大多数可用数据缺乏人工智能使用的适当许可，仅有互联网数据不足以满足这些专业应用。想象一下这样一个系统，基础模型公司可以立即利用全球范围内的各种培训数据，每个人都已经产生了有价值的数据。

日常运营中的长尾数据。只要他们的数据被使用，这些供应商就可以无缝地许可、出售或接收自动的“数据红利”微支付。

迫切需要这样一个系统，但由于三个挑战，这个问题仍未解决：

1. **供需匹配**：没有可扩展的机制将人工智能公司与全球数据供应商联系起来。
2. **知识产权和数据来源**：追踪经过版权许可的数据的来源仍然很困难，这给使用权带来了不确定性，以及
3. **数据估值**：没有为数据贡献者建立估值机制，以做出关于将其数据货币化的明智决策。

这些挑战为构建协调数据收集、验证起源并支持跨不同利益相关者（从领域专家和贡献者到注释者）的可扩展许可的系统创造了机会。Poseidon通过创建基础设施来协调人工智能数据受限的未来的数据供应和需求，从而解决了这一差距。

Poseidon的接近

Poseidon通过创建开放、分散和可扩展的数据层来协调数据供应和需求之间的流动，从而解决人工智能训练数据的挑战。其多层架构的设计以可扩展性、灵活性和可重用性为核心，支持不同的人工智能应用领域和协作经济模型。在更高的层次上，Poseidon提供了可适应的数据管道，用于管理数据集创建和激励机制的整个生命周期，使系统能够根据不断变化的数据需求进行动态调整。在较低的层次上，Poseidon依赖于一个模块化的、可重用的基础设施，该基础设施支持跨AI领域的的数据收集和协调。为了平衡效率 and 专业化，这个基础设施被组织成专门构建的碎片（子网），每个碎片都针对特定的AI域进行了优化。这种分片方法可确保资源得到有效利用，防止出现瓶颈，并允许根据不同领域的独特要求进行性能调整，同时仍能从共享基础层的一致性中受益。

数据管道

Poseidon的数据管道能够聚合、处理和准备高保真、生产就绪和权限清除的数据集，这些数据集是为各种人工智能应用程序定制的。

每个数据管道都基于指定数据要求和综合数据集构建过程（聚合、处理和注释）的工作流进行操作。任何供应方用户都可以参与数据管道，并承担工作流程中的各种角色，包括：

- **数据提供者**：根据指定的数据请求收集和传递数据。这包括各种数据格式，例如音频、视频和其他相关元数据。
- **数据处理**：对数据运行已定义的确定性过程。
- **数据标注**：根据规范对一条数据进行人工标注（人在回路）。

在本文中，“用户”是指除个人之外的更广泛的实体，包括分散式物理基础设施网络（DEPIN）应用程序、数据注释平台和其他类似实体。

每个工作流程都采用各种工具和技术来实现定义的质量目标。其中一些关键目标包括：

多样化的数据收集

人工智能模型在“长尾问题”中苦苦挣扎，它们在常见场景中表现良好，但在同样重要的罕见边缘情况下却失败了。例如，机器人模型可能擅长拿起杯子，但在不寻常的物体或照明条件下会遇到困难。

Poseidon的基础设施使数据收集者能够处理未充分表达的数据并将其货币化，从而解决这一问题。我们的平台为收集者提供工具，以识别、验证和打包AI公司需要的罕见场景。如果DePin网络在雨中收集了施工区的录像，Poseidon的基础设施会帮助他们处理这些数据，验证其质量，并与自动驾驶汽车公司联系，这些公司将为这种稀缺的训练数据付费。

严格的验证

高质量的数据对于训练人工智能至关重要，因为糟糕、嘈杂或不完整的数据会导致有偏差和不可靠的模型，而粗心或“懒惰”的数据准备等挑战会加剧错误并严重限制模型性能。为了确保数据的质量，Poseidon的基础设施提供了先进的验证工具，包括自动重复数据删除、标准化和过滤，以及AI支持的PII删除。除了基本的预处理，该平台还提供特定领域的验证，如验证音频质量和扬声器一致性

