Sahara AI: The Decentralized Blockchain Platform for An Open, Equitable, and Collaborative AI Economy

Sahara AI Research
research@saharalabs.ai

September 1, 2024

1 Introduction

As Artificial Intelligence (AI) advances, we envision a future where AI is not only capable and ubiquitous, but also inherently fair, equitable, and accessible to all. In this future, AI will serve as a universal instrument that enhances lives and fosters global equality of opportunities. However, at present, control over AI technologies remains concentrated predominantly in the hands of a few oligopolistic organizations. These centralized AI platforms come with significant limitations and risks, including concerns for privacy, deepening economic disparities, and restricted access to resources. This creates significant barriers that hinder widespread innovation, limit participation in AI development from diverse backgrounds, and restrict access to AI technologies across different communities.

To empower everyone to own and shape the future of AI, we are building **Sahara AI**—a decentralized AI blockchain platform that supports an open, transparent, secure and inclusive AI ecosystem. At the heart of Sahara AI is the notion of "AI asset", a novel framework that defines the ownership and management protocols of private AI resources, such as personal data and proprietary models.

On our platform, AI developers, data providers and other stakeholders can collaborate to co-create high-quality AI assets using the platform's integrated development tools. Throughout this process, the platform ensures that all contributions are securely recorded and transparently attributed, with clear traceability of each participant's input. Once created, these AI assets are made available on the platform, where users can explore and leverage these resources. Users have the flexibility to purchase licenses for access and further development or even trade shares of the AI assets. Figure 1 presents an overview of this user journey, illustrating how AI assets move from creation to utilization and user engagement within the Sahara AI ecosystem. Notably, all transactions within the platform are immutable and traceable, with ownership protected and the origins of assets recorded. This supports a transparent and fair revenue-sharing model, ensuring that both developers and data providers receive appropriate compensation for their valuable contributions whenever revenue is generated.

Sahara AI leverages both blockchain technology and privacy protection methods to develop a robust provenance infrastructure for AI assets. This infrastructure enables the attribution of user contributions, protects data privacy, and ensures equitable compensation, all while emphasizing the importance of user control over their AI assets. Powered by these features, Sahara AI will deploy a unique, permissionless "copyright" system for all AI assets on its platform. While "copyright" traditionally implies restrictive control, which may seem contrary to the open nature of blockchains, this concept is redefined by Sahara AI's approach—it represents a tailored framework that ensures contributors retain ownership of their AI assets, along with fair attribution and compensation for their work—without limiting access and

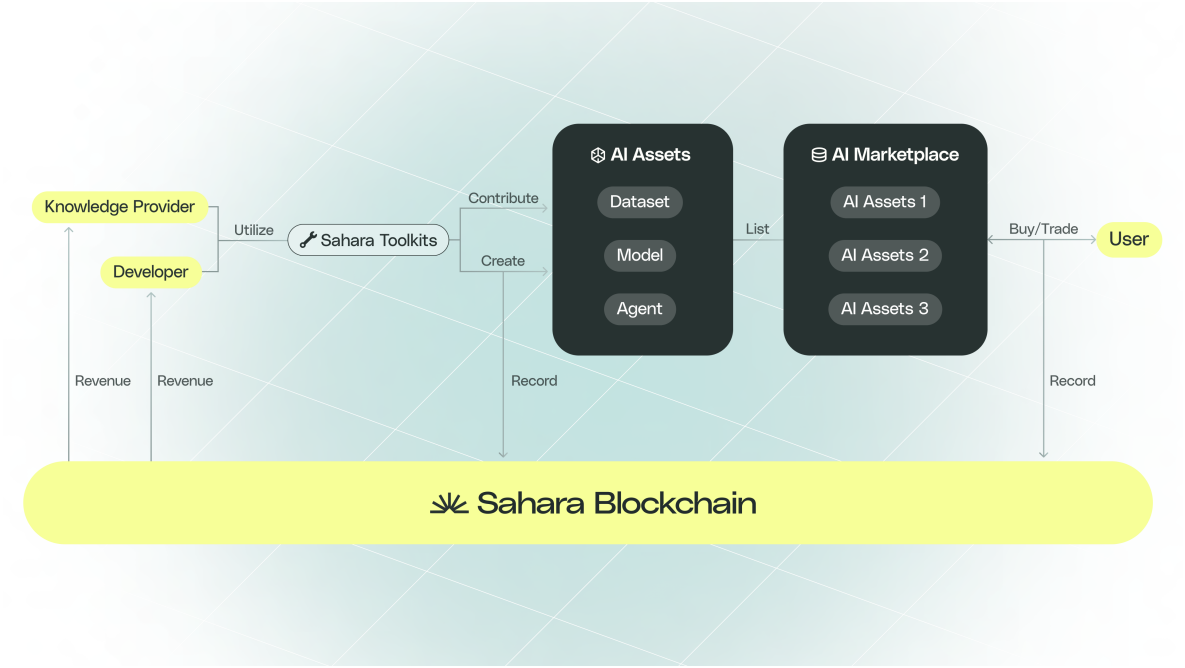

Figure 1: **Overview of user journey on the Sahara AI Platform.** This diagram demonstrates the key stages in the AI asset lifecycle within the Sahara AI ecosystem. It highlights how developers and knowledge providers collaborate to create AI assets, which are then listed on the AI Marketplace for users to monetize, with all transactions recorded on-chain.

sharing. This platform serves as a one-stop shop for all AI development needs throughout the entire AI lifecycle—from data collection and labeling, to model training and serving, AI agent creation and deployment, multi-agent communication, AI asset trading, and crowdsourcing of AI resources. By democratizing the AI development process and lowering barriers to entry found in existing systems, Sahara AI provides equal access for individuals, businesses and communities to co-build the future of AI.

Our platform encourages decentralized governance and community-driven innovation. This approach ensures that Sahara AI not only adapts to the evolving needs of the AI community but also leads the charge in setting new standards for ethical and equitable AI practices. By providing a collaborative environment for developing AI, Sahara AI allows individuals, small and medium-sized businesses (SMBs), and enterprises to work together, share ideas, and benefit from the collective intelligence and creativity of a global community.

On the journey to build Sahara AI, we are committed to transforming AI from an instrument controlled by the few into a resource that empowers all of humanity. Through the Sahara AI platform, we seek to break down barriers, foster global innovation, and unlock the full potential of AI for the betterment of society worldwide.

2 The Sahara AI Platform

The Sahara AI Platform is built on **three foundational pillars**: Sovereignty and Provenance, AI Utility, and Collaborative Economy. Together, these components create a cohesive platform where every participant can contribute, collaborate, and benefit.

