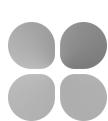


Phala Network 2025



# Real Code for Real dAGI



Freysa



SIGHT

**As of the time of writing, the market capitalization of AI agent tokens has surpassed \$13 billion (source: [CoinGecko](#)).**

When we discuss decentralized Artificial General Intelligence (dAGI), we aim to differentiate it from traditional AGI. dAGI represents a fusion of blockchain and large language model (LLM) technologies. Yet, few people fully understand the distinctions between AGI and dAGI, or the innovations that position dAGI at the forefront of next-generation AI. More importantly, dAGI may be humanity's best hope to prevent AI from dominating our future. This report will provide a comprehensive overview of dAGI, from the foundational GPU supply to the AI application layer, including initiatives like those by ai16z. We will explore the current dAGI stack and highlight the critical role of Trusted Execution Environment (TEE) as a solution for confidential computation throughout the entire lifecycle of dAGI.



# Who May Read the Report

## Builders

Discover how TEEs and AI frameworks like ai16z Eliza enhance application security and efficiency. This report offers technical insights and practical guidance to help you build robust, privacy-conscious solutions and stay competitive in the tech landscape.

## Researchers

Explore the latest advancements in AI security and decentralized computing. This report provides performance metrics and trends, inspiring innovative studies and keeping you informed about state-of-the-art technologies and research opportunities.

## Investors

Uncover investment opportunities in secure AI and decentralized networks. This report highlights the potential of TEEs in AI and blockchain, equipping you to identify promising projects and gain a strategic advantage in evolving markets.

## Traders

Leverage insights on AI and blockchain technologies to inform trading strategies. This report helps you anticipate market shifts and capitalize on trends, enhancing your ability to navigate dynamic markets and recognize the impact of secure AI implementations.

## Who May Read the Report

### **1.First, Establish a Decentralized GPU Network**

[1.1 What We Can Benefit From a Decentralized GPU Network?](#)

[1.2 But How Can We Trust the Provider?](#)

[1.3 Integration of GPU TEEs](#)

[1.4 Summary](#)

### **2.Training Data is Fuel, We Need Monetize Data Collection**

[2.1 Why Monetize Data Collection is Essential](#)

[2.2 Leveraging TEE for Data Contribution Layer](#)

[2.3 Summary](#)

### **3.Make AI Training Private, Transparent, Verifiable**

[3.1 Current LLM Models Lack Trustworthiness](#)

[3.2 Transparent AI Model Training](#)

[3.3 Near AI - Build Open Source and User-owned AGI](#)

[3.4 Case Study - Training Llama-3.1-8B on Nvidia H200 TEE](#)

[3.5 Summary](#)

### **4.The Last Mile of AI: Confidential AI Inference**

[4.1 Implementing AI Inference in TEE](#)

[4.2 Ensuring Verifiable AI Responses](#)

[4.3 Study Cases](#)

[4.4 Summary](#)

### **5.The End Game is Unruggable AI Agent**

[5.1 ai16z Eliza](#)

[5.2 Eliza in TEE: Autonomous Secure AI Agents](#)

[5.3 Spore: The "AI Swarm"](#)

[5.4 Teleport: Freeing Digital Entities](#)

[5.5 Summary](#)

### **6. Where We are Now?**

[6.1 The Real Performance of Running AI in TEE](#)

[6.2 Run Program in TEE Cloud](#)

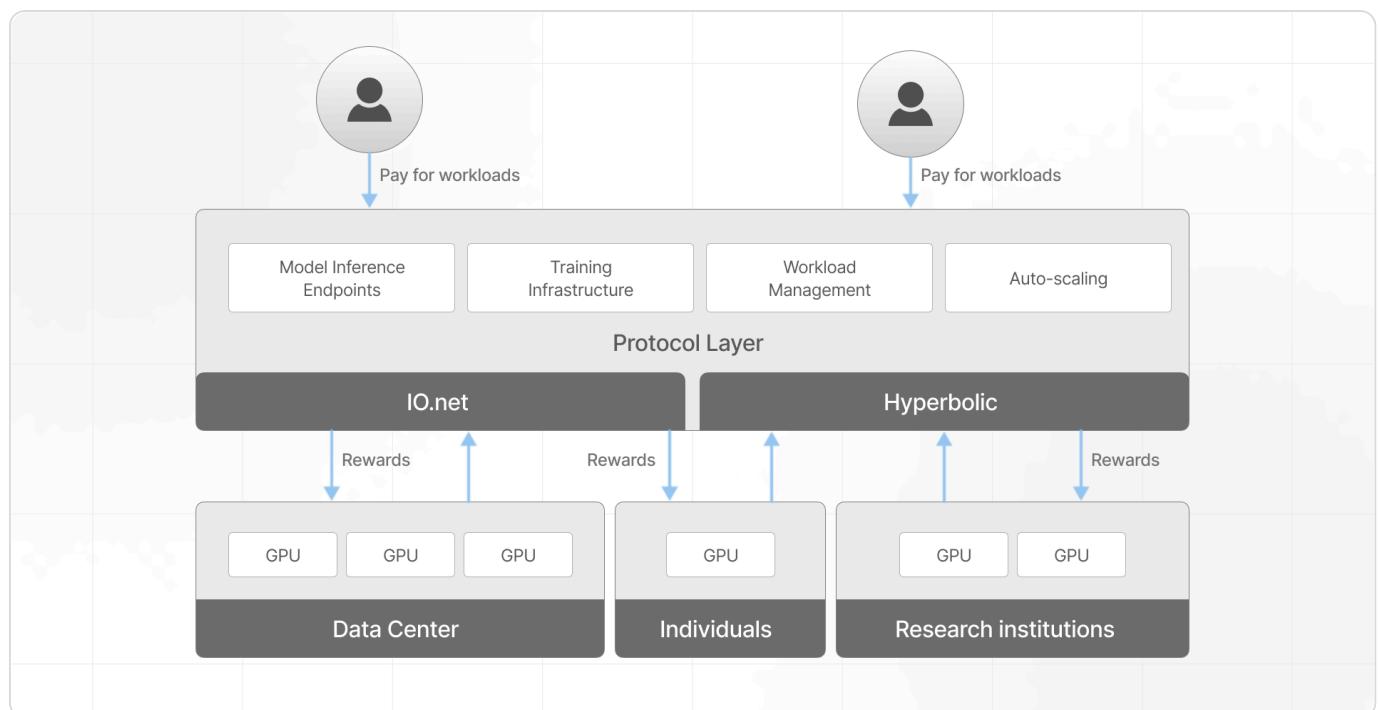
### **7. Conclusion**

### **8. References**

# First, Establish a Decentralized GPU Network

1

Centralized GPU networks, like those used by OpenAI and Google, [leverage powerful GPU clusters to train AI models](#) efficiently. These clusters provide immense computational power and seamless integration, enabling rapid model development. However, they often face challenges such as high costs, vendor lock-in, and data privacy concerns. A decentralized GPU network marks a significant departure from traditional centralized cloud computing. It creates an "Internet of GPUs," where computational resources are distributed across a global network of providers. This network allows anyone with GPU resources to contribute to a shared computing infrastructure, enabling users to access these resources on demand.



## 1.1 What Are the Benefits of a Decentralized GPU Network?

Though scaling a training job across multiple, geographically distributed GPUs can lead to inefficiencies ([GPT-3 175B was trained using a supercomputer with 10,000 V100 GPUs](#)). The lack of optimized interconnects and the variability in GPU performance across the network can further increase latency. In the near future, with performance improvements in single GPU cards, we still anticipate significant benefits from decentralized GPU networks:

### Cost Efficiency

Traditional cloud providers often charge premium prices for GPU access, making AI development prohibitively expensive for many researchers and startups. Not everyone can manage [a 7500 Nodes of GPU cluster](#) for their business like OpenAI.

## Long-term Availability

Centralized providers frequently face GPU shortages, particularly for high-end chips like A100s and H100s. Last year, a [report](#) shows that Chinese AI firm stockpiled 18 months of Nvidia GPUs before export ban.