语音数据，或使机器人公司能够根据其特定的培训要求定义自定义验证标准。该平台灵活的元数据结构允许购买者指定其确切的数据标准，验证工作流程会自动执行这些购买者定义的范围。

可扩展的数据注释

原始数据本身的价值有限，大多数训练数据集需要标签和注释才能用于人工智能训练。然而，现代生成人工智能模型所需的数据规模使得纯粹的人类注释变得越来越昂贵和耗时。由于在不同的使用案例中需要数百万个数据点，因此自动标记对于增强高质量的人工注释至关重要。

Poseidon的数据管道通过混合注释系统解决了这一扩展挑战。

流水线使人工智能辅助的预标记能够处理批量处理，

人在回路验证侧重于最关键或不确定的情况。注释任务可以分布在多个验证器之间，并使用基于共识的标签来确保准确性。对于有特定要求的买家，管道支持自定义注释工作流程和主动学习系统，这些系统可自动发送不确定的样本，以进行额外的人工审查。

权限已清除

对于需要确定使用权限和数据来源的人工智能公司来说，具有可验证来源的权限清除数据是必不可少的。传统的数据市场缺乏透明的来源跟踪，使得买家和卖家很难找到兼容的许可条款。

Poseidon的数据管道将数据集注册为Story的区块链（“ Story ”）上的IP资产，利用Story的可编程IP许可证（PIL）。这种链上注册建立了可验证的出处，这使得具有特定许可要求的买家能够轻松地提供兼容条款的卖家那里发现数据集，例如，买家是否需要商业使用权、归属要求或地理限制。随着时间的推移，流行的偏好类别演变为元数据标准。

子网络

数据管道部署在有目的的分片基础设施上，该基础设施由子网组成，这些子网由单个总体网络同步和保护。每个碎片都专门用于特定的AI领域，支持特定领域的优化，同时保持整个系统的互操作性。

每个子网都运行自己的数据管道、经济系统和验证机制，同时利用通用Poseidon处理管道技术和Story的共享基础设施来实现安全性、互操作性和IP管理的核心实用程序。

单独的子网是必要的，因为不同的AI域具有根本不兼容的基础设施要求。例如，聚合医疗数据需要高度特定的配置，通常包括安全的加密数据传输、强大的访问控制和身份验证、患者隐私措施（如匿名化或假名化）以及广泛的审核以监控数据的使用和更改。为了防止对数据的直接访问，处理通常在可信执行环境（TEE）中进行。

同时，由于每个大型视频训练数据点需要大量的自动化处理和注释，机器人应用需要高带宽。通过高带宽机器人基础设施运行医疗数据将产生不必要的隐私风险，而强制机器人数据通过TEE安全的医疗基础设施将产生令人望而却步的延迟和成本。

为了确保共享网络层的可靠性、可验证性和分散性，Poseidon建立在Story强大的基础设施之上。子网验证器将IP令牌放在Story上，这使得能够使用有限数量的高性能服务器，而不会影响分散化。这些IP-STAKED服务器执行子网的协议，并维护故事的公共可观察性和可挑战性。

除了安全和去中心化方面，Poseidon子网络还将故事用于其他特定目的。他们利用Story的分散式知识产权（IP）基础架构。这允许将数据集协作注册为可编程IP，实现诸如数据来源、许可框架、自动版税支付、随机性来源以及可选的数据安全存储等功能。此外，Story的IP Vault允许将机密文件与IP资产一起安全存储在链上，许可证持有者在获得IP许可证时可以自动访问数据。