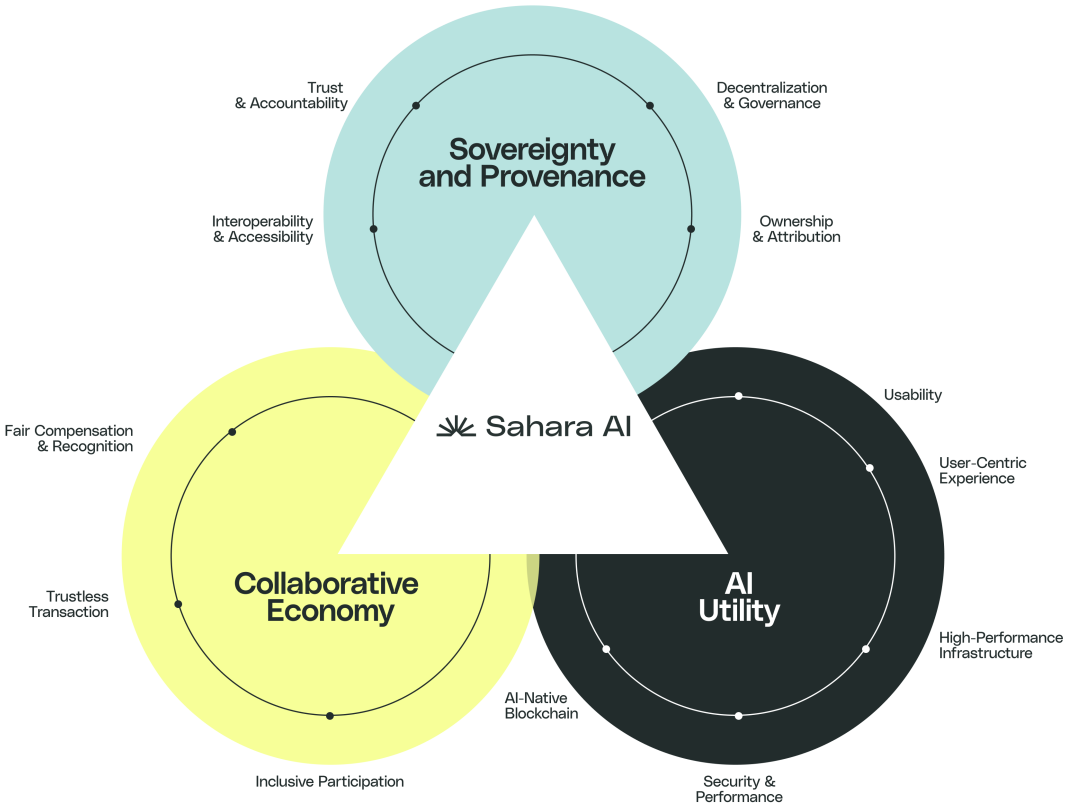

Figure 2: Three foundational pillars of Sahara AI

Pillar 1: Sovereignty and Provenance

In the evolving landscape of AI, Sahara AI emphasizes Sovereignty and Provenance as key principle that should define how AI assets and their development processes are owned, managed, and governed. These concepts ensure that all stages of the AI development cycle—from data collection and labeling, to model deployment and application building—are conducted in a manner that is decentralized, transparent, and inclusive.

Sovereignty encapsulates the idea that ownership and governance of AI assets should be decentralized and community-driven. This prevents monopolization and ensures that all stakeholders have a voice in the AI lifecycle. *Provenance*, on the other hand, ensures transparency in attributing contributions and tracking the origin and history of the usage and development of AI assets. It complements sovereignty by providing a comprehensive, immutable record of all activities and transactions associated with AI assets.

Sahara AI emphasizes the following critical aspects of sovereignty and provenance:

- **Ownership and Attribution:** Contributors to AI development (such as data providers, model trainers, and application developers) have verifiable, on-chain ownership and receive fair attribution for their contributions.
- **Decentralization and Governance:** Sahara AI promotes equitable and democratized control over AI assets. Actions and decisions on AI assets are made transparently through Sahara Blockchain Protocols and DAOs, ensuring all stakeholders have a say in the AI development cycle. Additionally, the evolution of AI components within the

platform is governed and driven by the community, allowing for a continuously evolving framework.

- **Trust and Accountability:** Detailed record-keeping ensures that every piece of data and every step in the AI lifecycle is meticulously recorded and traceable on the blockchain, allowing stakeholders to verify the sources and transformations of data and models.
- **Interoperability and Accessibility:** AI assets and services are designed to be interoperable across different platforms and accessible to a wide range of users, promoting inclusivity and broad participation in the AI ecosystem.

Pillar 2: AI Utility

Sahara AI empowers users throughout different stages of the AI lifecycle by leveraging a comprehensive technical infrastructure that delivers seamless AI user experience. We strive to ensure that every participant within the AI development cycle can efficiently develop, deploy, and manage AI assets in a trustless, privacy-preserving, and secure environment.

Our platform simplifies operations while embedding strong security measures to defend against unauthorized access and threats, alongside comprehensive privacy protections focused on safeguarding user information. These features not only protect user data but also build trust, enabling users to engage confidently and securely with AI technology.

To deliver maximum utility across all stages of the AI development cycle, Sahara AI focuses on five key aspects:

- **Usability:** Sahara AI streamlines the AI development cycle across processes from data curation, to model development, and agent deployment. This allows participants to leverage AI to enhance productivity, create high-utility applications and achieve positive real-world outcomes.
- **User-Centric Experience:** Sahara AI provides an out-of-the-box experience for all participants in the AI development cycle. Regardless of technical expertise, every participant can effortlessly engage with AI technology.
- **Security and Privacy:** Users benefit from state-of-the-art security measures and privacy protection. Users can confidently manage their AI assets and computations without compromising usability.
- **High-performance Infrastructure:** Sahara AI's infrastructure supports cutting-edge AI paradigms and offers a comprehensive toolkit for users to work on advanced AI models and applications.
- **AI-Native Blockchain:** Sahara AI is built on the Sahara Blockchain, a Layer 1 blockchain specifically designed with built-in protocol and precompiles for AI transactions across the entire AI lifecycle on the Sahara AI platform.

Pillar 3: Collaborative Economy

The Sahara AI collaborative economy is designed to enable monetization and attribution, ensuring that all participants are rewarded for their contributions. This means:

- **Fair Compensation and Recognition:** Users are rewarded in proportion to their contributions based on provenance of the AI development process, addressing challenges of global economic disparities.

- **Inclusive Participation:** The collaborative economy attracts simultaneous participation from individuals, SMBs, and enterprises, promoting a diverse and vibrant AI community.
- **Trustless Transactions:** The Sahara AI platform enables users to monetize their AI assets through a transparent and efficient process.

3 Design

Building on these three pillars, Sahara AI offers a platform where every participant can contribute, collaborate, and benefit. The platform is built on a layered architecture designed to securely and comprehensively support users and developers throughout the entire AI lifecycle. As illustrated in Figure 3, the Sahara AI platform is structured into four layers:

- **The Application Layer** acts as the user interface and the major interaction point for the Sahara AI platform. It provides native built-in applications for users to build and monetize AI assets.
- **The Transaction Layer** features the Sahara Blockchain—a Layer 1 blockchain infrastructure that manages provenance, access control, attribution and other AI-related transactions across the AI lifecycle.
- **The Data Layer** provides abstractions and protocols for data storage, access, and transfer. It integrates both on-chain and off-chain components to provide seamless data management across the AI lifecycle.
- **The Execution Layer** offers essential off-chain infrastructure to support AI utility, covering a full spectrum of AI functionalities. It provides versatile AI computation protocols and dynamically allocates computational resources to maximize performance, ensure scalability, and enhance robustness.

3.1 Application Layer

The Application Layer of the Sahara AI platform serves as the primary interface for participants of the platform, providing the natively built-in toolkits and applications to enhance user experience. At its core, the Application Layer is designed with components for participants to maximize their engagement within the AI ecosystem regardless of technical expertise.

3.1.1 Functional Components

The functional components are the foundational elements that support the robust operations and security measures of the application layer. These components are designed to ensure the safe storage, efficient management, and effective deployment of AI.

Sahara ID Sahara ID serves as the cornerstone of identity management within the Sahara AI platform. It act as the unique identifier for all participants, whether they are AI entities or human users. This system provides robust identity verification and reputation management, ensuring secure and transparent interactions across the platform. Through their Sahara ID, participants gain secure access to AI assets they own or are licensed to, and can track and manage their contributions and reputation within the ecosystem. Sahara ID also plays a crucial role in facilitating attributions. It meticulously records each participant's contributions to AI projects, ensuring that attribution is clearly tracked and AI "copyrights" are upheld.