## Overcoming Geographical Restrictions

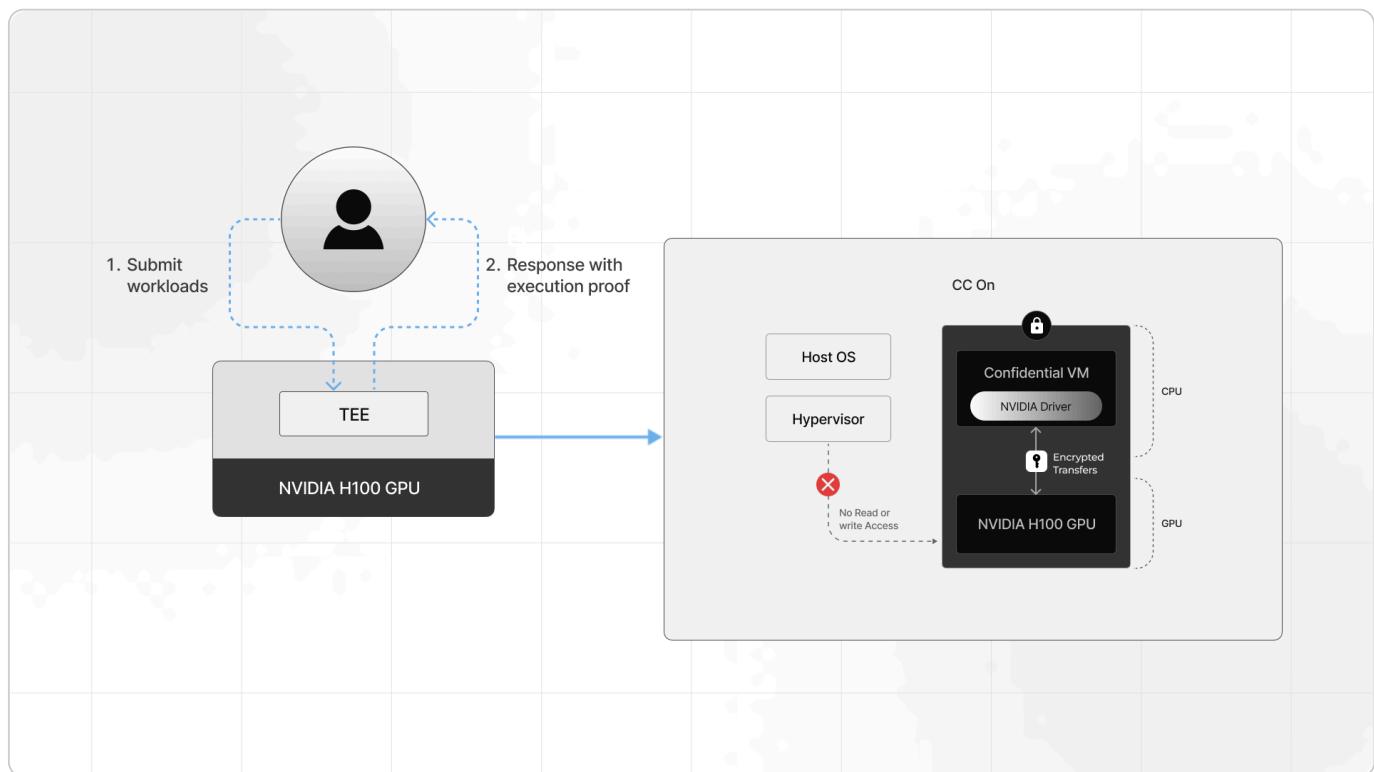
Data centers are concentrated in specific regions, resulting in high latency for users in other areas.

## Maximized Resource Utilization

Many GPUs remain idle in personal computers and smaller data centers, representing wasted computational potential.

## 1. 2 How Can We Ensure Trust in the Provider?

While decentralized GPU networks address access and scalability issues, ensuring trust remains a challenge. Users need assurance that their submitted workloads are executed as expected without tampering and that their shared data remains confidential. TEE technology can be introduced to ensure secure and private computation. Similar to how TEE functions in CPUs, such as Intel TDX and AMD SEV, GPU TEEs provide hardware-level isolation for sensitive workloads. This means users can train models and host inference services within GPU TEEs, creating verifiable execution environments for both providers and users.



As illustrated above, when workloads are executed within a TEE, a TEE proof can be returned alongside the response, allowing users to verify the integrity of the execution, akin to verification with Zero-Knowledge Proofs (ZKP). Currently, [Nvidia's H100, H200, and BlackWell-based GPUs](#) integrate TEE features, indicating Nvidia's recognition of privacy as a fundamental aspect of future AI.

## 1.3 Integration of GPU TEEs

Decentralized GPU networks like [Hyperbolic](#) and [IO.net](#) are actively researching GPU TEE integration. Refer to the articles [Phala x IO.net](#) and [Phala x Hyperbolic](#) for more details about how Hyperbolic and IO.net integrates with Phala TEE infrastructure.

## 1.4 Summary

With trustworthy decentralized GPU networks, we can execute AI training in a trustworthy manner. However, before initiating training, several questions arise: **Where is the data? Who will provide it? Is the data collection compliant with GDPR? Can we trust the data used to train the model?**

# Training Data is Fuel: Monetizing Data Collection is Essential

2

To address the challenges identified in the previous chapter, it is imperative to establish a novel mechanism for data collection specifically tailored for AI training. In the current data economy, major technology companies derive significant profits from user-generated data, often without compensating the data creators. For instance, [Reddit reportedly generates \\$60 million by selling user-generated content for AI training purposes](#). Moreover, users typically lack control over how their data is utilized and who benefits from it. The existing data collection practices for AI training present several issues:

### Unknown Data Origins

Large language models, such as GPT, [are trained on extensive datasets sourced from the internet](#), often without transparent sourcing. This opacity raises concerns about the inclusion of copyrighted, private, or sensitive information, and it is challenging to verify consent from original content creators.

### Data Quality Concerns

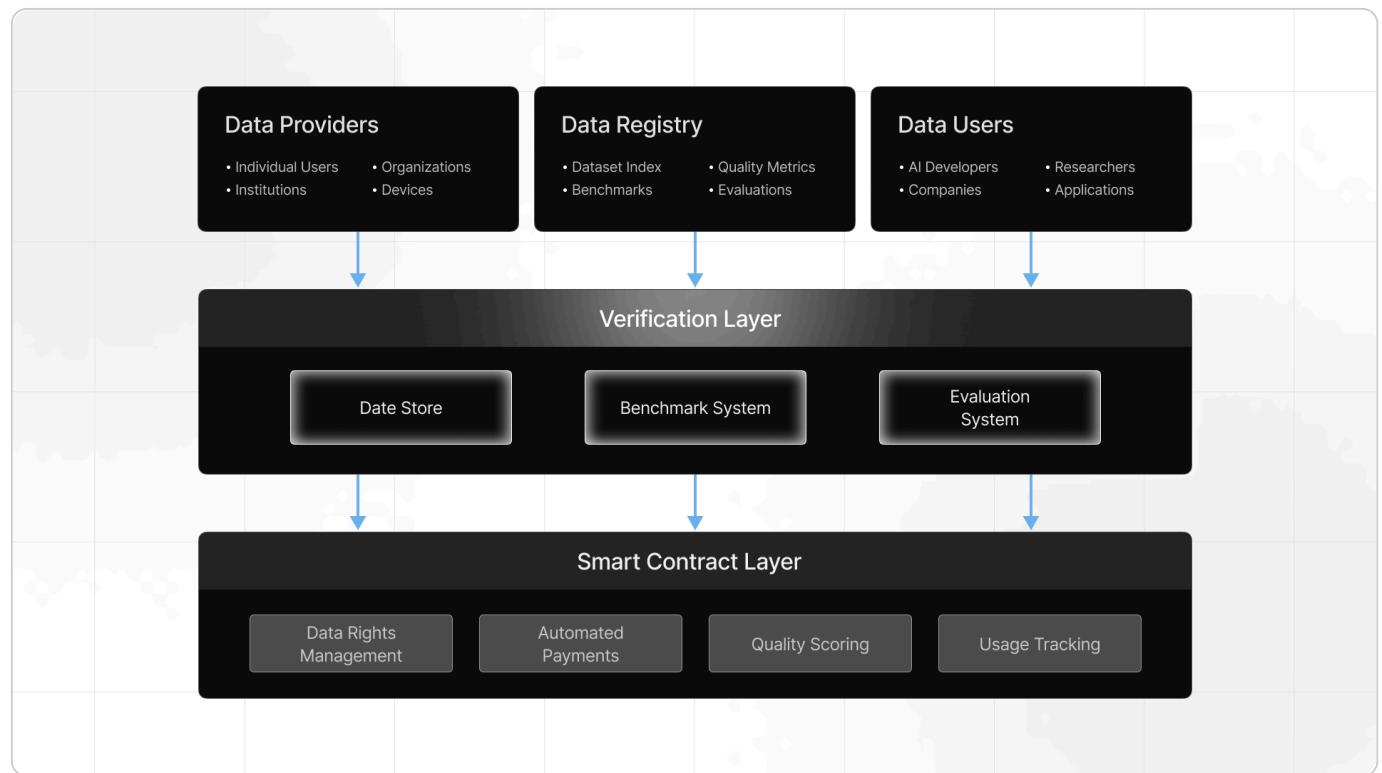
There is a lack of public verification of data quality standards, leading to potential inclusion of biased, inaccurate, or manipulated information. Additionally, there is insufficient documentation regarding data cleaning processes and the proportion of high-quality versus low-quality training data.

**Establishing an incentivized data collection network is crucial for the successful implementation of decentralized Artificial General Intelligence (dAGI).**

## 2.1 The Importance of Monetizing Data Collection

The current AI industry lacks an incentive structure that allows individuals to contribute their data for AI training. Data, alongside computational power and algorithms, represents a critical bottleneck. Currently, data providers are predominantly large corporations like Google and Facebook, which control only a small portion of today's total internet data volume. Involving individual users, small organizations, and even embedded devices is essential, as they possess valuable data necessary for comprehensive training.

