# Infera Litepaper

The rapid growth of AI has led to increased demand for scalable and efficient inference solutions. AI ecosystems face challenges such as high costs, single points of failure, and privacy concerns. Current AI inference solutions are hindered by scalability issues and high operational costs, with some offering flexibility in AI model inference while others remain closed-source.

Infera addresses these challenges by decentralizing AI inference through distributing a secure, scalable, and cost-effective solution. Utilizing a decentralized network, Infera distributes AI inference tasks to worker nodes. Worker nodes are Infera ecosystem participants that earn INFER tokens through supplying the network with compute power. This architecture enhances the network's reliability and performance by distributing work through-out a network without relying on a single node.

Leveraging blockchain technology, Infera offers a platform and access to on-chain AI inference for developers working with blockchains. Additionally, Infera is launching a Layer 3 dedicated to on-chain inference work and an SDK to provide a cost-effective and scalable solution. This enables developers to build AI-powered applications for both on-chain and off-chain use cases.

# 1. Introduction

## 1.1. Problem & Innovations

AI inference is the process of using a trained machine learning model to make predictions or decisions on new, unseen data, typically in real-time environments. With AI inference demand increasing due to user adoption, a multitude of problems surrounding AI have emerged.

**Model Restrictions** - Centralized AI inference providers often impose limitations on the customization and deployment of AI models. This restricts users' ability to fine-tune models for specific applications and hampers innovation.

**Economic Barriers** - High costs associated with AI inference services and the necessary hardware infrastructure present significant barriers to entry, especially for startups and individual developers.

**On-Chain Verification** - Having on-chain verification allows for the development of on-chain AI agents. However, posting large volumes of inference data on-chain is impractical due to high costs and scalability issues. This limits the feasibility of fully decentralized AI solutions that require frequent data access and updates.

**Accessibility & Privacy** - Access to advanced AI inference remains limited to well-funded organizations and specialized tech companies, creating a disparity in AI innovation and application across different sectors unless those enterprises have their own proprietary solution.

While infrastructure is becoming more abundant, key barriers still hinder the development of applications leveraging AI inference. By addressing key challenges such as model restrictions, on-chain storage, and accessibility, we envisions users using AI inference to build tools users will be able to overcome the current issues surrounding access to AI inference.

## 1.2. Vision and Mission

Infera aims to democratize AI inference through a decentralized, blockchain-based platform that leverages global computational resources. Our vision includes on-chain verification mechanisms for transparent, auditable AI computations, creating a trustless environment for reliable AI systems. Infera's AI agent framework enables developers to build sophisticated, autonomous agents that can perform tasks, make decisions, and interact within our ecosystem. This paves the way for a new generation of decentralized applications powered by verifiable, autonomous AI.

# 2. Architecture Overview

## 2.1 Inference Network

The network bundles computational tasks submitted by users which are then dynamically distributed among participating nodes. Nodes will be able to perform inference and pick up failed jobs from other nodes. This prevents issues with single points of failure and increases network stability for users and developers. The decentralized architecture facilitates communications between nodes, enabling efficient distribution and execution of AI inference tasks across the network.

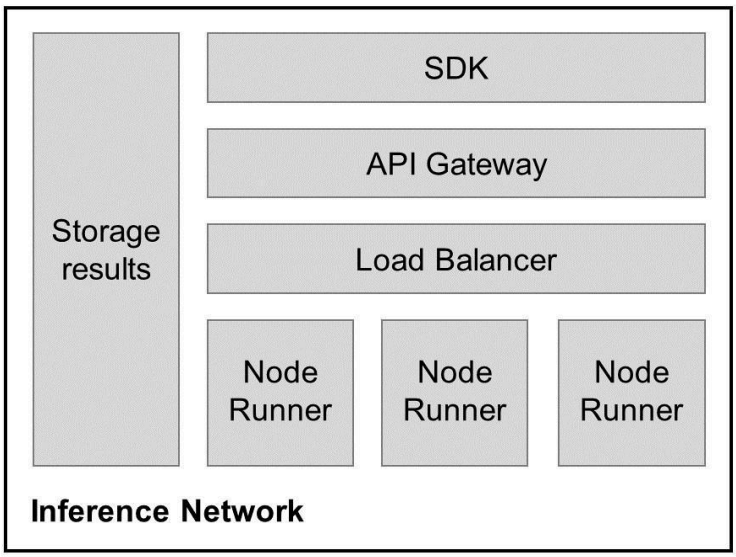


*Diagram 1 - Infera's inference network architecture*

Participants within the network include users and node runners. Demand is driven by users while node runners comprise the supply side of the network. Node runners are responsible for ensuring their GPUs and CPUs supplied have a stable operating environment to handle the inference requests generated by the users. Users will be able to access Infera through the SDK and API, accessing the inference network. Inputs submitted by users are sent to the load balancer from the API gateway, and are then routed to node runners. Users are able to obtain their results from the API Gateway