这种设计允许Poseidon子网络具有灵活性，并专注于创建可扩展和分散的数据管道，其中具有不同角色的不同用户集可以参与数据操作并共享新形成的数据集的所有权。

示例—音频转录子网络

正如前面所讨论的，Poseidon能够部署优化的子网，并在其上动态执行工作流。为了进行说明，让我们通过部署在子网上的示例工作流来检查数据集的过程。

子网设置

在这个例子中，子网运营商建立了一个优化的子网，为领先的人工智能语音模型公司收集和准备高保真语音和音频数据。为了做到这一点，子网运营商通过可卷起的智能合约锁定了故事区块链上的抵押品，以确保诚实行为。子网处理链外的交易或数据操作，将其聚合成批，并定期将加密承诺（如Merkle根）上传到Roll-Up智能合约进行结算和终结。

接下来，子网运营商设置存储层，该存储层保证可用性并被优化以处理各种音频格式的大数据文件的高吞吐量存储和检索。该过程的最后一步是子网运营商部署和运行将执行第一个工作流的计算引擎。请注意，子网为数据协作创建了一个经济模型，下面的工作流是工作流的一个示例——多个工作流可能部署在同一个子网上。

工作流设置

该工作流作为软件部署在子网之上，负责调度和执行各个步骤，不同的用户可以参与各个阶段。子网运营商可以通过设置参数（如奖励/惩罚机制）和强制执行元数据结构来决定如何确保质量。

数据采集步骤

该流程从定义数据要求并为用户准备一组数据收集任务的工作流开始。然后，用户可以选择参与并将其数据上传/提交到工作流。数据请求是专门设计的，通过涵盖不同的条件来解决数据不平衡问题。具有不同特征的数据点可能会产生不同的奖励值，这些奖励值将由子网运营商确定。在音频抄本工作流程的情况下，不同的语言和不同的上下文（例如，特定域中的文本）可能会收到不同的完成奖励。

数据验证步骤

在接收到数据之后，执行几个数据验证过程，以确保数据不是重复的、不是AI生成的，并且满足由子网运营商设置的所需特性。为实现此目标，此工作流使用基于共识的方法进行验证，其中每个数据点被随机分配给较小的用户子集（在提交前未知），并需要所有这些用户的批准。如果任何用户提交了对数据的关注，则会将更大的用户组分配给验证任务，其中多数投票确定数据是否有效。提供无效数据的数据提供者将失去奖励，并可能受到其他形式的处罚（在某些情况下，他们的赌注可能会被大幅削减）。在此工作流程中，数据验证是通过哈希技术的重复数据删除来完成的，然后是数据的自动转录以及与数据的引用部分的相似性比较。

数据处理步骤

一旦验证了一个数据点，它就会移动到管道中的下一个步骤，在该步骤中，将在顶部执行注释或预处理。对于该工作流程，不需要注释，但是执行一组预定过程来预处理和准备训练数据（例如，转码、去噪和离群点去除）以用于接下来的步骤。

IP注册步骤

一旦完成数据点的收集和验证，数据集将被打包并注册为Story Network上的知识产权资产。然后，它可以在一个开放的市场上上市，允许人工智能应用程序竞标许可证并获得访问数据的权限。

由于许多工作流程重复使用相似的步骤，Poseidon提供了一组工作流程模块，可用于为不同的应用程序构建新的工作流程。有关这些核心模块的概述，请参见附录A。

结论

Poseidon通过将数据转换为可编程、协作和可验证的资源，为下一个人工智能时代建立了一个分散的基础。通过其子网、模块化工作流程以及与Story的IP基础架构的集成，它能够大规模创建和管理专门的、正确清除的数据集，解锁组织中当前孤立的数据，并生成新的、特定于任务的资源，这些资源对于前沿应用程序至关重要。通过调整激励措施和确保来源，Poseidon将自己定位为新兴数字经济的基础设施和市场，为开放和可持续的人工智能发展提供支持。

致谢

作者衷心感谢Story Foundation和PIP Labs的许多团队成员做出的有见地的贡献和富有启发性的讨论。作者还感谢A16Z团队的Scott Kominers、Eddy Lazzarin、Andrew Tretyakov、Markos Georgiades和Noah Citron提供的宝贵反馈和见解。这是一个早期的草案，作者将感谢任何有兴趣提供意见或建议的人的反馈。