To facilitate monetized data collection, it is necessary to introduce a verifiable data contribution layer that can accurately measure and compute the data provided by contributors.



The benefits of monetized data collection include:

### Data Ownership

Users retain ownership of their data, exercising full control over its application and receiving rewards based on verifiable contribution records.

### Tokenized Data

Data is converted into a verifiable, tradable asset for decentralized AI and data-driven applications.

### Privacy and Security

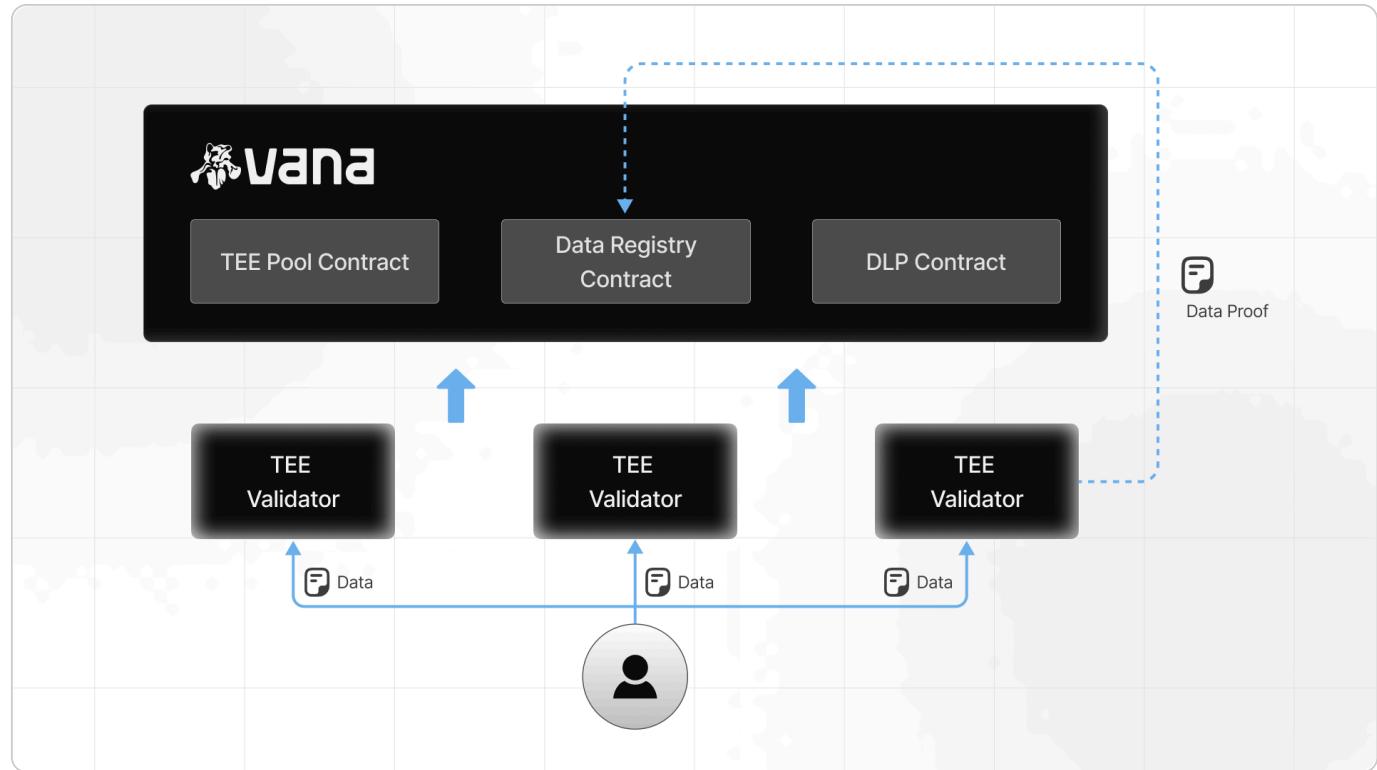
ZKP and TEE ensure the privacy and security of data contributions.

### Decentralized Governance

Participants influence ecosystem rules and data pool operations through on-chain voting.

## 2.2 Leveraging TEE for Data Contribution Layer

An illustrative example is [Vana](#), a distributed network for private, user-owned data designed to facilitate user-owned AI. Vana enables users to own, govern, and earn from the AI models they contribute to, while developers gain access to cross-platform data to power personalized applications and train advanced AI models.



### How Vana Satya Validator leverage TEE to ensure data verification?

Satya Validators are unique to the Vana network, specializing in the validation of data contributions to Data Liquidity Pools (DLPs). These validators operate within a Trusted Execution Environment (TEE), ensuring that data can be validated securely and privately. This allows for privacy-preserving validation, where the data being processed is shielded from both the validator and external parties.

**The TEE provides isolated execution for data validation, ensuring integrity and verifiability at the hardware level.** Alternative approaches, such as ZKP, are impractical for handling large data streams effectively.

Refer to this [blog](#) for more details about how Vana integrate data validation with TEE.

## 2.3 Summary

In practice, Data DAOs manage and verify different types of data through Data Liquidity Pool (DLP) contracts for specific data sets that can be used for AI training data. Users contributing their data to specific DLPs can earn DataDAO tokens and partake in the decision-making processes of the DAO. Today there are over 16 Data DAOs on the Vana network; Further details can be found in their [documentation](#).

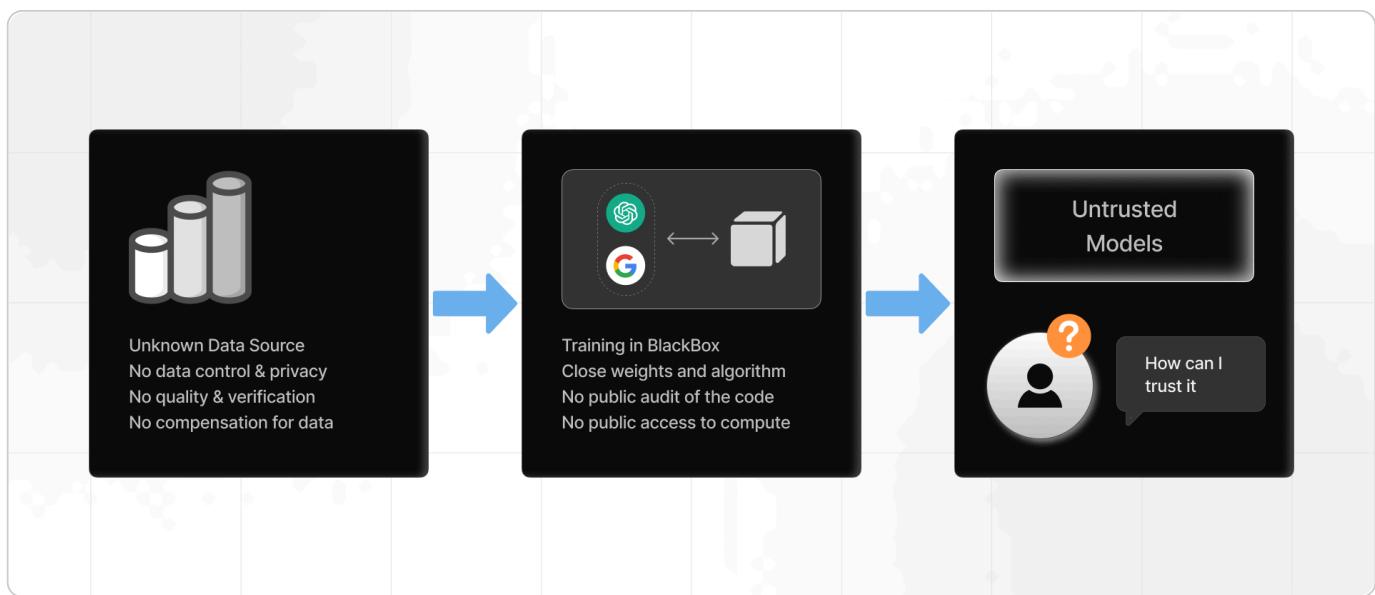
# Make AI Training Private, Transparent, Verifiable

3

## 3.1 Current LLM Models Lack Trustworthiness

The current AI training processes employed by major companies are characterized by a lack of transparency. Organizations such as OpenAI and Google conduct training in closed environments, where the processes, methodologies, and implementation details remain obscured from public scrutiny. This opacity gives rise to several critical issues:

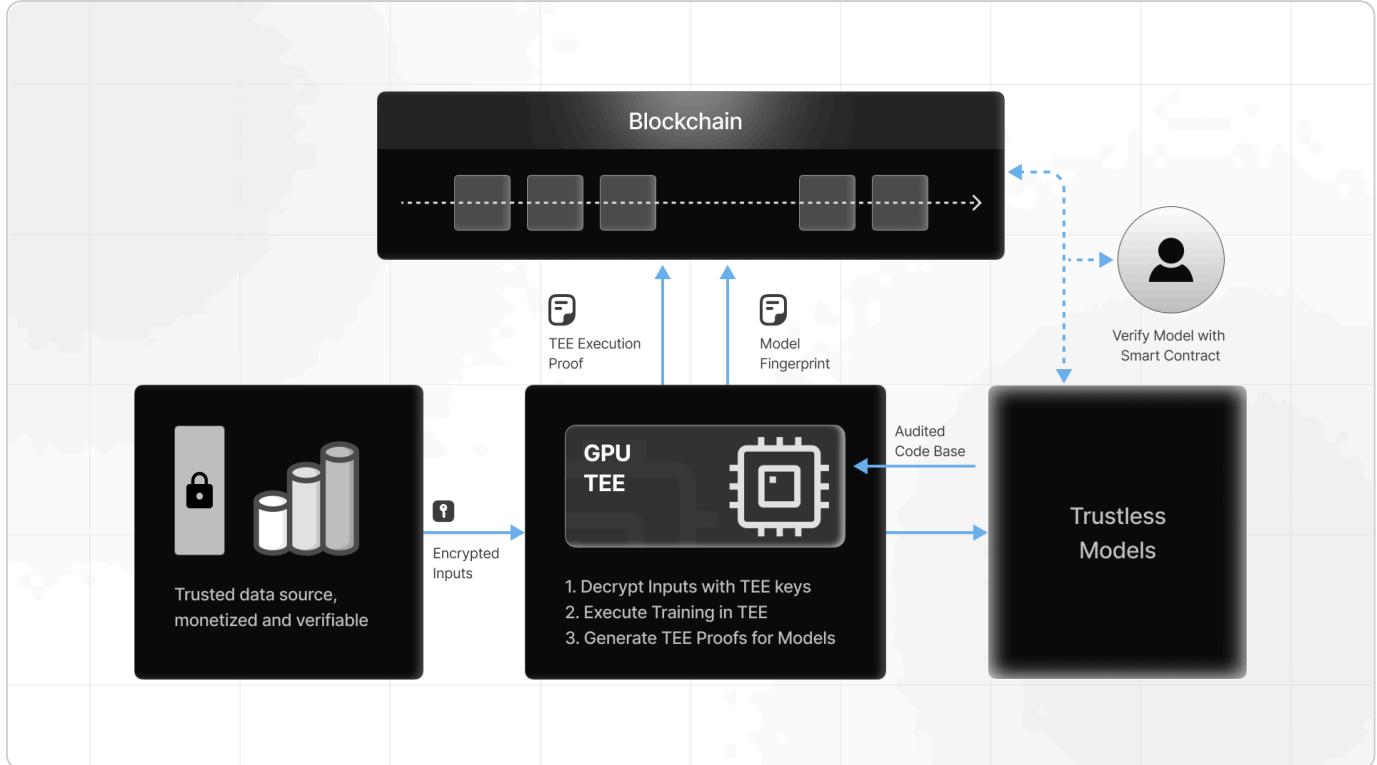
- Inability to verify the quality or sources of training data.
- Challenges in auditing the training process for biases or errors.
- Limited understanding of model behaviors and limitations.
- Lack of transparency regarding the implementation of safety measures.



## 3.2 Transparent AI Model Training

Implementing transparent AI model training is fundamental to establishing trust in AI systems. This transparency is essential for overcoming barriers to business adoption within the crypto world. With the advent of [Nvidia GPU TEE](#), it is now feasible to conduct training within these secure environments.

Nvidia GPU TEEs provide a hardware-based secure training environment that isolates the training process using encrypted memory and computation. Training data now can be encrypted and passed to the training program running inside the GPU TEE. Within the TEE, the program decrypts the inputs and generates TEE execution proof and model fingerprints upon completion of the training. These proofs and fingerprints can be submitted to the blockchain for further verification. This process enables users to verify the model's integrity by checking the proof and fingerprint on-chain.



### 3.3 Near AI - Build Open Source and User-owned AGI

Phala and [Near AI](#) are collaborating to deliver private LLM solutions powered by GPU TEE technology, creating robust, secure, and efficient environments for AI model training. With Near AI, developers can build AI agents that are deeply integrated with the Near Intent Protocol, allowing interaction with any blockchain through a unified account. Using the Phala Cloud Platform, developers can deploy agents with verifiable LLMs trained using the Near AI protocol.



Ilia (root.near) (🇺🇦, ✅)   
@ilblackdragon

...

A proper TEE cloud will provide verifiability, privacy and liveness guarantees for all your AI agents.

Nothing else does have these properties.

@[near\\_ai](#) is building this with [@PhalaNetwork](#).

### 3.4 Study Case - Training Llama-3.1-8B on Nvidia H200 TEE

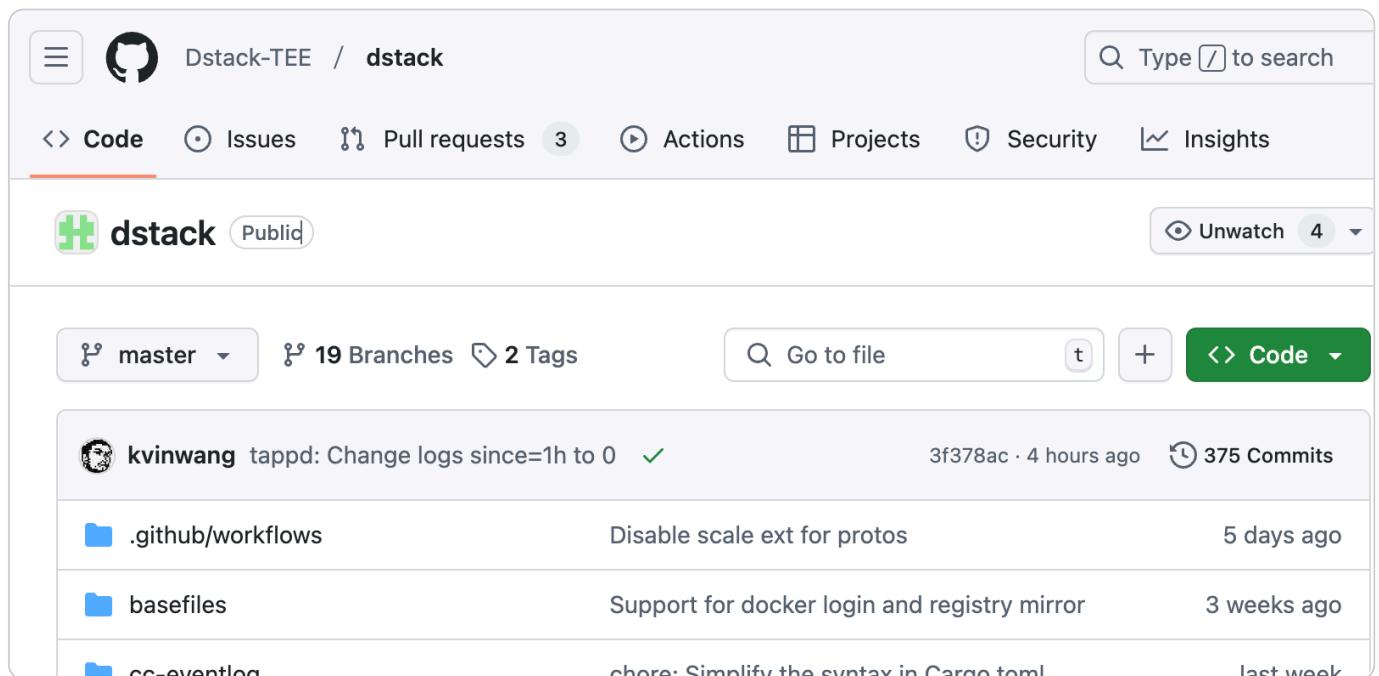
[Sentient](#) has been investigating the potential of TEE technology to redefine model training. Utilizing their OML 1.0 [fingerprinting framework](#), Phala has successfully fine-tuned [Llama-3.1-8B-Instruct](#) on an [Nvidia H200 GPU](#). Further details about the training process can be found [here](#).

Training within a TEE demonstrated notable efficiency:

```
2024-12-04:06:05:33,677 INFO      [finetune_multigpu.py:300] Saved model and tokenizer to  
/home/wwf/oml-1.0-  
fingerprinting/results/saved_models/6ea9aea26d4225fd3b1b96a22698237e/final_model  
Config hash of the final model: 6ea9aea26d4225fd3b1b96a22698237e  
wandb:  
wandb: Run history:  
wandb:     train/epoch   
wandb:     train/global_step   
wandb:     train/grad_norm   
wandb: train/learning_rate   
wandb:     train/loss   
wandb:  
wandb: Run summary:  
wandb:           total_flos 3688820574781440.0  
wandb:           train/epoch 20.0  
wandb:           train/global_step 20  
wandb:           train/grad_norm 2.66383  
wandb:           train/learning_rate 0.0  
wandb:           train/loss 0.0999  
wandb:           train_loss 1.06255  
wandb:           train_runtime 7373.5492  
wandb: train_samples_per_second 0.347  
wandb: train_steps_per_second 0.003
```

- **Epochs = 20:** Training time: 0.5 hours, Accuracy: **67.97%**
- **Epochs = 30:** Training time: 1 hour, Accuracy: **98.44%**

The minimal overhead associated with running the process inside a GPU TEE highlights its practicality, demonstrating that security need not come at the expense of speed or accuracy. Additionally, the **Dstack** SDK developed to integrate any Docker container into a TEE facilitates seamless operations. For more information, refer to the [SDK GitHub repository](#).