To incentivize participation, nodes contributing computational resources are rewarded with INFER tokens. This reward system encourages node workers to have stable uptimes to maintain a robust and active network. Additionally, the decentralized approach allows for greater flexibility and customization, including tools allowing users to deploy and fine-tune AI models to meet specific requirements without the constraints imposed by centralized providers.

## 2.2. Load Balancing

User inputs are dynamically distributed as batched tasks across active nodes on the network. By preventing overloading of individual nodes with the load balancer, Infera can maintain consistent performance even as the volume of tasks increases. The load balancer continuously monitors node performance and current workloads, adjusting the distribution of tasks in real-time. Infera will also be backstopping the network by providing worker nodes to the ecosystem to ensure that inference tasks on the network are completed.

Additionally, the load balancer is able to pick up dropped jobs from nodes and redistribute it across the network as a built-in redundancy to ensure that the network remains operational even in the event of node failures. If a node goes offline or experiences issues, the load balancer can quickly redistribute its tasks to other nodes, maintaining continuous operation and preventing disruptions.

## 2.3. On-chain Verification & Structure

A verified inference function (VIF) is a cryptographic primitive that produces a verifiable inference output from a given input and secret key. VIFs can provide proof that specific inference computations were performed correctly by generating verifiable random outputs associated with the computation. This proof can be stored and accessed on-chain, allowing anyone to verify that the inference results are genuine.To significantly reduce the cost of on-chain inference compared to existing methods, INFER will be implemented as the gas token for Infera's L3. This significantly reduces cost to use the network while also providing a dedicated environment for users to build on top of Infera.

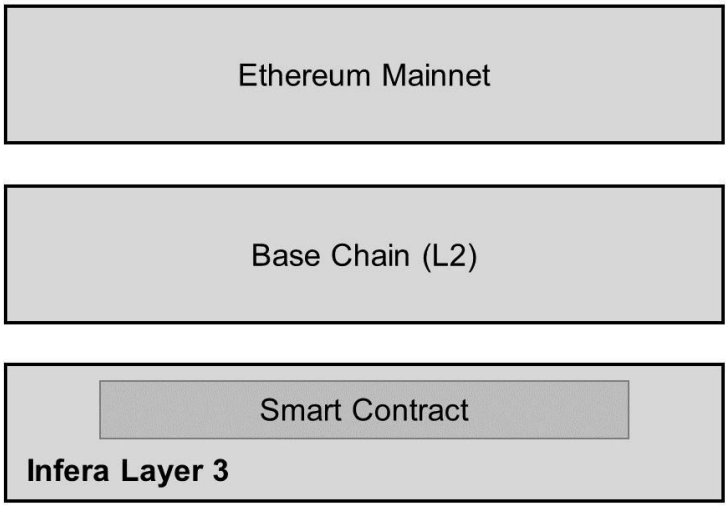


*Diagram 2 - Layered blockchain architecture with Infera's L3*

- **Ethereum Mainnet**

    - Ethereum Mainnet where all transactions get settled
- **Base Chain**
    - Aggregates and validates transactions from Layer 3.
    - Periodically commits these transactions to Ethereum
- **Infera Layer 3**
    - Optimized for specific use cases such as AI inference.
    - Inference network will be available via smart contract on Layer 3

# 3. Role of INFER Token

## 3.1. Gas Token for Transactions

INFER serves as the gas token for the L3, providing a cost-effective transaction mechanism compared to L2s and Ethereum mainnet. This approach is crucial for the network's economic viability, given the high volume of AI inference transactions. By using a dedicated L3 for on-chain inference, users avoid competition for blockspace on other chains, while still covering computational costs with INFER tokens for all transaction types.