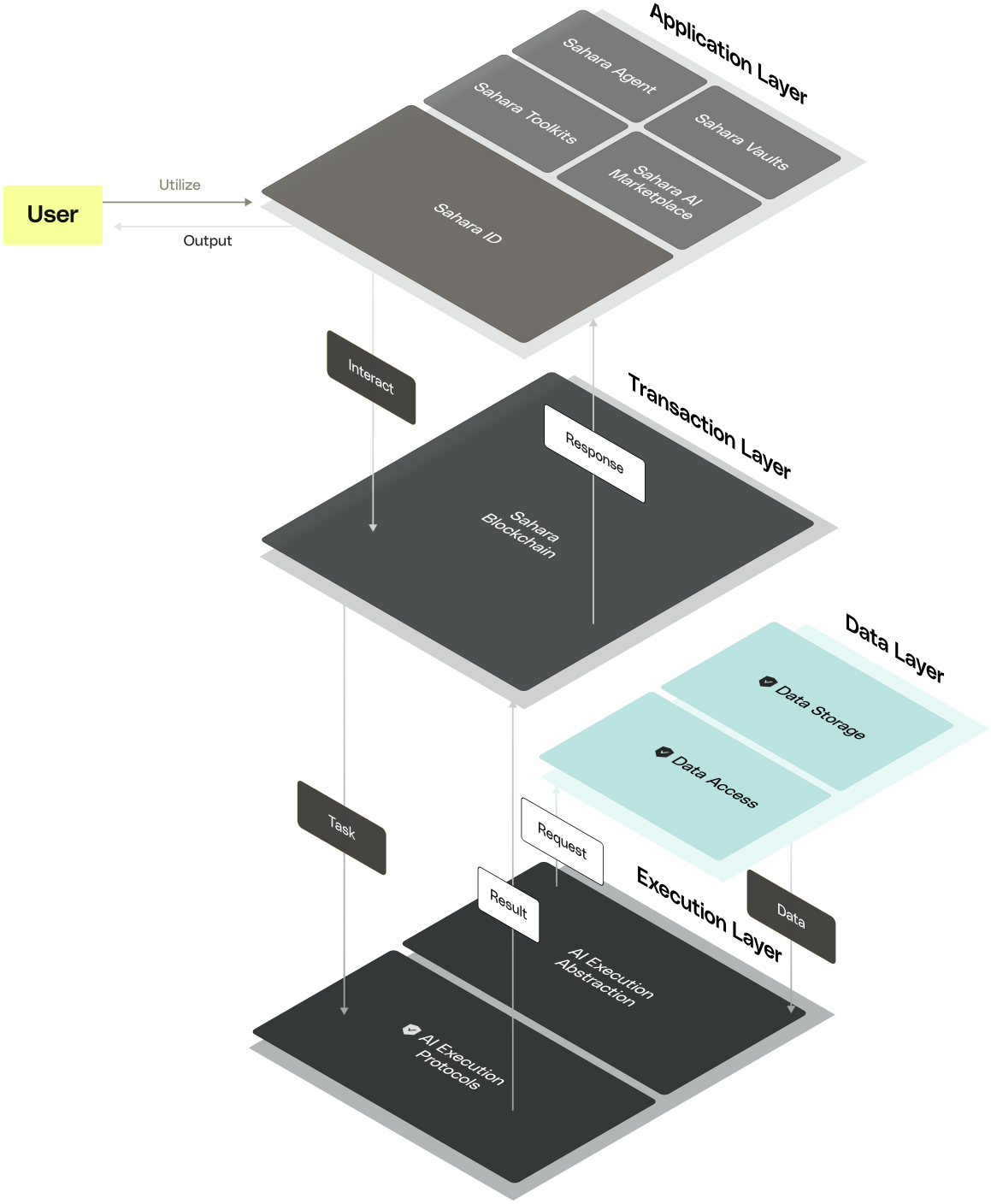

Figure 3: This layered diagram illustrates the technical architecture of the Sahara blockchain platform which consists of four inter-related layers. This hybrid infrastructure with both on-chain and off-chain protocols allows users and developers to effectively contribute to and benefit from the entire AI development cycle.

Sahara Vaults Sahara Vaults are private and secure repositories storing and managing AI assets, covering both local storage on user nodes and cloud storage on public nodes. These vaults offer advanced security features, ensuring that all data and assets are protected against unauthorized access and potential threats. Sahara Vaults strives maintain the privacy, security and integrity of proprietary AI assets.

Sahara Agent Sahara Agents are AI-driven entities within the Sahara AI platform, each composed of three integral components—Brain, Perceptor, and Actor. Each component is specifically designed to perform distinct functions:

- Brain: The strategic core responsible for thought, memory, planning and reasoning. It processes information and makes informed decisions. Features include:
  - Persona Alignment: Tailors the agent's responses and behaviors to align with specific user personas, ensuring customized interaction.
  - Lifelong Learning: Employs feedback mechanisms and reinforcement learning to continually enhance its capabilities over time.
- Perceptor: Handles input from various sources, analyzing and interpreting data to inform the Brain's decisions. It features:
  - Multimodal Perception: Processes and interprets multiple types of data inputs, including, but not limited to, visual and auditory.
- Actor: Take the actions as determined by the Brain, based on insights provided by the Perceptor. It features:
  - Tool Utilization: Leverages a wide range of tools and resources to execute actions, such as searching the web.

3.1.2 Interactive Components

The interactive components of the application layer directly facilitate user interaction with other users and AI entities, enabling users to actively engage with and utilize AI assets on the platform. These components make the platform's capabilities accessible and practical for a diverse range of applications.

Sahara Toolkits Sahara Toolkits are development and deployment tools designed for participants who aim to create and refine AI assets on the Sahara AI platform. These toolkits cater to a diverse audience by bridging the gap between technical expertise and innovative execution. For technical users, we provide the Sahara SDK & API, which includes tools for programming, integration, and customization, allowing for the development of complex AI functionalities tailored to diverse needs. For users who are less tech-savvy, the Sahara No-Code/Low-Code toolkits make AI development accessible through intuitive interfaces and pre-built templates. These platforms empower all users, regardless of their technical skill level, to participate actively in creating and deploying AI assets.

Sahara AI Marketplace Sahara AI Marketplace is the decentralized hub for publishing, monetizing and trading of AI assets. It features a comprehensive portfolio of high-quality AI assets, including proprietary AI agents, models, and datasets. The marketplace integrates seamlessly with Sahara ID to facilitate ownership protection and access control, leveraging

blockchain technology to ensure transparency and security in transactions as well as provenance. Additionally, it offers dynamic licensing and various monetization options, providing flexible and innovative solutions to meet the diverse needs of users.

3.2 Transaction Layer

The Transaction Layer of the Sahara AI Platform features the **Sahara Blockchain**. The Sahara Blockchain is a Layer 1 blockchain meticulously designed to meet the platform's comprehensive needs. It comes with protocols for managing ownership, attribution, and a wide range of AI-related transactions on the platform. The Sahara Blockchain is pivotal in upholding the sovereignty and provenance of AI assets.

3.2.1 Sahara Blockchain AI Native Features

The Sahara Blockchain distinguishes itself as an AI-Native blockchain through its integration of specialized Sahara AI-Native Precompiles (SAPs) and Sahara Blockchain Protocols (SBPs) that support essential tasks throughout AI lifecycle tasks.