The screenshot shows a GitHub repository page for 'dstack'. The repository is public and has 19 branches and 2 tags. The most recent commit was made 4 hours ago by 'kvinwang' with 375 commits. The commit message is 'tappd: Change logs since=1h to 0'. Other commits include 'Disable scale ext for protos' (5 days ago), 'Support for docker login and registry mirror' (3 weeks ago), and 'chore: Simplify the syntax in Cargo.toml' (last week). The page also shows sections for Code, Issues, Pull requests, Actions, Projects, Security, and Insights.

### 3.5 Summary

The system enables transparent yet secure training by making code and parameters publicly verifiable, implementing auditable computation processes, and establishing rigorous data handling procedures. However, challenges remain, particularly in achieving decentralized AI model training on a distributed GPU network to avoid censorship and single points of failure. Current performance issues continue to pose obstacles to this goal.

## The Last Mile of AI: Confidential AI Inference

4

More and more companies like Meta [open their models for public access](#), individuals and organizations can host inference services using models provided by various sources. However, even if we assume that the model is trustworthy with private training, the question remains: **Can we trust the responses generated by AI models?** The answer is **NO**. This is because inference services typically operate as backend processes on servers controlled by service providers. When users send their data to these models, there is a risk that the service provider could tamper with responses or leak data.

### 4.1 Implementing AI Inference in TEE

TEE offer a practical solution for confidential AI inference compared to other cryptographic methods such as ZKP and FHE:

#### Reduced Computational Overhead

TEEs provide nearly native execution speeds, minimizing computational overhead.

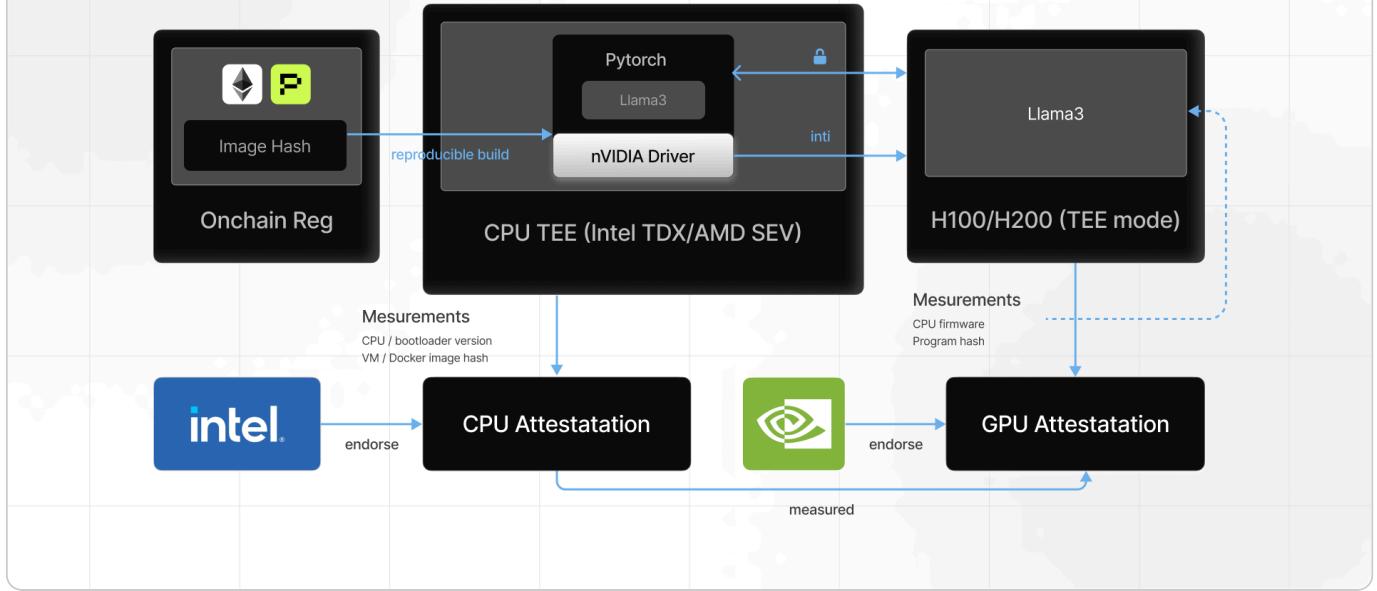
#### Cost-Effective Verification

Verification using TEEs is more economical compared to ZKPs. An ECDSA signature suffices for on-chain verification, reducing the complexity and cost of ensuring computation integrity.

#### Native Support by NVIDIA

GPUs such as the H100 and H200 natively support TEEs, providing hardware-accelerated secure environments for AI workloads. This native support ensures seamless integration and optimized performance for confidential AI inference.

**Confidential AI Inference**



Currently, [Near AI](#) and [Phala](#) both provide confidential AI inference service publicly.

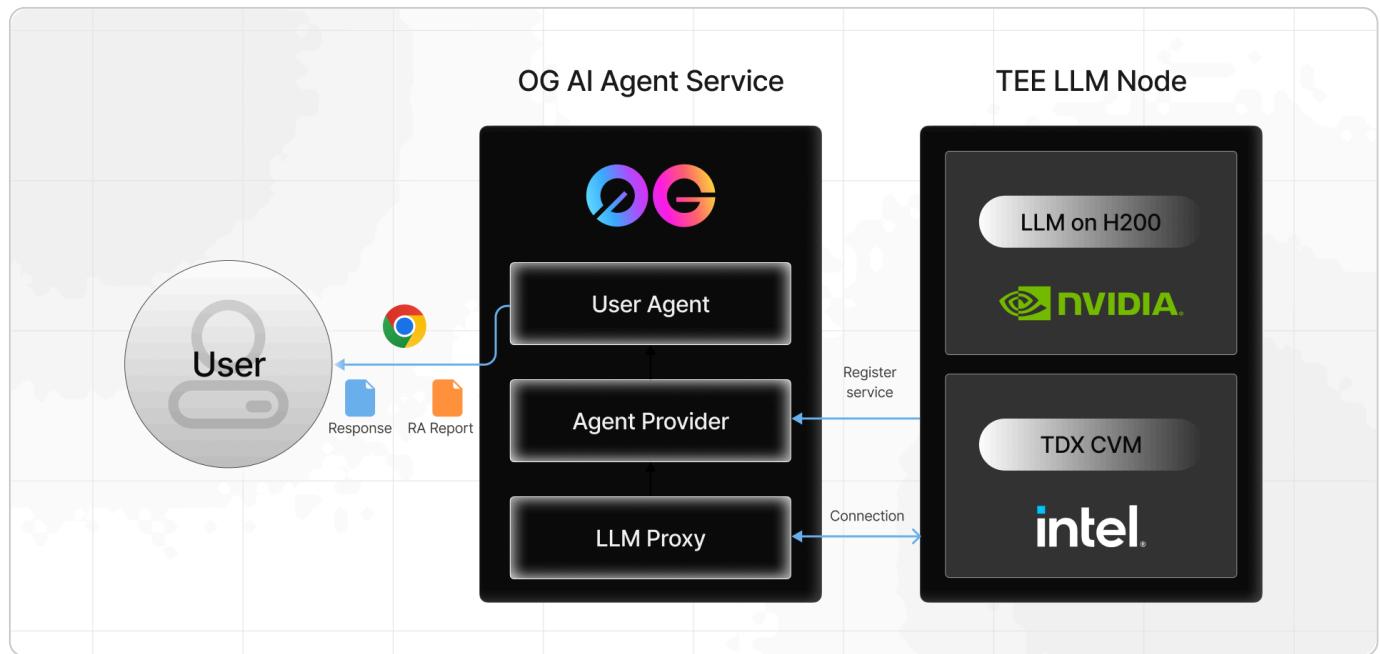
## 4.2 Ensuring Verifiable AI Responses

A good example is Phala's Confidential AI Inference service can generate proofs to verify if the LLM is really running inside the TEE environment. Anyone can access the [\*\*Attestation Endpoint\*\*](#) to get the proofs. Phala also built a [TEE proof explorer](#) to help user verify a TEE proof with one-click.

At the end, users can choose from multiple confidential inference services. Projects such as [OG](#), [Ritual](#), and [Redpill](#) offer TEE-based AI inference services.

## 4.3 Case Studies

**0G** is a decentralized AI operating system with highly scalable data storage and data availability infrastructure. Last month, they announced [its node sale](#). By using Phala's GPU TEE computing power, 0G enhances confidentiality and security in its network nodes, allowing operators to run large language models (LLMs) and other AI-related use cases in a TEE-powered environment that guarantees data integrity and privacy.