## 3.2.Node Participation Rewards

Node participation is critical to the Infera network’s functionality and scalability. To incentivize participation, Infera rewards node operators with INFER tokens for their contributions. These rewards are given to nodes that contribute to the network for AI inference task completion, ensuring that there is sufficient computational power to meet demand.

The reward mechanism operates on a performance-based model, evaluating nodes on their uptime, reliability, and task volume. Rewarding nodes with INFER tokens incentivizes continuous participation and aligns operator interests. This mechanism also enables node runners to operate on the L3 when contributing inference to the network. A smart contract will be responsible for the distribution of rewards based on historical node performance..

# 3.3 Distribution & Staking Mechanism

INFER is distributed to node runners through our smart contract and off-chain tracking system. This approach allows for a flexible and scalable reward system during the early stages of network deployment. Nodes report their operational metrics to the off-chain system, which then calculates the appropriate rewards.

As the network matures, Infera will transition to a more decentralized and trustless model by integrating on-chain verified staking-based rewards. Node operators will be required to stake a certain amount of INFER tokens to their node. This staking mechanism serves a dual purpose: it secures the network by ensuring that node operators have a vested interest in maintaining its integrity and it enhances the transparency and security of the reward distribution process.

# 4. AI Inference Verification and Performance

## 4.1. Off-Chain Verification Using Vector Similarity

Infera employs cosine similarity for verifying AI inference results off-chain, a method that is both efficient and accurate in ensuring the integrity of non-deterministic AI outputs. This technique is particularly useful for comparing the similarity between two outputs, ensuring that the results of AI computations remain consistent and reliable.

**Cosine Similarity Mechanism -** Cosine similarity calculates the similarity between two non-zero vectors by measuring the cosine of the angle between them. For AI inference, the vectors represent the model's output features. A cosine similarity close to 1 indicates that the two vectors are nearly identical, confirming the consistency of the AI inference. This method is highly effective in handling the inherent variability in AI outputs, particularly for non-deterministic models that may produce slightly different results due to stochastic processes.

**Implementation in Infera -** Infera's off-chain verification process involves comparing the inference results from multiple nodes. When a node completes an inference task, it generates a vector representing the output. This vector is then compared to vectors produced by other nodes for the same task. If the cosine similarity between these vectors is above a predefined threshold, the result is considered verified. This approach ensures that only consistent and accurate inference results are committed to the blockchain, maintaining the integrity of the system.

## 4.2. Node Slashing

To maintain the integrity and reliability of the Infera network, nodes will be required to stake an amount of INFER tokens to be eligible for earning emission rewards.

**Staking and Slashing Mechanism** -Nodes that participate in the AI inference process must stake a predetermined amount of INFER tokens. This staked amount acts as collateral, which can be forfeited if the node fails to meet the required performance standards. Specifically, nodes that provide inference outputs that do not fall within the acceptable safety margin, as determined by cosine similarity, will be subject to slashing. This means a portion of their staked tokens will be deducted as a penalty for producing unreliable results.

**Performance Monitoring** - In addition to the staking requirement, a reputation mechanism is in place to continuously monitor node performance. Nodes are evaluated based on their uptime, the accuracy of their inference outputs, and their overall contribution to the network. This reputation score is crucial in maintaining a high standard of reliability and trust within the network. Nodes that consistently underperform or produce outputs with low cosine similarity scores will see their reputation scores decline.

**Consequences of Poor Performance** -Nodes with poor performance are penalized through the slashing mechanism. If a node's outputs frequently fall outside the acceptable range, indicating a significant deviation from the expected results, it will face incremental slashing penalties. This ensures that only nodes that consistently meet the network's performance standards can continue to participate and earn rewards.

**Reputation and Ejection** - Should a node's reputation score fall below a certain threshold due to repeated poor performance or malicious behavior, it will be ejected from the network. This ejection process ensures that the Infera network remains robust and secure by removing unreliable or compromised nodes. The reputation mechanism thus acts as a self-regulating system that maintains the quality and reliability of the network over time.

**Transparency and Fairness** - The staking, slashing, and reputation mechanisms are all governed by smart contracts, ensuring transparency and fairness in their execution. These smart contracts automatically enforce the rules without the need for human intervention, reducing the risk of bias or error.

### 4.3. Infera API and SDK

The Infera API and SDK constitute the primary interface for accessing our decentralized AI inference network. Our initial release focuses on a Python SDK, chosen for its widespread adoption in the AI and data science communities. The API design allows for developers to access the network of available nodes to perform inference work for AI applications built on top of the network. Future development will extend support to other programming languages, with JavaScript being the next target platform to broaden accessibility for web-based applications.