附录A—工作流的核心构建模块

Poseidon提供了一系列可用于发布工作流的工作流模块。

安全数据存储模块

安全数据存储模块便于数据提供者安全地上传和控制对其数据的访问。它可以与各种存储提供商集成，并提供对其他模块的按需数据访问（例如，用于验证任务）。本模块只关注数据机密性，并将可用性保证委托给数据层。可以通过数据提供者的积极参与或通过利用密钥管理协议来配置该操作。该模块与IPVault完全兼容，无需不断重新加密数据或在存储位置之间传输数据。

数据验证模块

数据验证模块对于数据工件的安全和公平验证是必不可少的。通过验证任务在用户之间的随机和冗余分布来确保数据质量。它的运行方式是将数据片段随机分配给一组冗余的工作人员，然后这些工作人员对链外数据执行预定义的验证任务。在末端运行专门的共识机制，在链上比较对结果的承诺。为防止工人串通，事先不透露随机性和分配细节。

这促进了社区驱动的质量控制，坚持高标准并阻止低于标准的贡献。Poseidon利用Trident Consensus进行验证，而不是将每个任务分配给每个用户并比较结果，这是一种具有高复制成本和有限可扩展性的方法。三叉戟共识是一个两步链上共识过程，需要一个无偏见和不可预测的随机性来源。在第一步期间，任务被分配给随机工人的较小子集（“ n ”中的“ m ”），并且需要所有工人一致同意（一个诚实节点假设）。如果任何工作人员报告不同的结果，则合意移动到第二个步骤，并且工作流现在将相当大的一组工作人员分配给任务（升级路径），并需要多数投票来确定最终状态。

这种设计确保了可扩展的验证过程。初始工作者分配保证至少一个诚实节点，其中分配的数量是基于工作流要求来平衡成本和数据质量的灵活参数。让我们考虑一个有' N '个工作者的场景，其中最多' $N/3$ '可能是恶意的，并且随机性不是预先确定的。验证错误结果的成功共谋概率为 $(1/3)^M$ ，其中' M '表示分配的数据验证器的数量。可以根据所需的数据质量灵活地调整参数。例如，如果目标是成功验证无效数据的机会小于或等于 10^{-3} ，则工作流生成器将需要分配7个随机工作人员。这个比例可以进一步调整，以在性能和质量之间取得平衡。例如，如果Workflow Builder将“ M ”设置为14，则成功攻击的可能性将降至 10^{-6} 以下。未能正确完成任务或触发第二阶段共识的用户可能会失去奖励，并最终被大幅削减股份。

安全自动处理模块

该模块通过用户提供的可信执行环境（TEE）实现安全的离链自动处理。TEE是隔离的、硬件支持的环境，通过防止未经授权的访问来确保计算过程中的数据隐私。该模块旨在与各种验证机制（主要是Trident验证模块）灵活集成。可选地，该模块可以被设置为采用高级密码证明，例如零知识证明（ZK-Proof），以在不暴露底层数据的情况下验证离链计算的准确性和完整性。例如，利用该模型的工作流可以结合ML模型来区分由ML模型生成的数据和由用户唯一捕获的数据。一个单独的过程可以评估数据点的新颖性。

数据保护模块

数据保护模块通过使数据上传者能够在其数据中嵌入指纹来解决整个数据管道中的数据泄漏问题。该模块采用最先进的加密技术和激励机制来防止数据从工作流中泄漏。它为数据点生成冲突安全指纹，这对于在人工注释任务期间识别数据泄漏至关重要，尤其是当这些任务无法在可信执行环境（TEE）中得到保护时。此外，该模块的高级加密技术甚至可以抵抗旨在发现和破解这些指纹的用户共谋。

IP注册模块

该模块将最终数据集注册为Story上的IP资产，其中数据访问密钥也附加到IP保险库，并且可以由注册IP资产的未来许可证持有人访问。