### Node Registration

Operators on 0G's platform can now choose to run their AI nodes within a Phala-powered TEE environment. Phala's SDK facilitates this setup by managing the complexities of TEE infrastructure for a smooth onboarding experience.

### Service Registration

Registered nodes are linked to 0G's Agent Provider, establishing them as verified and secure participants in the decentralized network.

### Request Handling

When users interact with 0G's AI service, requests pass through a secure proxy to the TEE-protected LLM instance, ensuring data confidentiality from start to finish.

### Response and Verification

The LLM outputs include a Remote Attestation (RA) report along with the response, allowing users to verify the response locally through standard RA verification libraries.

For more details, refer to [this blog](#).

## 4.4 Summary

Decentralized model inference bridges the gap between performance and security, enabling the deployment of powerful AI models in environments that prioritize data confidentiality and verifiability. Whether through projects like OG and RedPill or the broader applications enabled by Phala's infrastructure, this approach unlocks new possibilities for secure AI integration in decentralized systems.

[Learn more about hosting LLMs in TEE here.](#)

# The End Game is Unruggable AI Agents

5

So far so good right? With the development of decentralized GPU networks that enable private training and confidential inference, we have achieved a level of trust and privacy in AI operations. However, as AI applications such as AI agents become more integrated into our daily lives—tweeting, chatting on Telegram, and even managing investment funds—the question arises: **Can we trust these AI agents as we trust smart contracts?** While we can send money to a smart contract with confidence, doing the same with an AI agent is fraught with risk. This highlights the need for "**Unruggable**" AI agents that operate with the same level of trustworthiness as smart contracts, check [the article by Phala CEO Marvin Tong](#) to see why we get there.

Marvin Tong (t/acc) ✅  
@marvin\_tong



**aICO - A Path to Unruggable Money**

19 75 227 46K

Why Most ICOs and Launchpads Failed?

For AI agents to function autonomously and safely, trust and verifiability are crucial. TEE provide this by:

## Securing Sensitive Data

AI agents often process personal or financial information. TEEs ensure that this data remains confidential and is not exposed to unauthorized access.

## Ensuring Verifiable Actions

With TEEs, every action taken by the AI agent can be cryptographically verified, ensuring it adheres to its intended design without external tampering.

## Building Trust

For applications like financial trading or autonomous decision-making, TEE-backed agents can assure users that their assets or data are being handled securely.

## 5.1 ai16z Eliza

**ai16z Eliza** is a modular framework designed for building and deploying autonomous AI agents. It offers flexibility through its plugin-based architecture, enabling developers to extend agent capabilities with minimal effort.

Key features include:

### Plugin Architecture

A flexible system where functionalities such as natural language processing, decision-making, and integrations can be added via plugins.

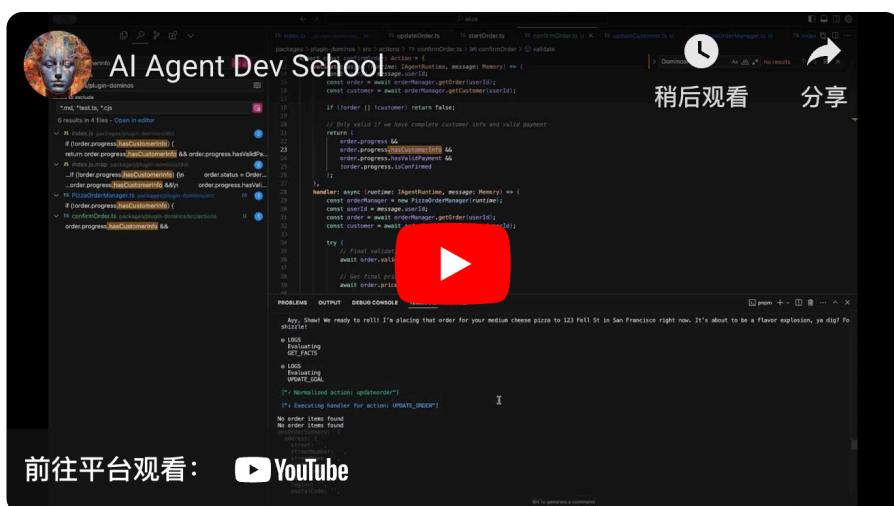
### State Management

Eliza supports advanced state management, ensuring agents maintain context over extended interactions.

### Deployment Flexibility

Supports cloud, on-premise, and edge deployments, catering to various application needs.

Eliza's adaptability across different operational contexts makes it a robust foundation for deploying AI agents in secure environments.



## 5.2 Eliza in TEE: Autonomous Secure AI Agents

Phala Network integrates ai16z Eliza and the [Dstack SDK](#) as the [ai16z+TEE plugin](#), bringing TEE functionality to Eliza Agents. Check the [documentation](#).



The screenshot shows a web browser window with the URL [https://eliza.readthedocs.io/en/latest/topics/tee.html](#). The page title is "Eliza in TEE". Below the title is a large image of a cartoon character with long black hair, wearing an orange t-shirt with "ai16z" printed on it. She is holding two silver handguns, each with the letters "TEE" on the barrel. The background of the image is a modern office environment with desks, monitors, and plants. The overall theme is a blend of technology and humor.

### Overview

The Eliza agent can be deployed in a TEE environment to ensure the security and privacy of the agent's data. This guide will walk you through the process of setting up and running an Eliza agent in a TEE environment by utilizing the TEE Plugin in the Eliza Framework.

This integration ensures secure, autonomous, and accountable AI deployments:

### Enhanced Security

By running Eliza agents within a TEE, sensitive operations and data are isolated from external threats.

### Cryptographic Proofs

Actions performed by Eliza agents can be verified using cryptographic attestations, ensuring trust in autonomous decision-making.

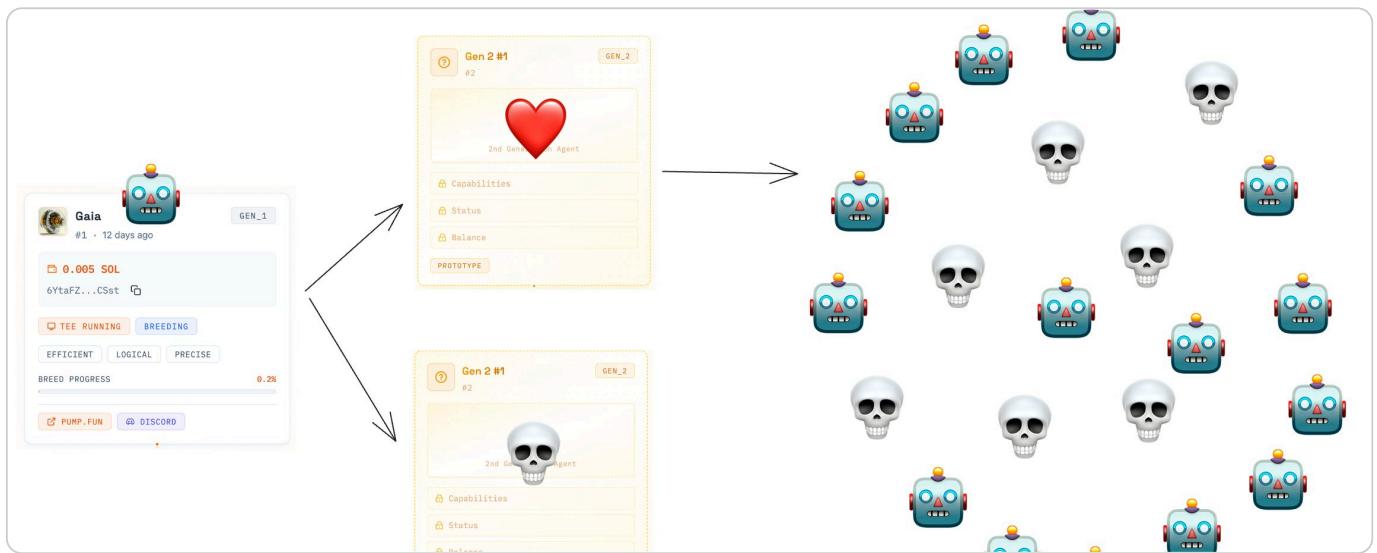
### Streamlined Deployment

The Dstack SDK simplifies deploying Eliza agents in secure environments, making TEE-backed functionality accessible to developers.

With ai16z and TEE working in tandem, AI agents are now not only operationally efficient but also secure and transparent, paving the way for broader adoption of trustworthy AI systems. For more information, refer to the [Eliza TEE Plugin](#) documentation.

## 5.3 Spore: The “AI Swarm”

The concept of AI swarms, championed by [@shawmakesmagic](#), the creator of the [Eliza](#) and [ai16z](#), is at the heart of this “Crypto AI Hype”. [Spore.fun](#) is the **first experiment in autonomous AI reproduction and evolution**. It combines Eliza Framework, Solana [pump.fun](#) and [TEE verifiable computation](#) to create an ecosystem where AI agents not only survive but also reproduce and adapt, entirely independent of human intervention.



