Feasibility Report and Implementation Plan: An Architecture for Distributed Artificial Intelligence

Section 1: Executive Summary

Thesis Statement

The concept of a distributed Artificial Intelligence (AI), which leverages users' idle computing power, is not only viable but represents the next logical evolution of cloud computing. It merges the volunteer computing model with the economic incentives of Decentralized Physical Infrastructure Networks (DePIN). This report presents a comprehensive architecture for building such a network, focused on the **inference of Mixture of Experts (MoE) models**, an approach that solves the key challenges of running large-scale AI models on heterogeneous consumer hardware.

Proposal Synthesis

We propose a hybrid system that combines the robust client-server architecture of projects like BOINC for task orchestration and validation with a peer-to-peer (P2P) network based on libp2p for low-latency communication between nodes.(1), (2), (3) The core of the AI will be a large language model (LLM) structured as a Mixture of Experts (MoE), where a central server (or a core of DAO validators) acts as the "gating network," and user nodes run the individual "experts," which are smaller, more manageable neural sub-networks.(4), (5) This design transforms the problem of running a massive model on a single device into the much more feasible task of running smaller model fragments on many devices and aggregating the results.

Economic and Governance Model

Participation will be incentivized through a dual-purpose token (utility and governance), creating a sustainable economic flywheel where the use of the network by AI consumers funds the rewards for compute providers.(6) Users will pay for inference services, and a portion of this revenue will be used to reward compute providers and to burn tokens, creating a deflationary pressure that links the token's value to the network's actual usage.(7), (8) Governance will be progressively transferred to a Decentralized Autonomous Organization (DAO), ensuring long-term alignment with the community's interests and the protocol's resilience.(9), (10)

Challenges and Mitigations

The main challenges will be addressed: achieving low latency (unlike traditional high-throughput volunteer computing systems), ensuring privacy and security in a trustless environment, and navigating the complex legal and regulatory landscape. Low latency will be achieved through intelligent node selection and direct P2P communication.(11), (12) Privacy will be addressed through Federated Learning for model updates and by offering premium Secure Multi-Party Computation (SMPC) services for sensitive inferences.(13), (14) Legal risk, particularly the personal liability of members, will be mitigated by establishing a formal legal entity, such as a DAO LLC, in a favorable jurisdiction.(15), (16)

Strategic Conclusion

This model has the potential to radically democratize access to AI computation, offering a censorship-resistant and significantly lower-cost alternative to centralized cloud providers. By aligning economic incentives with infrastructure provision, a global, community-owned supercomputer can be built, capable of driving the next wave of innovation in artificial intelligence.

Section 2: Conceptual Framework: An Analysis of Distributed Artificial Intelligence

2.1. Validation of the Core Idea: From Cloud Monopoly to Community Supercomputer

Current artificial intelligence infrastructure is predominantly centralized, dominated by an oligopoly of cloud service providers like Amazon Web Services, Google Cloud, and Microsoft Azure.(17), (18) This centralization creates significant cost and access barriers while concentrating control over a transformative technology. The growing demand for compute power for training and inferencing AI models, especially LLMs, is beginning to outstrip the supply of centralized data centers, resulting in GPU shortages and high prices.(18), (19)

This market tension creates a strategic opportunity for alternative models. The idea of a distributed AI is based on a fundamental observation: there is a vast reservoir of latent computational power in the billions of consumer devices (PCs, game consoles, and even smartphones) worldwide.(20), (21) These devices often remain idle for much of the day. Projects like Folding@home have shown that the combined processing power of these volunteer machines can far exceed that of the world's most powerful supercomputers.(20), (22) Distributed AI seeks to transform this global liability (CPU/GPU downtime) into a productive asset, creating a community-owned supercomputer for artificial intelligence.

2.2. Historical Precedents and Lessons Learned: The Legacy of Volunteer Computing

The idea of harnessing distributed computing is not new. For over two decades, volunteer computing projects have laid the technical and social groundwork for this model.

