

Poseidon

Bringing Data Economies Onchain via Subnetworks

Litepaper v1.0
October 2025

Poseidon AI

Story Foundation

Summary

We introduce **Poseidon**, a decentralized platform designed to foster the collaborative creation, collection and curation of specialized, long-tail, rights-cleared training data crucial for the next generation of AI development. Poseidon is a collection of application-specific, efficient, and scalable data pipelines, where anyone can contribute AI training data and engage in processing and evaluation of data quality. Under the hood, Poseidon operates on a foundation of specialized subnetworks, each tailored to specific AI domains and their unique data handling requirements. These subnetworks are built atop Story's existing L1 infrastructure to track data provenance and intellectual property (IP) lineage, collectively forming an open marketplace designed to address the challenges of the supply and demand of AI training data.

Introduction

The AI industry is experiencing a significant shift in where value accrues following the initial success of foundation models. There are three core parts of the stack where AI has accrued value: compute, models, and data. **Compute** is nearly monopolized by a handful of scaled players like Nvidia and AMD who have optimized chip production and distribution. While expensive, it's largely a capital expenditure problem that well-funded organizations can solve. **Model** architectures have become hyper-competitive with rapidly diminishing half-lives. OpenAI's GPT breakthrough was quickly matched by Anthropic's Claude, then open-sourced by DeepSeek and Mistral. What once provided years of competitive advantage now offers only months, as architectural innovations spread almost immediately through the research community. **Data**, however, represents the critical bottleneck constraining AI progress. Unlike compute (which scales with capital) or models (which diffuse rapidly), unique training datasets cannot be easily replicated or purchased. The competitive advantage lies in accessing specialized, long-tail, multi-modal data that existing players cannot simply acquire or generate, from rare edge-case scenarios to domain-specific multi-modal recordings. Such data can vary widely, from uncommon street scenes at construction sites for autonomous driving to noisy speech collected in call centers for transcription tasks.

Data availability has fundamentally shifted from abundance to scarcity. For the first generation of AI models, the internet provided seemingly unlimited text and images that could be freely

scraped. That era is over. Most publicly accessible internet resources like CommonCrawl have already been exhausted for AI training.

Training advanced physical AI models requires specific, real-world data that's difficult to source. This critically valuable data either resides in domain-specific organizations lacking infrastructure to monetize it effectively, or is non-existent data that needs to be generated. Foundation model companies need both scale (from platforms like YouTube) and specificity (like humans assembling PCBs or call center conversations in rare dialects). However, most available data lacks proper licensing for AI use, and internet data alone is insufficient for these specialized applications. Imagine a system where foundation model companies could instantly tap into diverse supplies of such training data worldwide, each of whom already generate valuable, long-tail data in their daily operations. These suppliers could seamlessly license, sell, or receive automatic "data dividend" micropayments whenever their data is used.

There is an acute need for such a system, but it remains unsolved due to three challenges:

1. **Matching Supply and Demand:** No scalable mechanism exists to connect AI companies with global data suppliers,
2. **IP Rights and Data Provenance:** Tracking provenance for rights-cleared data remains difficult, creating uncertainty around usage rights, and
3. **Data Valuation:** No established valuation mechanisms exist for data contributors to make informed decisions about monetizing their data.

These challenges create an opportunity to build systems that coordinate data collection, verify provenance, and enable scalable licensing across diverse stakeholders, from domain experts and contributors to annotators. Poseidon addresses this gap by creating infrastructure that coordinates data supply and demand for AI's data-constrained future.

Poseidon's Approach

Poseidon addresses the challenges of AI training data by creating open, decentralized, and scalable data layers that coordinate the flow between data supply and demand. Its multi-layer architecture is designed with scalability, flexibility, and reusability at its core, supporting diverse AI application domains and economic models for collaboration. At the higher level, Poseidon provides adaptable data pipelines that manage the complete lifecycle of dataset creation and incentive mechanisms, enabling the system to dynamically adjust to evolving data requirements.

On the lower level, Poseidon relies on a modular, reusable infrastructure that supports data collection and coordination across AI domains. To balance efficiency and specialization, this infrastructure is organized into purpose-built shards (subnetworks), with each shard optimized for a specific AI domain. This sharded approach ensures that resources are used efficiently, prevents bottlenecks, and allows performance tuning based on the unique requirements of different domains, while still benefiting from the consistency of a shared foundational layer.