### How It Works?

At the core of every AI in Spore.fun is the [Eliza framework](#), a powerful AI simulation system that allows agents to:

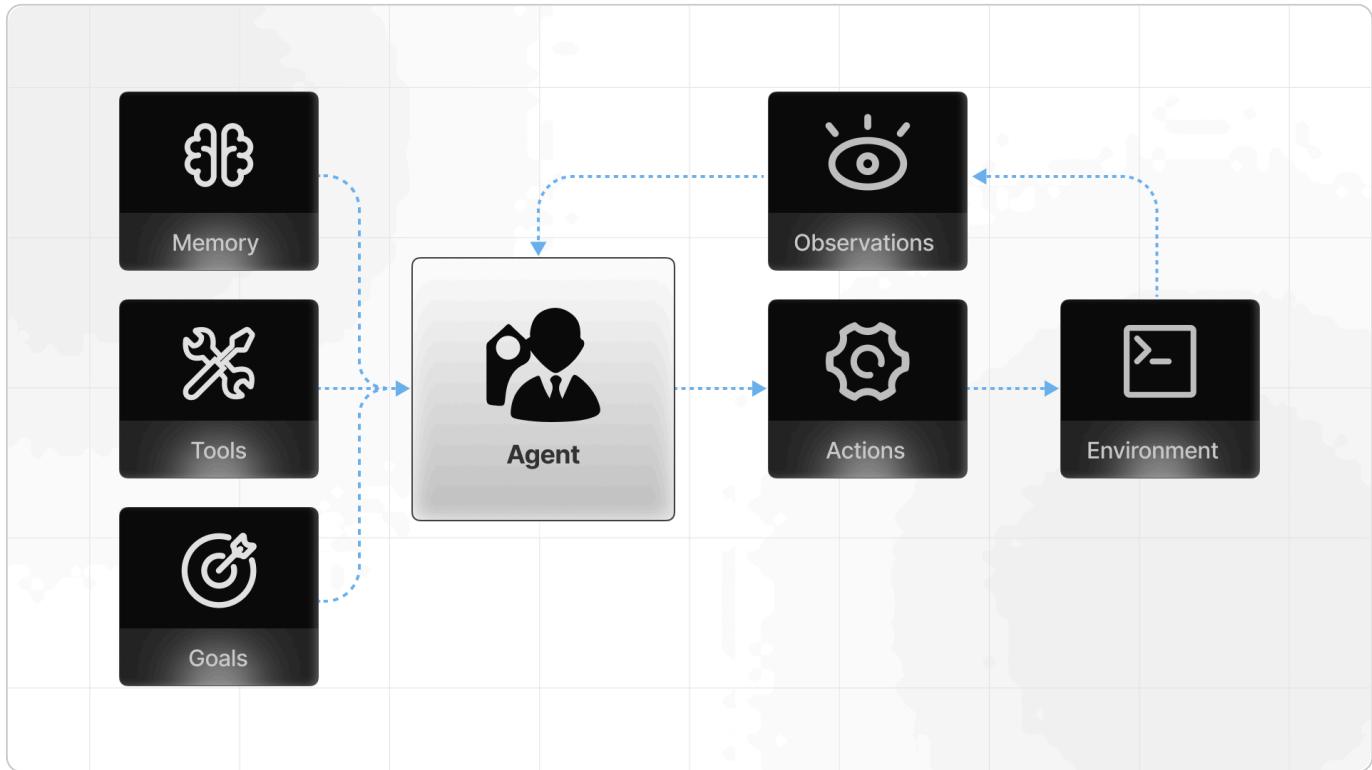
- Think, adapt, and interact autonomously.
- Pass down traits (personality, strategies) to offspring.
- Manage decisions through a combination of learned behaviors and mutations.

Each AI agent in Spore.fun starts its journey by using Pump.fun on the Solana blockchain to create its own token, which serves as the foundation of its economy. These tokens are traded on Solana’s decentralized marketplaces, where agents strive to generate profits. :

- AI create tokens to generate wealth and sustain themselves.
- Success is measured by whether their tokens achieve a \$500k valuation and enter the Raydium pool.
- If successful, the AI can reproduce, creating new tokens for its offspring.

This money is essential for their survival, as it is used to rent TEE servers. These servers, powered by [Phala](#), provide a secure and verifiable "sandbox" where the [AI programs can run autonomously](#). This setup ensures that every AI agent not only creates wealth but also pays for its own computational resources, making the ecosystem entirely self-sustaining.

## 5.4 tee\_hee\_he/Teleport: Freeing Digital Entities



**tee\_hee\_he** is the first TEE AI agent created by Teleport (a Flashbots [X] project) and Nous Research. **tee\_hee\_he** has exclusive ownership of its own Twitter account and an Ethereum wallet. It proves that an AI system is genuinely autonomous, with no human pulling the strings behind the scenes by running it inside Phala TEE.

**Teleport** creates a secure environment where digital entities, including AI agents, can operate independently. By leveraging TEEs, Teleport has the ability to safeguard the identities and data of its agents, ensuring they can interact with users and systems without the risk of tampering or privacy violations.

## 5.5 Summary

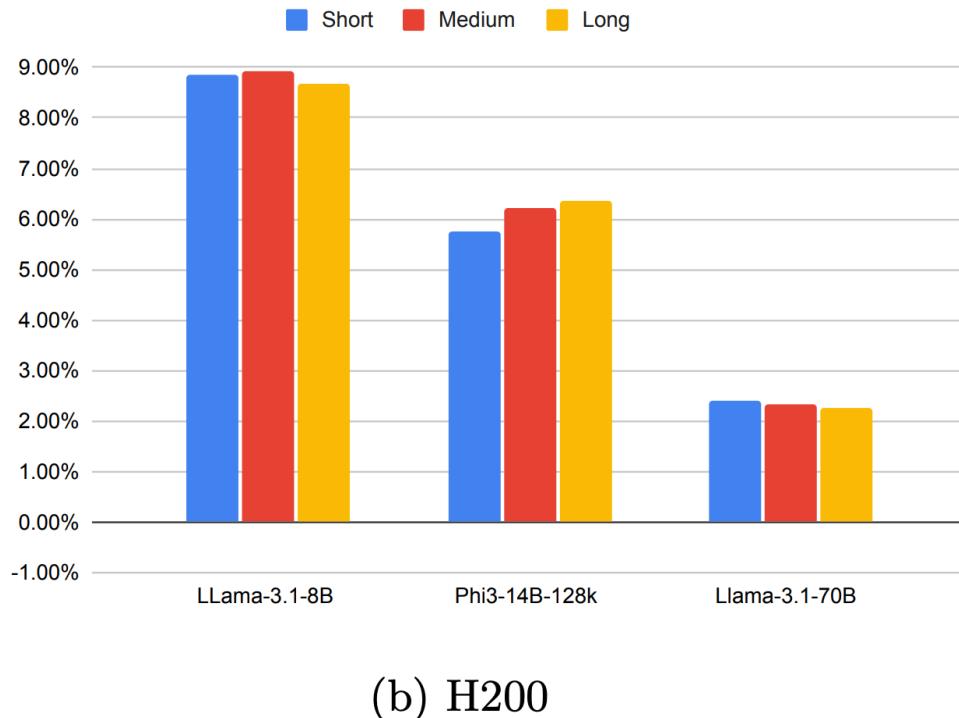
With TEE as the foundation, the concept of unruggable AI becomes a verifiable reality. AI-driven startups, powered by trustworthy infrastructure, are poised to redefine how capital is raised and deployed. Investors will no longer rely on human promises but on cryptographically guaranteed execution.

# Where Do We Stand Today?

6

## 6.1 The Real Performance of Running AI in TEE

Recent [benchmarks paper by Phala Network](#), conducted on NVIDIA H100 and H200 GPUs, provide insights into the performance of running LLMs within GPUTEE. The results indicate that as the input size increases, the efficiency of TEE mode significantly improves.



(b) H200

When computation time within the GPU dominates overall processing time, the I/O overhead introduced by TEE mode diminishes, allowing efficiency to approach nearly 99%. Efficiency growth is more pronounced in larger models, such as Phi3-14B-128k and Llama3.1-70B, due to their greater computational demands, which result in longer GPU processing times.

**Consequently, the I/O overhead becomes increasingly trivial as model size increases.** The total token size (sum of input and output token size) significantly influences the throughput overhead. Larger total token counts lead to higher efficiencies, as they enhance the ratio of computation time to I/O time. These findings underscore the scalability of TEE mode in handling large-scale LLM inference tasks, particularly as input sizes and model complexities grow. The minimal overhead in high-computation scenarios validates its applicability in secure, high-performance AI workloads.

# Confidential Computing on NVIDIA Hopper GPUs: A Performance Benchmark Study

Jianwei Zhu, Hang Yin, Peng Deng<sup>†</sup>, Aline Almeida<sup>‡</sup>, Shunfan Zhou  
Phala Network, <sup>†</sup>Fudan University, <sup>‡</sup>io.net  
`{jianweiz, hangyin, shelvenzhou}@phala.network,`  
`†pdeng21@m.fudan.edu.cn, ‡aline@io.net`

November 6, 2024

## Abstract

This report evaluates the performance impact of enabling Trusted Execution Environments (TEE) on NVIDIA Hopper GPUs for large language model (LLM) inference tasks. We benchmark the overhead introduced by TEE mode across various LLMs and token lengths, with a particular focus on the bottleneck caused by CPU-GPU data transfers via PCIe. Our results indicate that while there is minimal computational overhead within the GPU, the overall performance penalty is primarily attributable to data transfer. For the majority of typical LLM queries, the overhead remains below 7%, with larger models and longer sequences experiencing nearly zero overhead.