**Metrics for Success**

- Number of API calls and SDK downloads.
- Developer satisfaction and engagement levels.
- Growth in the number and quality of third-party applications and services built on Infera.

## 5. Use Cases

### 5.1. AI-Driven dApps

Infera is designed to support decentralized applications (dApps) that require real-time AI inference through its API. This capability is essential for applications such as recommendation systems and automated decision-making tools.

**Developer API and Model Diversity**

**API**: Infera offers a developer API that enables access to various AI models, allowing developers to integrate AI inference into their dApps.

**Model Integration**: The platform supports the integration of future AI models, expanding the range of possible use cases and enhancing adaptability to emerging technologies.

**Practical Applications**:

**Telegram Bots**: Developers can create intelligent bots for Telegram that leverage on-chain AI inference to provide real-time responses and personalized interactions.

**Web-Based Programs**: Web applications can utilize Infera's AI capabilities to deliver enhanced user experiences through personalized content and services.

**SDK Utilization**: The Infera SDK enables developers to build programs that seamlessly integrate on-chain AI inference, ensuring efficient and scalable AI-driven functionalities.

## 5.2. AI Research and Agent Framework

Infera will deploy an agent framework built on top of the inference network. This framework will enable users to build custom AI tools that combine multiple nodes for specialized computations.

For example, developers can leverage Infera to create AI-driven market-making algorithms and trading strategies. By utilizing real-time data and AI inference, these agents could optimize trading decisions, enhance liquidity, and stabilize markets. Researchers can build and deploy these models on-chain, ensuring transparency and trust through verifiable and immutable blockchain transactions and strategies.

Another application is on-chain loan monitoring and management. Developers can use Infera to deploy AI agents that continuously assess risk and predict default probabilities based on certain preset factors. This enables protocols to implement proactive measures, such as adjusting interest rates or restructuring loans, thereby minimizing risk and improving portfolio performance. The AI-driven approach enhances risk prediction accuracy and reduces manual effort in loan monitoring.

Infera's agent framework follows a perceive-think-act cycle:

- Perceive: Processes input data and updates the knowledge base.
- Think: Uses the decision engine to determine the next action.
- Act: Executes the chosen action and learns from the result.

To use this framework, developers would need to download our SDK and integrate our built-in functions for accessing the Infera network. The SDK provides easy-to-use methods for:

- Initializing an AI agent with specific capabilities and models
- Connecting to the Infera network and managing node interactions
- Submitting inference tasks and retrieving results
- Updating the agent's decision-making parameters

Our SDK will be available in python for developers to integrate Infera into their project. With just a few lines of code, developers can create sophisticated AI agents that leverage the power of Infera's decentralized inference network.

### 5.3. Enterprise AI Solutions

By integrating Infera's AI inference network, enterprises can develop and deploy sophisticated AI agents tailored to their business needs, enhancing operational efficiency and decision-making processes.

**Scalable AI Deployment** - Infera's platform allows enterprises to scale their AI deployments efficiently. The decentralized nature of the network means that enterprises can tap into a distributed pool of computational resources, reducing the need for expensive dedicated hardware.

**Customization and Flexibility** - The flexibility of Infera's architecture allows enterprises to customize the selection of AI models to meet specific requirements. Whether it's for predictive analytics, customer behavior analysis, or automated support systems, enterprises can build and deploy bespoke AI solutions that integrate with their existing system.

**Privacy -** Private nodes operate within the enterprise's local network perimeter ensuring that raw data never traverses public networks. Private nodes can operate in data center level infrastructure or consumer grade equipment.

## 6. Conclusion

In conclusion, by leveraging blockchain technology and distributed systems, we propose a solution that addresses key challenges in AI deployment, including scalability, security, and accessibility.

Key contributions of this work include:

1. A scalable architecture for distributed AI inference
2. Load balancing engine optimized for distributing AI workloads
3. An incentive structure promoting on-chain AI ecosystems

Further research is required to fully realize the potential of this technology. Future work will focus on optimizing network performance, enhancing security measures, and conducting large-scale deployments to validate our approach in real-world scenarios. As the AI landscape continues to evolve, Infera's decentralized platform has the potential to democratize access to AI capabilities, fostering innovation across various industries and research domains.