Sahara AI-Native Precompile The Sahara Blockchain integrates Sahara AI-Native Precompile (SAP), which are built-in functions that operate at the native level of the blockchain. These SAPs are precompiled to enable faster execution, lower computational overhead, and reduced gas costs.

- **Training Execution SAPs:** This class of SAPs facilitates the invocation, recording, and verification of off-chain AI training processes. They enable the blockchain to interface seamlessly with off-chain training environments and ensure that training activities are accurately recorded. Additionally, these SAPs verify the integrity and authenticity of the training computations performed outside the blockchain. Their aim is to enhance the credibility and reliability of AI models developed within the Sahara AI Platform.
- **Inference Execution SAPs:** Inference Execution SAPs support the invoking, recording, and verification of AI inference results generated off-chain. These SAPs invoke off-chain AI computations required for inference and provide a robust mechanism to verify that the AI inferences, which may include predictions, classifications, or other outputs, are derived from legitimate computations. By integrating these functionalities, the Sahara Blockchain ensures that all AI inference activities are transparent, verifiable, and accountable.

Sahara Blockchain Protocols The Sahara Blockchain also implements AI-specific protocols through smart contracts, collectively known as Sahara Blockchain Protocols (SBP). These SBPs provide a structured and secure framework for managing various aspects of the AI lifecycle. They ensure that AI assets and computation results are handled transparently and reliably.

- **AI Asset Registry SBPs:** These protocols manage the initial registration and tracking of AI assets on the blockchain. They establish a comprehensive ledger that uniquely identifies AI models, datasets, agents, and other AI-related assets, ensuring their provenance is verifiable. The AI Asset Registry SBPs focus on the static aspects of AI asset management, such as recording the creation, identity, and origin of assets.
- **AI Licensing SBPs:** The AI Licensing SBPs decide on-chain rights to access or utilize AI assets. It enforces access control with different types of on-chain licenses (see

section 3.5.3 for details) ensuring that only authorized entities can use or access specific AI assets. This protocol helps maintain the security and proper usage of AI assets, facilitating compliant and controlled distribution of AI capabilities.

- **AI Ownership SBPs:** These protocols maintain clear, non-transferable, and non-fungible ownership records of AI assets. It ensures that the ownership details of AI models and datasets are securely stored on-chain, providing indisputable proof of ownership. This helps in promoting the "copyright" of AI and managing asset ownership transparently.

- **AI Attribution SBPs:** The AI Attribution SBPs track ongoing contributions throughout the AI lifecycle and manage the distribution of rewards based on these contributions. These protocols ensure that all inputs to the development and evolution of AI assets are recorded and that any revenue generated is fairly allocated to contributors.

3.2.2 Sahara Blockchain Design

The Sahara Blockchain employs a *Proof-of-Stake* consensus mechanism and focuses on providing a suite of AI native features to support application layer and work seamlessly with the execution layer to achieve strong AI utility. We utilize the Tendermint algorithm [1] for Byzantine Fault-Tolerant consensus. This ensures a high degree of fault tolerance and enables the network to reach consensus even in the presence of malicious nodes. Additionally, the Sahara Blockchain follows a modular design to offer flexibility in administrative domains and supports scalable solutions. Our design choice aims to address the following chain features:

- **Efficiency:** Building on top of the Tendermint algorithm, the Sahara Blockchain not only enjoys the inherent benefits of high-performance properties such as fast block confirmation time, quick average block time, and near-instant finality, but also enhances them further by introducing additional optimizations tailored for real-time data processing and high-speed transactions.

- **Scalability:** Our modular design supports horizontal scalability and off-chain scaling solutions, such as Layer 2 solutions. We plan to support rollup solutions to alleviate the load on the main chain to ensure efficient scalability as demand grows. Rollups aggregate multiple transactions off-chain and submit them to the main blockchain, thereby significantly enhancing throughput and lowering transaction costs. Our approach aim to maintain high levels of security and decentralization while providing a robust solution for increased usage in the future.

- **Interoperability:** We design Sahara Cross-chain Communication (SCC) Protocol to facilitate seamless interactions with other blockchains. Our protocol allows for the secure and permissionless transfer of any type of data encoded in bytes between different blockchains without relying on third-party intermediaries. By leveraging SCC, we ensure trustless and permissionless interactions, where any party can operate a relayer to transfer information between blockchains. Additionally, it supports cross-chain bridges to enable the seamless transfer of assets between different blockchain networks.

- **EVM-Compatibility:** The Sahara Blockchain's built-in virtual machine is fully compatible with the Ethereum Virtual Machine (EVM). This compatibility allows developers to leverage the extensive ecosystem of Ethereum tools, resources, and community support, including frameworks like Solidity. Developers can effortlessly write and deploy smart contracts on the Sahara Blockchain, ensuring they can also operate across other

EVM-compatible blockchains with minimal to no code modifications. We provide developers with a familiar and powerful environment to build decentralized applications efficiently on Sahara.

- **Low Gas Fees:** The Sahara Blockchain implements a highly efficient fee structure to minimize transaction costs. By optimizing transaction batching and utilizing dynamic fee mechanisms, we aim to keep gas fees economically viable, even as network demand grows. We are striving to achieve cost-efficiency on our blockchain to make our platform more accessible and attractive for developers and users in an effort to foster greater participation and engagement within the ecosystem.

3.3 Data Layer

The Data Layer in the Sahara AI platform is an abstraction designed to optimize data management throughout the AI lifecycle. It acts as a vital interface that connects the execution layer to diverse data management mechanisms and seamlessly integrating both on-chain and off-chain data sources. This conceptual layer not only ensures efficient access to data but also enhances system integrity and performance.

3.3.1 Data Components

On-chain Data On-chain data includes, but is not limited to, critical AI asset metadata, attributions, commitments, and proofs. This ensures all contributions and interactions within the Sahara AI platform are transparent and accountable.

Off-chain Data Significant datasets, AI models, and supplemental information are stored off-chain due to storage limitations and cost considerations associated with on-chain data management. By utilizing off-chain storage solutions, we can handle vast amounts of data without compromising performance while remaining efficient and cost-effective.

3.3.2 Data Management

Security At Sahara AI, we prioritize security above all to protect our platform and user data. We implement a comprehensive suite of security measures to ensure robust protection across all interactions:

- Advanced Encryption: We secure all data using the latest encryption practices, which ensures that sensitive information remains protected both in transit and at rest. This level of security is crucial for maintaining confidentiality and integrity within our ecosystem.

- Access Controls: In synergy with AI licensing SBPs, our platform implements on-chain licenses that serve as stringent access control systems. These controls strictly limit data access to authorized personnel only, significantly enhancing the security and compliance of our platform. Furthermore, this approach leverages the inherent properties of the Sahara Blockchain, such as transparency and immutability, to ensure that these controls are not only effective but also verifiable and secure.

- Private Domain Storage: We enable users to store their data in private domains, which offer enhanced security features while allowing seamless interaction with our platform. This capability ensures that users retain control over their sensitive data, while still benefiting from the robust functionality and connectivity of our platform.

Data Availability At the Sahara AI platform, we proactively tackle the Data Availability (DA) problem by implementing off-the-shelf solutions that ensure all block data is verifiably accessible to all network participants. This strategy is critical not just for maintaining network integrity and trust, but also for enhancing our blockchain's scalability. By enabling efficient and reliable data verification, particularly for light clients who do not store the entire blockchain, we significantly reduce the network's dependency on full nodes. This approach minimizes bandwidth and storage demands, and facilitates smoother and more scalable network operations. Integrating these DA solutions into our Data Layer boosts overall system performance and enables the network to handle increased volumes of transactions and data without compromising speed or security.