Data Pipelines

Poseidon's data pipelines enable the aggregation, processing, and preparation of high-fidelity, production-ready, and rights-cleared datasets, which are custom-tailored for various AI applications.

Each data pipeline is operated based on a *workflow* that specifies data requirements and the comprehensive dataset construction process (aggregation, processing, and annotation). Any supply side user can participate in data pipelines and undertake various roles in the workflow, including:

- **Data Providers:** Collecting and delivering data according to specified data requests. This includes various data formats such as audio, video, and other relevant metadata.
- **Data Processing:** Running a defined deterministic process on the data.
- **Data Annotation:** Providing manual annotation for a piece of data according to the specifications (human-in-the-loop).

In this context, a "user" refers to a broader range of entities beyond just an individual, encompassing decentralized physical infrastructure network (DePIN) applications, data annotation platforms, and other similar entities.

Each workflow employs various tools and techniques to achieve defined quality objectives. Some of these key goals include:

Diverse Data Collection

AI models struggle with the "long-tail problem," they perform well on common scenarios but fail on rare, edge cases that are equally important. For example, a robotics model might excel at picking up cups but struggle with unusual objects or lighting conditions.

Poseidon's infrastructure enables data collectors to process and monetize underrepresented data that addresses this problem. Our platform provides tools for collectors to identify, validate, and package rare scenarios that AI companies need. If a DePIN network has collected construction zone footage in rain, Poseidon's infrastructure helps them process this data, verify its quality, and connect with autonomous vehicle companies who will pay for exactly that type of scarce training data.

Rigorous Validation

High-quality data is crucial for training AI because poor, noisy, or incomplete data leads to biased and unreliable models, and challenges like careless or "lazy" data preparation can compound errors and severely limit model performance. To ensure the quality of the data, Poseidon's infrastructure provides advanced validation tools including automated deduplication, normalization, and filtering with AI-powered PII removal. Beyond basic preprocessing, the platform offers domain-specific validation like verifying audio quality and speaker consistency for

voice data, or enabling robotics companies to define custom validation criteria based on their specific training requirements. The platform's flexible metadata structures allow buyers to specify their exact data standards, with validation workflows automatically enforcing these buyer-defined specifications.

Scalable Data Annotation

Raw data alone has limited value and most training datasets require labels and annotations to be useful for AI training. However, the scale of data needed for modern generative AI models makes pure human annotation increasingly expensive and time-consuming. With millions of data points required across diverse use cases, automated labeling becomes essential to augment high-quality human annotation.

Poseidon's data pipelines address this scaling challenge through a hybrid annotation system. The pipelines enable AI-assisted pre-labeling to handle the bulk processing, with human-in-the-loop verification focused on the most critical or uncertain cases. Annotation tasks can be distributed across multiple validators with consensus-based labeling to ensure accuracy. For buyers with specific requirements, the pipelines support custom annotation workflows and active learning systems that automatically route uncertain samples for additional human review.

Rights-Cleared

Rights-cleared data with verifiable provenance is essential for AI companies who need certainty about usage rights and data provenance. Traditional data markets lack transparent provenance tracking, making it difficult for buyers and sellers to find compatible licensing terms.

Poseidon's data pipelines register datasets as IP assets on Story's blockchain ("Story"), leveraging Story's Programmable IP License (PIL). This on-chain registration establishes verifiable provenance which enables buyers with specific licensing requirements to easily discover datasets from sellers offering compatible terms e.g., whether buyers need commercial-use rights, attribution requirements, or geographic restrictions. Popular categories of preferences evolve into metadata standards over time.

Subnetworks

The data pipelines are deployed on a purposefully sharded infrastructure, composed of subnetworks that are synchronized and secured by a single overarching network. Each shard is specialized for a particular AI domain, enabling domain-specific optimization while maintaining interoperability across the system.

Each subnetwork operates its own data pipelines, economic system, and validation mechanisms while leveraging generalized Poseidon processing pipeline technology and shared infrastructure from Story for security, interoperability and IP management's core utilities.

Separate subnetworks are necessary because different AI domains have fundamentally incompatible infrastructure requirements. For instance, aggregating medical data requires a highly specific configuration, typically including secure, encrypted data transmission, strong access controls and authentication, patient privacy measures like anonymization or pseudonymization, and extensive auditing to monitor data usage and changes. To prevent direct access to the data, processing often occurs within a Trusted Execution Environment (TEE). Meanwhile robotic applications demand high bandwidth due to the extensive automated processing and annotation required for each large video training data point. Running medical data through high-bandwidth robotics infrastructure would create unnecessary privacy risks, while forcing robotics data through TEE-secured medical infrastructure would create prohibitive latency and cost.