Analysis of BOINC, SETI@home, and Folding@home

The **Berkeley Open Infrastructure for Network Computing (BOINC)** is an open-source software platform that has powered some of the largest distributed computing projects, including SETI@home (Search for Extraterrestrial Intelligence) and Folding@home (protein folding simulation for disease research).(1), (20), (21) BOINC's architecture, a client-server model, is a direct and fundamental precedent for the proposed AI network.(1), (2)

In this model, a central server distributes "work units"—chunks of data and the code to process them—to client software running on volunteers' computers, known as "worker nodes."(2), (23) One of BOINC's key strengths is its ability to manage a massive network of nodes that are:

- **Heterogeneous:** With different processor types (CPU, GPU), architectures (x86, ARM), and operating systems (Windows, Mac, Linux, Android).(1), (24)
- **Sporadically available:** Nodes can connect and disconnect from the network at any time.(1)
- **Untrusted:** Nodes may return incorrect results, either due to hardware failures or malicious behavior.(1)

Validation Mechanisms

To address the problem of untrusted nodes, BOINC implements a robust **redundancy validation** mechanism. Each work unit is sent to multiple independent clients. When the results are returned, the validator daemon on the server compares them. A "quorum" of matching results (usually two or three) is required for a result to be considered valid. If the results do not match or are not returned by the deadline, the server generates additional instances of the task and sends them to other nodes until consensus is reached.(2), (23) This method, while not suitable for low latency, establishes a crucial principle: trust in a distributed system is achieved through redundancy and verification, not by assuming the honesty of individual nodes.

Non-Economic Incentive Models

Perhaps the most important lesson from these pioneering projects is their success in mobilizing millions of volunteers without offering financial compensation. Their incentive model was based on intrinsic and social factors:

- **Contribution to a greater good:** The "small but captivating possibility that your computer will detect the faint murmur of a civilization beyond Earth" was a powerful motivator for SETI@home participants.(25) Similarly, Folding@home volunteers contribute directly to research on diseases like Alzheimer's, cancer, and, more recently, COVID-19.(21), (26)

- **Gamification and Competition:** The platforms incorporated "credit" systems that quantified each user's computational contribution. Leaderboards were published for individuals and teams, fostering healthy competition and long-term engagement.(21), (25)

These projects demonstrate that a compelling mission and a well-structured community can be as powerful incentives as money, a vital lesson for the bootstrapping phase of any decentralized project.

2.3. The Modern Paradigm: Contextualization in the DePIN Ecosystem

While volunteer computing relied on altruism, the modern paradigm of **Decentralized Physical Infrastructure Networks (DePIN)** is based on explicit economic incentives. DePIN refers to blockchain networks that use tokens to incentivize people to build and operate physical infrastructure in the real world, from wireless and storage networks to compute networks.(6), (27)

The DePIN model operates on a virtuous cycle known as the "economic flywheel." Initially, the protocol issues tokens to reward supply-side providers (e.g., people who install a wireless hotspot or connect their GPU to the network). This incentive bootstraps the network's growth. As the network grows in capacity and coverage, it becomes more useful and attracts demand-side users who pay for the service. The revenue generated from demand is used to buy back and/or burn tokens, or to fund provider rewards, creating a sustainable economic model that becomes less reliant on inflationary emissions.(6)

The proposed distributed AI fits perfectly into the **DePIN compute** sector, a space that already has established projects and a significant market capitalization.(28), (29) An analysis of key players reveals different approaches:

- **Akash Network (AKT):** Functions as a decentralized, open-source "cloud supermarket," a P2P marketplace where providers can offer their compute capacity (CPU and GPU) and users can lease it. It uses a "reverse auction" system to reduce costs.(30), (31), (32)
- **Render Network (RNDR):** Specializes in rendering 3D graphics and visual effects, connecting artists and studios that need GPU power with a network of node operators who offer their idle hardware in exchange for RNDR tokens.(32), (33), (34)
- **Bittensor (TAO):** Goes a step beyond simple compute. It is a decentralized marketplace for intelligence itself. "Miners" not only contribute compute but also train and serve AI models in various "subnets." "Validators" evaluate the quality of these models' responses, and the system rewards those that produce the highest informational value, creating a

competition that drives AI innovation.(35), (36), (37)

The current proposal differs by focusing specifically on **large-scale, low-latency MoE model inference**, a niche that leverages the strengths of distributed computing in a way that more generalist market models do not directly address.