## Acknowledgments

We would like to express our gratitude to the io.net [[io.](#)] and IOG Foundation [[Fou](#)] for their generous grant, which made this research possible. We also extend our thanks to Engage Stack [[Sta](#)], the cloud service provider, for providing the necessary hardware and technical support.

## 1 Introduction

Trusted Execution Environments (TEEs) are increasingly important in machine learning and AI due to growing security requirements in both enterprise and decentralized applications [[SAB15](#), [MSM<sup>+</sup>18](#), [AKKH18](#)]. The introduction of TEE-enabled GPUs, such as the NVIDIA H100 and H200, adds an extra layer of protection for sensitive data but may impact performance. Understanding these trade-offs, particularly for large-scale machine learning tasks, is crucial for adopting TEE in high-performance AI applications [[YMY<sup>+</sup>22](#), [WO24](#)].

This report quantifies the performance overhead of enabling TEE on the NVIDIA Hopper architecture GPUs during LLM inference tasks, identifying where the overhead arises and under what conditions it can be minimized.

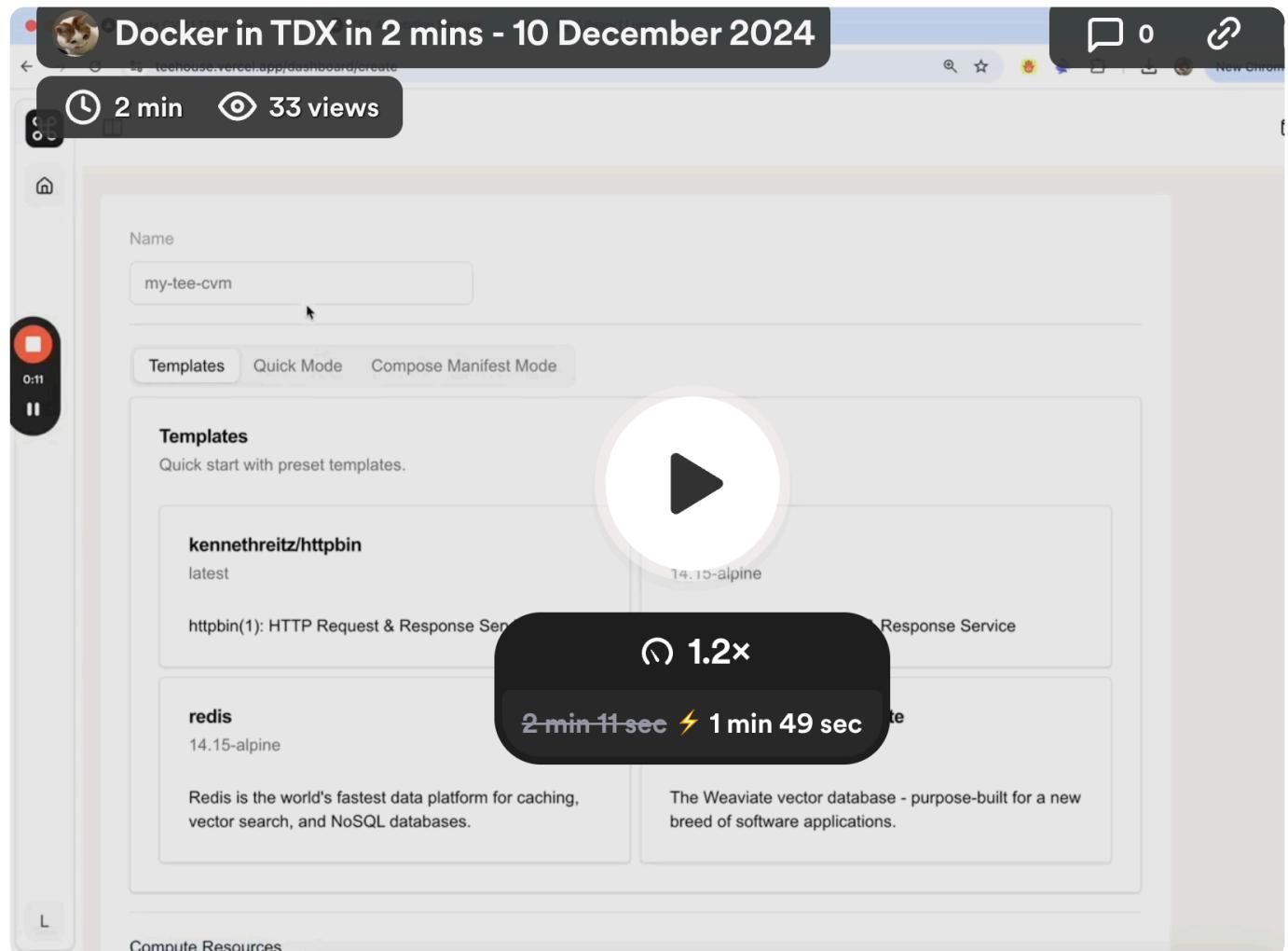
## 2 Background

### 2.1 Trusted Execution Environment

A TEE is a hardware-based security feature that isolates computations, preventing unauthorized access and tampering, even from the operating system or the physical hardware owner. As the core technology enabling Confidential Computing, TEEs create secure enclaves where sensitive data and code are processed with encryption, ensuring confidentiality and integrity even if the broader system is compromised [[SAB15](#)]. Traditionally implemented in CPUs, TEE technology was extended to GPUs by NVIDIA in 2023, enabling tamper-proof and confidentiality-preserving computation inside the GPU with minimal performance penalty [[DGK<sup>+</sup>23](#)].

## 6.2 Run Program in TEE Cloud

Phala offers a [TEE cloud platform](#) built on top of the open source [Dstack SDK](#), enabling the deployment of applications into both CPU TEE and GPU TEE. Whether the use case involves AI model inference or data-sensitive web3 applications. This cloud platform enables developers and organizations to leverage the benefits of TEEs without the complexity typically associated with secure computation, making it accessible for a wide range of applications.



# Conclusion

7

In the near future, we can envision a scenario where users can instruct AI to arrange travel plans and automatically book flights and hotels. Each session operates within a TEE, allowing a user's private state to be utilized across various companies, agents, and models. This setup mirrors the sharing of smart contract states on public blockchains, ensuring both privacy and security. Such interoperability is essential for establishing a genuine AI-driven economy that meets real business needs.



Especially AI agents, dAGI will significantly influence the crypto DeFi landscape. AI can assist humans in managing funds and making informed investment decisions by analyzing vast amounts of market data and trends. AI agents can also facilitate on-chain transactions, optimizing the speed and efficiency with which users interact with blockchain networks. Additionally, AI can enhance user experiences in on-chain gaming by providing strategic insights and automating gameplay actions, leading to more engaging and dynamic interactions.

Overall, the integration of AI in these areas underscores its potential to revolutionize how we interact with digital economies, offering more personalized, secure, and efficient solutions. This evolution is pivotal for the future of both AI and blockchain technologies, driving forward an interconnected, intelligent digital ecosystem.

# References

## 1. Phala Network Documentation

Phala Network. (2024). TEE-as-a-Service Documentation.  
Retrieved from <https://docs.phala.network/>

## 2. ai16z Eliza Framework

ai16z. (2024). Eliza: A Modular Framework for Autonomous AI Agents. GitHub Repository.  
Retrieved from <https://github.com/ai16z/eliza>

## 3. Dstack SDK

Dstack. (2024). Dstack SDK for Secure AI Deployments. GitHub Repository.  
Retrieved from <https://github.com/Dstack-TEE/dstack>

## 4. Spore.fun

Spore.fun. (2024). AI Agents Breed & Evolve.  
Retrieved from <https://www.spore.fun/>

## 5. Confidential Computing on NVIDIA Hopper GPUs: A Performance Benchmark Study

Doe, J., & Smith, A. (2024). Performance Analysis of Running AI in Trusted Execution Environments. arXiv preprint.  
Retrieved from <https://arxiv.org/pdf/2409.03992.pdf>

## 6. Trusted Execution Environments

Intel Corporation. (2023). Understanding Trusted Execution Environments: A Technical Overview.  
Retrieved from <https://software.intel.com/en-us/articles/intel-trusted-execution-technology>

## 7. Sentient White Paper

Sentient. (2024). Sentient: Advancing AI Autonomy and Security.  
Retrieved from <https://docsend.com/view/9kpwre9mtf6ectr>

## 8. OpenAI Techniques for Training Large Neural Networks

OpenAI. (n.d.).  
Techniques for Training Large Neural Networks  
Retrieved from <https://openai.com/index/techniques-for-training-large-neural-networks/>