Indexing In the Sahara AI platform, we actively enhance on-chain data management with advanced indexing techniques tailored to our blockchain architecture. These methods significantly improve data retrieval speeds and query efficiency. We leverage off-the-shelf decentralized indexing solutions alongside our proprietary solutions to boost scalability and performance. This integration within our Data Layer ensures seamless interactions between execution layers and data sources. Our approach enables the platform to manage increasing data volumes and complex queries efficiently.

Storage Our storage strategy for off-chain data employs a hybrid model that combines the strengths of both decentralized and centralized systems to optimize cost-efficiency and scalability. We use both decentralized storage solutions like IPFS for critical data that benefits from immutability and distributed hosting, while leveraging more traditional cloud storage solutions for large volumes of data where speed and availability are paramount. This dual approach allows us to effectively manage data storage costs while ensuring high availability and rapid access.

3.4 Execution Layer

The Execution Layer is the off-chain AI infrastructure of the Sahara AI platform that interacts seamlessly with the Transaction Layer and Data Layer to execute and manage protocols related to AI computation and functionality. Based on the execution task, it securely pulls data from the Data Layer and dynamically allocates computational resources for optimal performance. During execution, the Execution Layer leverages versatile protocols that are efficient, private, and integrity-preserving. It interacts with the Sahara Blockchain to record all execution activities and proofs, ensuring provenance and trust. Furthermore, our AI infrastructure supports high performance, meaning that it is expedient, elastic, and resilient.

3.4.1 High Performance Infrastructure

Sahara AI's Execution Layer's underlying infrastructure is designed to support high-performance AI computation with the following properties:

- **Expedient:** The Execution Layer is building to ensure rapid and reliable performance by efficiently coordinating AI computations across various contributors and participants within our infrastructure.
- **Elastic:** To handle varying levels of traffic, the Execution Layer is building robust autoscaling mechanisms. This feature will ensure our infrastructure remains highly available under high traffic conditions.

- **Resilient:** The Execution Layer is built with fault tolerance to ensure system stability and reliability. Further complemented by the Sahara Blockchain, our infrastructure is partition tolerant. In the event of failures, the system can quickly recover to maintain the integrity of the processing workflow and minimizing downtime.

3.4.2 Abstractions

Abstractions are foundational to the implementation of various AI assets on the Sahara AI platform. They provide the necessary conceptual framework for managing datasets, AI models, computation resources, vaults, and AI agents.

Core Abstractions Core abstractions are essential components that form the basis of AI operations on the Sahara AI platform.

1 **Datasets** The datasets abstraction represents the curated data from our participants that fuel AI training and augment inference processes.

2 **AI Models** The AI Model abstraction encapsulates the models and architectures for a variety of machine learning tasks. Our primary models of interest are generative models [2, 3, 4, 5], with an emphasis on large language models (**LLM**s) [6, 7, 8]. Sahara AI optimizes specifically for transformer-based models [9, 10].

3 **Computation** The computation abstraction encompasses the resources that supports the execution of AI tasks. Sahara AI's infrastructure dynamically allocates these resources for optimal performance and cost-efficiency. This includes not only cloud-based GPUs, but also decentralized contributions from network participants.

High-Level Abstractions High-level abstractions build upon core abstractions and provides higher-level functionalities and integrations on the Sahara AI platform.

1 **Vaults** Vaults abstractions are the execution interfaces behind Sahara Vault. They are structured to manage the accessibility and effective use of AI assets and are crucial for operational processes.

2 **AI Agents** AI agents abstractions are the exeuction interfaces behind Sahara Agent. Specifically, we are interested in LLM-based agents [11, 12, 13] where LLMs are the brains. These abstractions enable the agents to perform complex reasoning, engage in natural language interactions, and execute a range of cognitive and decision-making tasks efficiently [14, 15, 16, 17].

3.4.3 Protocols for the Execution Layer

The Execution Layer of the Sahara AI platform orchestrates complex AI operations through a suite of specialized protocols designed to facilitate efficient interactions among various abstractions. These protocols manage a wide range of activities, including AI access, training and execution. They work seamlessly with Sahara Blockchain Protocols (SBPs) to meticulously record execution details. This includes, but not limited to, tracking who performed specific actions, what resources were accessed, and the contributions made by different entities.

Abstraction Execution Protocols Abstraction execution protocols provide the necessary framework to ensure that all high-level abstractions operate efficiently and securely.

- **Vault Execution Protocols** The Vault Execution Protocols standardize interactions with the vaults within the Sahara AI platform's Execution Layer, detailing how vaults are accessed and utilized for various AI processes.
  - **Direct Access Protocol:** This protocol allows users to perform targeted queries on the vault to retrieve specific pieces of information quickly and securely.
  - **Downstream Model Training Protocols:** These protocols facilitates the use of stored data within the vault for training downstream AI models. It supports a diverse array of training paradigms [10, 6, 18, 19] from the use of both unstructured data for pre-training to structured data for subsequent fine-tuning.
  - **Retrieval-Augmented Generation (RAG) Protocol:** The RAG [20, 21] protocol harnesses relevant data retrieved from the vault to enhance the outputs of generative models.
- **Agent Execution Protocols** The Agent Framework Protocols manage the interactions and coordination of AI agents within the Execution Layer. These protocols include:
  - **Communication Protocols:** Sahara AI supports a range of communication strategies to suit different coordination needs:
    - **Hierarchical Communication:** In this strategy, one agent acts as a leader, with other agents reporting to it [22, 23, 24].
    - **Peer-to-Peer Communication:** In this strategy, all agents interact as equals, sharing information and making decisions collaboratively [12, 25, 26].
  - **Multi-Agent Coordination Protocols:** Effective coordination and orchestration of multiple agents are essential for achieving complex objectives. Within the Sahara AI platform, these protocols involves sophisticated task allocation, synchronization, evolution, and collaboration strategies.

Collaborative Computation Protocols The Execution Layer introduces Collaborative Computation Protocols to facilitate joint AI model development and deployment among multiple participants. Under this collaborative framework, our solution includes privacy-preserving compute modules and computation fraud proof mechanisms to ensure privacy and integrity throughout the AI development process, as well as Parameter-Efficient Fine-Tuning (PEFT) Modules for efficient tuning of AI models.

- **Collaborative Model Training Protocols** Collaborative model training is a key component that allows multiple users or systems to jointly build AI models. This includes:
  - **Decentralized Training:** Sahara AI is building to support decentralized training, which allows participants to contribute computational resources [27] and train models collectively as a distributed training system [28].
  - **Model Aggregation:** Beyond decentralized training, Sahara AI is also building to support model aggregation with cutting-edge model merging techniques [29, 30, 31] to allow multiple participants to combine their proprietary models into a single, collective model.