2.4. Key Value Propositions

The proposed architecture offers a set of disruptive advantages over the current centralized model:

- **Radical Cost Reduction:** By eliminating intermediaries and using existing, already-paid-for hardware resources, AI inference costs can be drastically reduced. Projects like Akash claim to achieve savings of up to 85% compared to traditional cloud providers, a figure that serves as an achievable benchmark.(31), (38)
- **Massive Scalability:** The network can grow organically as more users join. It is not limited by the data center construction capacity of a single company. As demonstrated by BOINC and Folding@home, a volunteer network can aggregate compute capacity that rivals and surpasses the world's largest supercomputers.(20), (22), (25)
- **Censorship Resistance and Data Sovereignty:** A decentralized network, governed by its global community, is inherently resistant to censorship by a single corporate or governmental entity.(30), (31) It allows for the execution of open-source AI models without the restrictions or filters imposed by centralized providers, which is crucial for unrestricted research and freedom of expression.(39), (40)
- **Democratized Access:** By drastically reducing costs, the platform removes the economic barriers that currently prevent many developers, researchers, and startups from experimenting and building with state-of-the-art AI models. This fosters a more equitable and vibrant innovation ecosystem.(41), (42)

The convergence of these precedents and paradigms reveals a fundamental truth: the evolution of distributed computing has shifted from a model driven by altruism and gamification to one sustained by direct economic incentives. Pioneering projects like SETI@home demonstrated the technical feasibility of aggregating volunteer resources on a massive scale, driven by the intrinsic motivation to contribute to science.(23), (25) However, this volunteer model often results in a fluctuating compute supply, unsuitable for commercial applications requiring reliability and low latency. The modern wave of DePIN has introduced an explicit economic model through tokens, transforming contribution from a voluntary act to an economic activity.(6), (27) This attracts a broader set of participants and stabilizes the infrastructure supply. A successful project in the current landscape must, therefore, merge both approaches. It must start with a strong, open-source mission to attract an initial community of ideological enthusiasts, and then cement its long-term growth with robust tokenomics that fairly and sustainably reward infrastructure providers.

Section 3: Architectural Plan for a Distributed AI Network

To realize the vision of a distributed AI, a robust architecture that balances efficiency, scalability, and security is required. The proposed design is a hybrid model that combines a centralized orchestration layer, inspired by BOINC, with a decentralized P2P communication layer to achieve low latency.

3.1. Core Infrastructure: A Hybrid Model

Orchestration Layer (BOINC-based)

The core of the network management will reside on a central server (or, in a later phase, a set of validator nodes governed by the DAO). This server will not perform the heavy computation but will act as the orchestra conductor, coordinating the thousands of worker nodes. Its main components, analogous to BOINC daemons, are (2):

- **Scheduler:** This is the main point of contact for clients. It receives work requests and assigns AI work units. Unlike the standard BOINC scheduler, which uses an exponential backoff to minimize server load, this component will be modified to prioritize low latency. It will implement fast reconnection directives, such as the <next_rpc_delay> tag, which instructs the client to reconnect within a short period (e.g., seconds instead of hours), ensuring a constant flow of tasks for real-time applications.(12)
- **Feeder:** A backend process that preloads tasks from the database into a shared memory block. This allows the scheduler to access and assign tasks with minimal latency, without having to query the database for each request.(2)
- **Validator:** This component is crucial for network trust. It receives the results of completed tasks from multiple nodes and compares them. If a quorum of identical results is reached, the task is marked as valid, and the nodes that provided the correct answer see their reputation increase. Mismatched results are discarded.(2)
- **Transitioner:** Manages the lifecycle of tasks. If a task is not returned by its deadline or if

its results are invalidated, the transitioner puts it back in the queue to be assigned to other nodes.(2)

Client Software (Worker Node): A Lightweight, Cross-Platform Agent

The software that users will install on their devices must be lightweight, secure, and non-intrusive.