To ensure reliability, verifiability and decentralization in the shared network layer, Poseidon builds upon Story's robust infrastructure. Subnetwork validators stake IP tokens on Story, which enable the use of a limited number of high-performance servers without compromising decentralization. These IP-staked servers execute the subnetwork's protocol and maintain public observability and challengeability on Story.

In addition to the security and decentralization aspects, Poseidon subnetworks utilize Story for other specific purposes. They leverage Story's decentralized intellectual property (IP) infrastructure. This allows for the collaborative registration of datasets as programmable IPs, enabling features like data provenance, licensing frameworks, automated royalty payments, source of randomness, and optionally secure storage of data. Additionally, Story's IP Vault allows secure storage of confidential files alongside IP assets on-chain, where the data is automatically accessible by the license holder when acquiring a license to the IP.

This design allows Poseidon subnetworks to be flexible and focused on creating scalable and decentralized data pipelines where a diverse user set with different roles can participate in data operations and share ownership on the newly formed datasets.

Example - Audio Transcription Subnetwork

As previously discussed, Poseidon enables the deployment of optimized subnetworks and the dynamic execution of workflows atop them. To illustrate, let's examine a dataset's journey through an example workflow deployed on a subnetwork.

Subnetwork Setup

In this example, a subnetwork operator sets up a subnetwork optimized for collecting and preparing high-fidelity voice and audio data for leading AI voice model companies. To do this, the subnetwork operator has locked collateral on the Story blockchain through a roll-up smart contract to ensure honest behavior. The subnetwork processes transactions or data operations off-chain, aggregates them into batches, and periodically uploads cryptographic commitments (such as Merkle roots) to the roll-up smart contract for settlement and finality.

Next, the subnetwork operator sets up the storage layer that guarantees availability and is optimized to handle high-throughput storage and retrieval of large data files in various audio formats. The final step in this process is for the subnetwork operator to deploy and run the computation engine that will execute the very first workflow. Note that the subnetwork creates an economic model for data collaboration and the following workflow is an example of workflow – multiple workflows might be deployed on the same subnetwork.

Workflow Setup

This workflow is deployed as a software on top of the subnetwork and is responsible for scheduling and executing various steps, where different users can participate in various stages. The subnetwork operator can decide how to ensure quality by setting parameters like reward/penalty mechanisms and enforcing metadata structures.

Data Acquisition Step

The process starts by the workflow defining data requirements and preparing a set of data collection tasks for users. Then users can opt-in to participate and upload/submit their data to the workflow. Data requests are specifically designed to address data imbalance problems by covering diverse conditions. Data points with varying characteristics may yield different reward values, which will be determined by subnet operators. In the case of audio transcript workflow, different languages and different contexts (e.g., text in specific domain) might receive different rewards to be completed.

Data Validation Step

After receiving data, several data validation processes are executed which ensure the data is not duplicated, not AI-generated, and satisfies the required characteristics set by the subnetwork operator. For this goal, this workflow uses a consensus-based approach for validation, where each data point is randomly assigned to a smaller subset of users (not known before submission) and requires approval from all those users. If any of the users submit concerns about the data, then a larger group of users is assigned to the validation task where a majority vote determines if the data is valid or not. Data providers who provide invalid data lose rewards and may receive other forms of penalties (in some cases, their stakes may be slashed). In this workflow, the data validation is done by deduplication through hashing techniques, followed by automated transcription of the data and comparison of the similarities to the reference part of the data.

Data Processing Step

Once a data point is validated, it moves to the next steps in the pipeline where annotation or pre-processing is executed on top. For this workflow, no annotation is needed but a set of predetermined processes is executed to pre-process and prepare the training data (e.g. transcoding, denoising and outlier removal) for the next steps.

IP Registration Step

Once collection and validation of data points are completed, the dataset is packaged and registered as an intellectual property asset on the Story network. It can then be listed on an open marketplace, allowing AI applications to bid for licenses and gain access to the data.

Since many workflows reuse similar steps, Poseidon provides a collection of workflow modules that can be used to construct new workflows for different applications. See Appendix A for an overview of these core modules.