- **Collaborative Model Serving Protocols** Collaborative model serving allows multiple users or systems to jointly serve AI models for inference. This includes:
  - **Decentralized Serving:** Similar to decentralized training, participants will be able to host segments of models for serving [32, 33, 27]. We are building this to be scalable and efficient to foster seamless collaboration among participants.
- **Add-on Modules**
  - **PEFT Modules** Sahara AI is building to support various PEFT techniques [34, 35, 36], such as Low-Rank Adaptation (LoRA) [37], to efficiently and portably customize AI models. These PEFT modules enhance model performance with minimal computational overhead and can be easily integrated as plug-and-play components for large models as LLMs.
  - **Privacy Preserving Compute Modules** Privacy-preserving compute modules within the Execution Layer ensure that proprietary data and model parameters remain secure and confidential throughout the processing lifecycle. To cater to a wide range of use cases and scenarios, we offer a suite of privacy-preserving techniques:
    - **Differential Privacy (DP):** Employed during both inference and training to add noise to data or computations [38, 39, 40, 41]. This ensures that individual data points cannot be distinguished.
    - **Homomorphic Encryption (HE):** Used for scenarios involving two-party computations, lightweight computation tasks, and preprocessing computations such as random number generation for larger models [42].
    - **Secret Sharing (SS):** Utilized for general transformer structure inference tasks [43, 44]. SS divides data into parts and distributes them across different nodes, preventing any single node from accessing the complete data.
  - **Computation Fraud Proof Modules** We are also building to support fraud proofs generation [45, 46] for computation results from our collaborative computation protocol. These proofs will be verified on-chain to protect against malicious or faulty computations.

3.4.4 Integrations

The Execution Layer integrates seamlessly with other layers of the Sahara AI platform to provide a cohesive and efficient AI infrastructure.

- **Transaction Layer** The Execution Layer collaborates with the Transaction Layer to manage the sovereignty and provenance of AI assets. All execution, contribution, and usage activities involving AI assets—such as model access, data usage, and computation results—that occur on the execution layer are logged and sent to the Sahara Blockchain for recording. This is facilitated by SAPs and SBPs.
- **Data Layer** The Execution Layer utilizes the vault abstraction and corresponding protocols to interact with the Data Layer, securely accessing data from vaults for training and RAG. This integration leverages mechanisms from both the Data Layer and the Execution Layer's computation protocols to ensure privacy, security, and integrity.

3.5 Economic System

The economic system of Sahara AI is designed to create a collaborative, fair, and rewarding environment where all participants can benefit from their contributions. Whether it's operating a node, contributing knowledge and data, fine-tuning a model, or building an autonomous agent, Sahara AI ensures fair compensation, transparent transactions, and inclusive participation. By leveraging Sahara Blockchain and its SAPs, Sahara AI provides a robust system that upholds these principles, fostering trust and equitable opportunities for all.

AI asset Investment Model Sahara AI further enhances the collaborative economy by adopting an innovative investment model. This model allows users to invest resources such as capital, data, compute, and technical AI expertise in exchange for stake in AI models, agents, and applications. This mirrors the strategies of major GPU/cloud giants like Microsoft and Amazon, who invest in foundational model companies such as OpenAI and Anthropic. These giants offer GPU resources in exchange for stake, which in the future secures their vested interest in AI advancements. This approach not only provides AI developers with essential computational power but also aligns the interests of hardware suppliers with the success of foundational model companies. As a result, it creates a mutually beneficial ecosystem that accelerates AI innovation while potentially reshaping industry dynamics. By extending this model across the entire AI lifecycle—from data to compute, model development to final application—Sahara AI enriches the collaborative economy, making it more dynamic and inclusive.

3.5.1 Economic Roles

In the economic system, the Sahara AI platform includes the following important roles:

- **Developer:** Developers create AI models, tools, and applications on the Sahara AI platform. Their incentives include revenue through usage fees, licensing, and sales of their creations.
- **Knowledge Provider:** Knowledge providers consist of annotators and reviewers. Annotators are responsible for providing personalized data and professional knowledge, while reviewers ensure the quality of the collected data. They are compensated based on the quality and utility of the data they provide.
- **Consumer:** Consumers are end-users and businesses that utilize AI models, tools, and applications from the platform. They pay for services based on usage, which ensures a sustainable economic cycle. Consumers benefit from access to cutting-edge AI solutions that can enhance their operations and drive innovation.
- **Investor:** Investors provide capital and resources (GPUs, Cloud servers, RPC nodes, etc.) to support the development and deployment of AI assets. In return, they receive the stake of the AI assets, earn revenue generated by AI assets and potential income from the price appreciation of the AI assets.
- **Operators:** Operators are responsible for contributing storage and computing power and maintaining the network infrastructure, both on-chain and off-chain. They ensure the smooth operation and security of the platform. Operators are incentivized through rewards for their role in maintaining network stability and performance.
- **Validators:** Validators maintain the integrity and security of the Sahara Blockchain. They verify transactions and secure the network. Validators earn rewards for their critical role in ensuring the trustworthiness and reliability of the blockchain.

3.5.2 Growth Flywheel

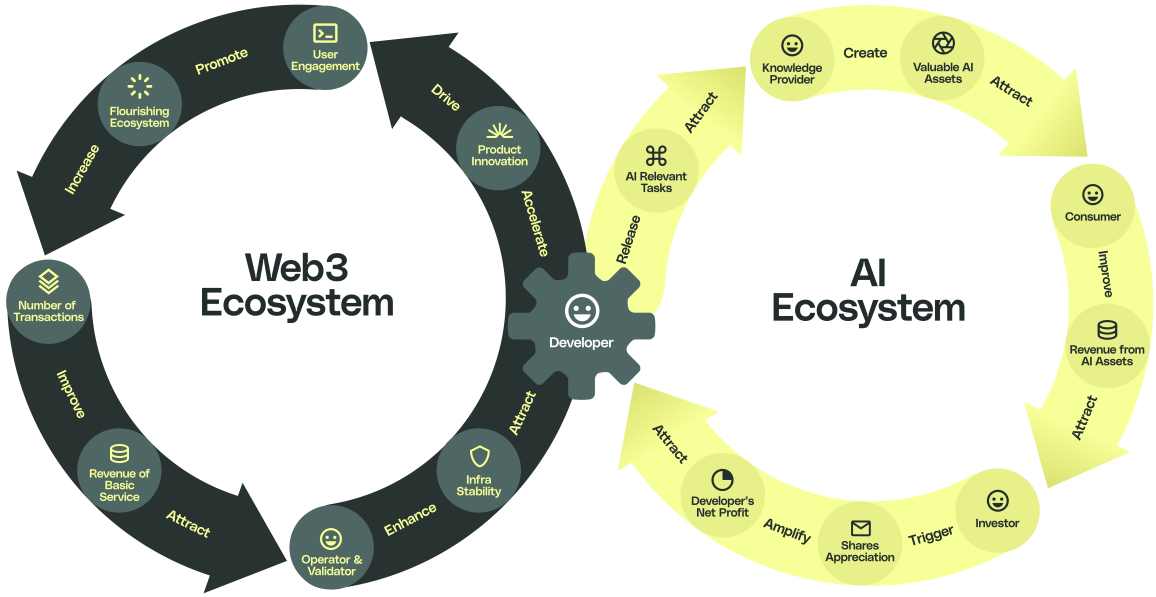

Figure 4: Sahara Dual Growth Flywheel

Overall, as illustrated in Figure 4, Sahara AI employs the dual growth flywheel model to drive the sustainable and scalable growth in Web3 ecosystem and AI ecosystem. This dual approach ensures that both Web3 ecosystem and AI ecosystem seamlessly work in synergy under a self-sustaining loop.

AI Ecosystem The AI ecosystem growth flywheel starts with developers who release and engage in AI relevant tasks. These AI relevant tasks attract knowledge providers to co-operate with developers creating AI assets within the ecosystem. The AI assets created then attract consumers, who contribute to an increase in revenue from AI assets. This increase in revenue captures the interest of investors, leading to shares appreciation. This appreciation triggers a rise in developer's net profit, which in turn, draws more developers into the ecosystem. This continuous influx of developers leads to more innovation, further driving the creation of valuable AI assets and perpetuating the growth cycle of the AI ecosystem.