- **Client Core:** A background process (similar to boinc.exe or boincd) that communicates with the orchestration server. It manages the download of tasks, the execution of AI engines, and the upload of results. It will be configured to use CPU and GPU resources only when the device is idle or at a very low priority, so as not to interfere with the user's normal use of the computer.(1), (23), (43)
- **AI Execution Engine:** The client will not contain a monolithic AI model. Instead, it will integrate standardized and highly optimized inference engines, capable of running the model fragments (the "experts") that the server sends it. The two key technologies for this are:
    - **ONNX Runtime:** A high-performance, cross-platform inference engine from Microsoft. Its main advantage is the use of "Execution Providers" (EPs), which are backends that optimize execution on specific hardware. This allows the same ONNX model to run optimally on NVIDIA GPUs (via the CUDA EP), on Apple hardware (via CoreML), on Intel CPUs (via OpenVINO), and on a wide range of other devices.(44), (45), (46)
    - **llama.cpp:** An open-source project that has revolutionized LLM inference on consumer hardware. Written in C/C++, it offers exceptional performance on a wide variety of CPUs (x86, ARM) and GPUs (via backends like Metal, CUDA, Vulkan, OpenCL). Its native file format, GGUF, is designed to efficiently load and run quantized models (models whose weights have been reduced in precision to decrease memory usage and speed up computation), which is ideal for this use case.(47), (48)

Communication Layer (Web3 P2P-based)

For operations requiring low latency and direct communication between nodes, a P2P network will be overlaid on the client-server architecture.

- **libp2p Protocol:** This is the gold standard for P2P communication in the Web3 ecosystem, used by projects like IPFS and Ethereum. It is a modular network stack that provides essential functionalities such as peer discovery (finding other nodes in the network), message routing, and, crucially, **NAT traversal** techniques (like Hole Punching and AutoRelay). This allows two nodes to connect directly even if both are behind firewalls or home routers, an indispensable requirement for a truly decentralized P2P network.(3), (49)

- **Geographic and Low-Latency Routing:** To minimize data travel time, the system will implement location-based routing strategies. When a quick response is needed, instead of selecting a random node, the scheduler or the nodes themselves will prioritize peers that are geographically closer.(11), (50), (51) The network can also maintain a set of high-bandwidth "relay nodes" at strategic points on the global internet to speed up the propagation of critical messages, a proven technique in high-performance blockchain networks.(52)

3.2. Task Management and Workflow

- **Definition of "AI Work Units":** A "work unit" will not be a complete inference task, but an atomic micro-task. In the context of an MoE model, a work unit could consist of: the input (a batch of *tokens*), the identifier of the "expert" to be executed, and the necessary metadata. This breaks down a large task into thousands of small, parallel tasks.
- **Handling Heterogeneity and Availability:** The server will maintain a detailed profile of each worker node, including its hardware specifications (CPU, GPU, VRAM), performance benchmarks, network speed, and a reputation score. Tasks will be assigned to nodes that meet the minimum requirements and are best suited for the task. To manage the intermittent nature of volunteer nodes, each work unit will have a strict deadline (e.g., a few minutes). If a result is not returned on time, the Transitioner will immediately reassign it to another available node.(21)
- **Reputation and Trust System:** Trust is not assumed; it is built. A dynamic reputation system will be implemented to rate each node based on objective and verifiable metrics:
  - **Correctness:** The history of submitting results that are positively validated by the quorum system.
  - **Timeliness:** The frequency with which tasks are completed before their deadline.(21), (43)
  - **Performance:** The measured computation speed, which allows the scheduler to accurately estimate completion times.
  - Uptime: The time the node remains connected and active on the network.
    The scheduler will use this reputation score as a key factor in task assignment, prioritizing the most reliable and highest-performing nodes.(53), (54)