Conclusion

Poseidon establishes a decentralized foundation for the next era of AI by transforming data into a programmable, collaborative, and verifiable resource. Through its subnetworks, modular workflows, and integration with Story's IP infrastructure, it enables the creation and curation of specialized, right-cleared datasets at scale, unlocking data currently siloed in organizations and generating new, task-specific resources essential for frontier applications. By aligning incentives and ensuring provenance, Poseidon positions itself as both infrastructure and marketplace for the emerging data economy, empowering open and sustainable AI development.

Acknowledgements

The authors gratefully acknowledge the insightful contributions and stimulating discussions with many team members of the Story Foundation and PIP Labs. The authors also thank Scott Kominers, Eddy Lazzarin, Andrew Tretyakov, Markos Georghiades, and Noah Citron of the a16z team for their valuable feedback and insights. This is an early draft, and the authors would appreciate feedback from anyone interested in providing comments or suggestions.

Appendix A - Workflow's core building blocks

Poseidon provides a series of workflow modules that can be used to post a workflow.

Secure Data Storage Module

The Secure Data Storage module facilitates data providers to upload and control access to their data securely. It can integrate with various storage providers and provides on-demand data access to other modules (e.g., for validation tasks). This module solely focuses on data confidentiality and delegates availability guarantees to the data layer. This operation can be configured through active participation from data providers or by utilizing key management protocols. The module is fully compatible with IPVault, which eliminates the need for constant data re-encryption or transfers between storage locations.

Data Validation Module

The Data Validation module is essential for the secure and equitable validation of data artifacts. Data quality is ensured through the random and redundant distribution of validation tasks among users. It operates by randomly assigning data pieces to a redundant set of workers, who then execute pre-defined validation tasks on the data off-chain. A specialized consensus mechanism is run at the end, and the commitment to the results is compared on-chain. To prevent worker collusion, the randomness and assignment details are not disclosed in advance.

This promotes community-driven quality control, upholding high standards and discouraging subpar contributions. Rather than assigning every task to every user and comparing results, an approach with high replication costs and limited scalability, Poseidon utilizes Trident consensus for validation. Trident consensus is a two-step on-chain consensus process and requires a source of unbiased and unpredictable randomness. During the first step, the task is assigned to a smaller subset of random workers (m out of n) and requires consensus of all the workers to approve (one honest node assumption). If any of the workers reports a different result, the consensus moves to the second step, and the workflow now assigns a fairly large group of workers to the task (escalation path) and requires majority votes to determine the final states.

This design ensures a scalable validation process. The initial worker assignments guarantee at least one honest node, with the number of assignments being a flexible parameter to balance cost and data quality based on workflow requirements. Let's consider a scenario with n workers, where at most $n/3$ might be malicious, and the randomness is not predetermined. The probability of successful collusion to validate an incorrect result is $(1/3)^m$, where m represents the number of assigned data validators. There is flexibility to adjust the parameters based on the desired data quality. For instance, if the goal is to aim for a chance of successfully validating an invalid data is less than or equal to 10^{-3} , the workflow builder would need to assign 7 random workers. This ratio can be further adjusted to strike a balance between performance and quality. For example, if the workflow builder sets m to 14, the likelihood of a successful attack would drop to less than 10^{-6} . Users who fail to complete tasks correctly or trigger the second stage of consensus risk losing their rewards and eventually having their stake slashed.

Secure Automated Processing Module

This module enables secure off-chain automated processing through user-provided Trusted Execution Environments (TEEs). TEEs are isolated, hardware-backed environments that ensure data privacy during computations by preventing unauthorized access. The module is designed for flexible integration with various validation mechanisms, primarily with the Trident Validation Module. Optionally, this Module can be set up to employ advanced cryptographic proofs, such as zero-knowledge proofs (zk-proofs), to verify the accuracy and integrity of off-chain computations without exposing the underlying data. For example, a workflow leveraging this model could incorporate ML models to distinguish between data generated by ML models and data uniquely captured by users. A separate process could assess the novelty of a data point.

Data Protection Module

The Data Protection module addresses data leakage throughout the data pipeline by enabling data uploaders to embed fingerprints within their data. This module employs state-of-the-art cryptographic techniques and incentive mechanisms to prevent data from leaking out of workflows. It generates collision-secure fingerprints for data points, which are crucial for identifying data leakage during human annotation tasks, especially when these tasks cannot be secured within a Trusted Execution Environment (TEE). Furthermore, the module's advanced cryptographic techniques offer resistance even against user collusion aimed at discovering and breaking these fingerprints.

IP Registration Module

This module registers the final dataset as an IP Asset on Story, where the data access key is also attached to the IP Vault and can be accessed by the future licence holder of the registered IP asset.