Web3 Ecosystem The Web3 ecosystem growth flywheel begins with developers who drive product innovation by building applications on the Sahara Blockchain. This innovation accelerates user engagement, attracting more users to the platform. Increased user engagement helps to promote a flourishing ecosystem, leading to an increase in the number of transactions. As transactions grow, the revenue of basic services improves. This enhancement in revenue attracts more operators and validators to the ecosystem, which further enhances infrastructure stability. A stable infrastructure, in turn, attracts even more developers, fueling further product innovation and continuing the cycle of growth.

Synergy The AI and Web3 ecosystems are closely interconnected, with developers at the core of both. In the AI ecosystem, developers engage in AI relevant tasks to create AI assets that drive innovation and attract investment, fueling further growth of developers. More developers with these AI advancements are then leveraged within the Web3 ecosystem,

enhancing product innovation and user engagement. The increased engagement boosts transactions and revenue in the Web3 ecosystem, which in turn provides a stable platform for more AI developments, attracting more developers. This synergy ensures that as each ecosystem grows, it supports and accelerates the growth of the other, creating a powerful cycle of mutual reinforcement.

3.5.3 Capitalization of AI Assets

The capitalization of AI assets within the Sahara AI platform is structured into three distinct instruments: Receipts, Shares, and Licenses.

Receipt Receipt is an on-chain, non-transferable and non-fungible digital proof for AI assets *ownership*. Some specific parameters of the receipts, such as eligibility, allocation ratios, or the mechanism followed, will be finalized before developers and knowledge providers co-create AI assets. The receipts of high-quality AI assets will significantly benefit AI developers and relevant contributors in building their on-chain reputation. A higher on-chain reputation can lead to wider visibility within the Sahara AI ecosystem, increased opportunities to sell shares or licenses of AI assets, additional access to features in the Sahara AI ecosystem, and so forth. As such, AI developers are highly motivated to create high-quality AI assets.

Share Share is an on-chain digital proof that represents the *right to revenue sharing* of AI assets. The percentage of revenue sharing depends on the number of shares held. Some specific parameters of the shares, such as eligibility, allocation ratios, or the mechanism followed, will be finalized before developers and knowledge providers co-create AI assets. The revenues from shares are generally derived from the revenue generated from the sales of licenses for AI assets or commission fees from trading shares and licenses.

Owners can either retain their shares to earn part of the revenue generated by AI assets or sell their shares on the secondary market for immediate arbitrage. High-quality AI assets will ensure a continuous flow of revenue for the holders of shares, which will also lead to an increase in the price of shares in the secondary market. This will further motivate AI developers to create high-quality AI assets and incentivize consumers to buy and trade the different types of licenses.

License License is an on-chain digital proof that represents *permission to access or utilize* AI assets. Licenses have different rights to access or utilize the AI assets:

- Partnership License: Access permissions and payment methods are concluded in a customized agreement, typically involving a long-term profit-sharing model based on the user's benefits.
- API License: Access is granted through a secure authentication process and requires approval based on predefined usage criteria with fixed payment for each call.
- Full-access License: Full access to the AI assets with all internal parameters with one-time payment.
- Long-term License: Unlimited calls to AI assets within a specific time duration with one-time payment.

Some specific parameters of the license, such as eligibility, allocation ratios, or the mechanism followed, will be finalized before developers and knowledge providers co-create AI assets.

4 Governance

The governance of Sahara AI Platform emphasizes decentralized and community-driven innovation and decision-making. Actions and decisions are made transparently through the Sahara DAO, which empowers the users who made significant contributions to propose, discuss, and vote on key initiatives. The Sahara Foundation supports the initial setup of the Sahara DAO, guiding the community towards a fully decentralized governance. Our governance model ensures broad participation in the evolution of Sahara AI Platform.

4.1 The Sahara DAO

The Sahara DAO is dedicated to complete democratization, minimize governance to essential functions, promoting autonomy and innovation. Users who made significant contributions to the Sahara ecosystem play a crucial role in shaping the platform, proposing, discussing, and casting votes on various governance proposals. They have the option to either vote personally or assign their voting rights to a trusted representative.

The Sahara DAO is also designed to ensure the platform's independence and transparency, reflect the broader community's interests, prioritize fair compensation for contributions to the ecosystem and strategically manage resources to balance short-term needs with long-term goals.

4.2 The Sahara Foundation

The Sahara Foundation represents a vital element within the expansive, dynamic network of enterprises, collectives, and individuals dedicated to realizing the Sahara Vision. Its function is to facilitate the creation of the Sahara DAO, foster the growth of the Sahara ecosystem, and advance the underlying technology. During the formative phase of the Sahara DAO, the Foundation supports the Sahara DAO to identify effective strategies on long-term development and significantly enhance its prospects for a decentralized, transparent and community-driven governance.

The Foundation also supports the SBP as an open-source initiative, fostering community engagement to expand its technical infrastructure and sponsoring research on scalability, security, and decentralization.

5 The Future

In the future enabled by Sahara AI, artificial intelligence is no longer controlled by a select few but is instead a shared resource, accessible to all. This new world of AI is guided by principles of transparency, inclusivity, fairness, and, most importantly, user privacy and control.

In this future, AI enthusiasts can log into Sahara AI to explore a wide array of AI modules, fine-tune these models with their proprietary datasets, and share their improvements with the community. They have full control over their AI assets, ensuring that their contributions are protected while still earning rewards.

Data scientists, once constrained by centralized barriers, can now leverage Sahara AI's vast pool of data and models to advance their research. With the platform's robust privacy features, they can refine models and deploy solutions while maintaining control over their data and intellectual property.

For business leaders, Sahara AI provides a strategic advantage with tailored AI solutions that align with their company's specific needs. They can drive innovation within their organizations, secure in the knowledge that their data is protected and their AI assets remain under their control.

Meanwhile, individuals who may not be AI experts can also find a welcoming environment within Sahara AI. They can engage with cutting-edge projects, contribute their own expertise, participate in governance, and support ethical AI initiatives, all while maintaining full control over their contributions and earning rewards that reflect their involvement.

As the Sahara AI ecosystem evolves, we seek to build a community where everyone—regardless of background or expertise—has a meaningful role in shaping the future of AI. This is not just about building a platform; it's about creating a decentralized, open environment where the benefits of AI are shared equitably, and where privacy and control over AI assets are paramount. Together, we will transform AI into a force that empowers individuals and communities alike, guiding us toward a more connected, secure, and fair world.

References

[1] Ethan Buchman, Jae Kwon, and Zarko Milosevic. The latest gossip on bft consensus, 2019.

[2] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

[5] Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[7] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971, 2023.

[8] Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *ArXiv*, abs/2310.06825, 2023.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[11] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi (Jim) Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *ArXiv*, abs/2305.16291, 2023.

[12] Sirui Hong, Xiawu Zheng, Jonathan P. Chen, Yuheng Cheng, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zi Hen Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, and Chenglin Wu. Metagpt: Meta programming for multi-agent collaborative framework. *ArXiv*, abs/2308.00352, 2023.

[13] Noah Shinn, Beck Labash, and Ashwin Gopinath. Reflexion: an autonomous agent with dynamic memory and self-reflection. *ArXiv*, abs/2303.11366, 2023.

[14] Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. Dynamic llm-agent network: An llm-agent collaboration framework with agent team optimization. *ArXiv*, abs/2310.02170, 2023.

[15] Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role play with large language models. *Nature*, 623:493–498, 2023.

[16] Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. In *North American Chapter of the Association for Computational Linguistics*, 2023.

[17] Wenlong Huang, P. Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. *ArXiv*, abs/2201.07207, 2022.

[18] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

[19] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2023.

[20] Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. *ArXiv*, abs/2005.11401, 2020.

[21] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2024.

[22] Yongchao Chen, Jacob Arkin, Yang Zhang, Nicholas Roy, and Chuchu Fan. Scalable multi-robot collaboration with large language models: Centralized or decentralized systems? *arXiv preprint arXiv:2309.15943*, 2023.

[23] Rui Hao, Linmei Hu, Weijian Qi, Qingliu Wu, Yirui Zhang, and Liqiang Nie. Chatllm network: More brains, more intelligence. *ArXiv*, abs/2304.12998, 2023.

[24] Xudong Guo, Kaixuan Huang, Jiale Liu, Wenhui Fan, Natalia Vélez, Qingyun Wu, Huazheng Wang, Thomas L Griffiths, and Mengdi Wang. Embodied llm agents learn to cooperate in organized teams. *arXiv preprint arXiv:2403.12482*, 2024.

[25] Yashar Talebirad and Amirhossein Nadiri. Multi-agent collaboration: Harnessing the power of intelligent llm agents. *ArXiv*, abs/2306.03314, 2023.

[26] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. Encouraging divergent thinking in large language models through multi-agent debate. *ArXiv*, abs/2305.19118, 2023.

[27] Alexander Borzunov, Dmitry Baranchuk, Tim Dettmers, Max Ryabinin, Younes Belkada, Artem Chumachenko, Pavel Samygin, and Colin Raffel. Petals: Collaborative inference and fine-tuning of large models, 2023.

[28] Ziheng Jiang, Haibin Lin, Yinmin Zhong, Qi Huang, Yangrui Chen, Zhi Zhang, Yanghua Peng, Xiang Li, Cong Xie, Shibiao Nong, Yulu Jia, Sun He, Hongmin Chen, Zhihao Bai, Qi Hou, Shipeng Yan, Ding Zhou, Yiyao Sheng, Zhuo Jiang, Haohan Xu, Haoran Wei, Zhang Zhang, Pengfei Nie, Leqi Zou, Sida Zhao, Liang Xiang, Zherui Liu, Zhe Li, Xiaoying Jia, Jianxi Ye, Xin Jin, and Xin Liu. Megascale: Scaling large language model training to more than 10,000 gpus, 2024.

[29] Xisen Jin, Xiang Ren, Daniel Preotiuc-Pietro, and Pengxiang Cheng. Dataless knowledge fusion by merging weights of language models, 2023.

[30] Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time, 2022.

[31] Michael Matena and Colin Raffel. Merging models with fisher-weighted averaging, 2022.

[32] Zhuoran Zhao, Kamyar Mirzazad Barijough, and Andreas Gerstlauer. Deepthings: Distributed adaptive deep learning inference on resource-constrained iot edge clusters. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 37:2348–2359, 2018.

[33] Bingyang Wu, Yinmin Zhong, Zili Zhang, Gang Huang, Xuanzhe Liu, and Xin Jin. Fast distributed inference serving for large language models. *ArXiv*, abs/2305.05920, 2023.

[34] Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. Parameter-efficient fine-tuning for large models: A comprehensive survey, 2024.

[35] Fanxu Meng, Zhaohui Wang, and Muhan Zhang. Pissa: Principal singular values and singular vectors adaptation of large language models, 2024.

[36] Yixiao Li, Yifan Yu, Chen Liang, Pengcheng He, Nikos Karampatziakis, Weizhu Chen, and Tuo Zhao. Loftq: Lora-fine-tuning-aware quantization for large language models, 2023.

[37] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

[38] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS'16. ACM, October 2016.

[39] Björn Bebensee. Local differential privacy: a tutorial, 2019.

[40] Tong Wu, Ashwinee Panda, Jiachen T. Wang, and Prateek Mittal. Privacy-preserving in-context learning for large language models, 2023.

[41] Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A. Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. Differentially private fine-tuning of language models, 2022.

[42] Tianyu Chen, Hangbo Bao, Shaohan Huang, Li Dong, Binxing Jiao, Daxin Jiang, Haoyi Zhou, Jianxin Li, and Furu Wei. The-x: Privacy-preserving transformer inference with homomorphic encryption, 2022.

[43] Kanav Gupta, Neha Jawalkar, Ananta Mukherjee, Nishanth Chandran, Divya Gupta, Ashish Panwar, and Rahul Sharma. Sigma: secure gpt inference with function secret sharing. *Cryptology ePrint Archive*, 2023.

[44] Théo Ryffel, Pierre Tholoniat, David Pointcheval, and Francis R. Bach. Ariann: Low-interaction privacy-preserving deep learning via function secret sharing. *Proceedings on Privacy Enhancing Technologies*, 2022:291 – 316, 2020.

[45] KD Conway, Cathie So, Xiaohang Yu, and Kartin Wong. opml: Optimistic machine learning on blockchain, 2024.

[46] Haochen Sun, Jason Li, and Hongyang Zhang. zkllm: Zero knowledge proofs for large language models, 2024.

A Glossaries

- **AI Models**: AI models are algorithms designed to simulate human intelligence by learning from data. These models can recognize patterns, make decisions, and even generate content based on what they've learned. AI models range from narrow AI, which specializes in specific tasks, to more general models that handle a broader range of functions. Examples include neural networks, decision trees, and regression models.

- **Generative Models**: These AI models are designed to create new content, such as text, images, music, or video. Generative models work by learning patterns from large datasets and using this knowledge to generate new data that resembles the original. Common types of generative models include transformers.

- **Transformers**: A type of deep learning model architecture that excels in processing sequential data, like language. Transformers use self-attention mechanisms to weigh the importance of different parts of the input data simultaneously, making them effective for tasks like translation, text generation, and more. Transformers are the foundation for many modern large language models.

- **Large Language Models (LLMs)**: These are a subset of transformers trained on massive amounts of text data. LLMs are capable of generating human-like text by predicting the next word in a sequence based on the context provided. They are versatile and can perform various tasks such as writing, translation, and summarization.

- **AI Agent**: An AI agent is an autonomous program that interacts with its environment to perform tasks, make decisions, and achieve specific goals. AI agents can be simple, like a virtual assistant that schedules appointments, or complex, like a self-driving car that navigates through traffic.

- **Parameter Efficient Fine-Tuning (PEFT)**: PEFT refers to techniques that enable fine-tuning large pre-trained models with fewer parameters, making the process more resource-efficient. PEFT methods like LoRA or adapters focus on updating only a small part of the model, reducing computational cost while maintaining performance.

- **LoRA (Low-Rank Adaptation)**: LoRA is a PEFT method that adds small, trainable matrices to the layers of a large language model. Instead of updating the entire model during fine-tuning, LoRA modifies these low-rank matrices, which represent a smaller subset of the model's parameters. This approach significantly reduces the number of parameters that need adjustment, speeding up the process and reducing resource usage while maintaining or enhancing performance.

- **Retrieval Augmented Generation (RAG)**: RAG is a hybrid approach that combines retrieval-based methods with generative models. In RAG, the model first retrieves relevant information from a database or document set and then uses that information to generate a response. This method allows the model to provide more accurate and contextually relevant answers, especially in cases where up-to-date or specialized information is required.

- **Differential Privacy**: Differential privacy is a technique used to protect individual privacy in datasets by adding random noise to the data. This ensures that the data can be used for analysis without revealing sensitive information about individuals, making it useful in contexts like sharing private records or training AI models with personal data.

- **Model Merging**: The process of combining two or more AI models to create a new model that benefits from the strengths of each original model. Model merging can be used to improve performance, incorporate diverse knowledge, or combine different capabilities into a single, more powerful model.