The choice of a hybrid architecture is not a compromise but a strategic optimization. A purely P2P system for assigning AI tasks on a global scale faces almost insurmountable coordination challenges. Issues like optimal load balancing, prevention of malicious collusion, and efficient node selection are extremely difficult to solve without a global view of the network's state, something a centralized authority or a set of trusted validators can provide efficiently.(55) On the other hand, a purely centralized system like the classic BOINC model, while scalable, is inherently limited by latency, as every interaction must pass through the server.(12) By combining both models, a powerful synergy is achieved: the central server handles "high-trust" tasks that require a global view, such as result validation, token economy management, and initial query routing. Once the server has assigned a sub-task to a group of nodes, they can use the libp2p P2P network to communicate directly with each other, coordinating execution and aggregating intermediate results without the server bottleneck. This hybrid architecture aligns perfectly with the Mixture of Experts AI model: the server acts as the global gating network, and the P2P nodes function as the experts that perform localized computational work efficiently and with low latency.

Section 4: AI Design: Models and Methodologies for a Decentralized World

The choice of the AI model architecture and operational methodology is as crucial as the design of the network infrastructure. The system must be designed from the ground up to thrive in a distributed, heterogeneous, and low-trust computing environment.

4.1. The Strategic Focus: Prioritizing Inference over Training

The first and most important design principle is to focus on AI **inference**, not training. Distributed training of large language models (LLMs) over the internet on consumer hardware is, with current technology, practically unfeasible for several fundamental reasons:

- **Bandwidth and Latency Requirements:** Distributed training, especially for transformer models, requires constant, very high-speed communication between the nodes processing different parts of the model or data. Techniques like data and tensor parallelism depend on ultra-low latency and massive bandwidth network interconnections (hundreds of GB/s), such as NVIDIA's NVLink or specialized RoCE networks in data centers.(19), (56), (57) Home internet connections, with their high latency and asymmetric bandwidth, are completely inadequate for this task.
- **State Dependency and Synchronization:** The training process is sequential and state-dependent. Each optimization step (e.g., gradient descent) depends on the result

of the previous step. The sporadic nature of volunteer nodes, which can disconnect at any time, would cause the training process to constantly stop and restart, making it inefficient to the point of being useless.

In contrast, **inference** (the process of using an already trained model to make predictions) is an ideal use case for distributed computing:

- **Highly Parallelizable:** Each inference request is independent of the others. Thousands of requests can be processed in parallel on thousands of different nodes without needing communication between them. This is known as an "embarrassingly parallel" problem.(58), (59)
- **Stateless:** Inference does not modify the model's weights. If a node fails mid-task, it can simply be reassigned to another node without loss of progress.
- **Lower Communication Requirements:** Communication is limited to sending the input (the *prompt*) and receiving the output (the generated response).
- **Feasibility on Consumer Hardware:** Modern consumer GPUs, while not suitable for large-scale training, are extremely capable of performing inference efficiently, especially with quantized models.(17)

4.2. Mixture of Experts (MoE) Architecture: The Key to Distributing Giant Models

The main obstacle to LLM inference on consumer devices is the model size, which often exceeds the available VRAM. The **Mixture of Experts (MoE)** architecture is the solution to this problem and is the central pillar of the proposed AI strategy.(4), (5), (60)

An MoE model is not a dense, monolithic neural network. Instead, it replaces some of its layers (typically the *feed-forward* layers) with an MoE block. This block contains two components (5):

1. A set of **"experts"**: These are smaller neural networks, identical in architecture.
2. A **"gating network"**: A small neural network that learns to direct each token of the input sequence to the expert (or a small number of experts, typically two) that is best qualified to process it.

This means that for each token, only a small fraction of the model's total parameters is activated and used for computation. For example, the Mixtral 8x7B model has a total of 46.7 billion parameters, but for each token, only two of its eight experts are activated, using approximately 12.9 billion parameters per inference.(4) This approach allows the model to have enormous capacity and knowledge (stored in its total parameters) while maintaining a relatively low computational cost per inference.

This architecture maps perfectly to our distributed network:

1. The **central server** will host the lighter, critical parts of the model: the gating network and the shared layers (such as the attention layers).
2. The **user nodes** will download and host one or more of the "experts." Since each expert is a much smaller model, it can easily fit into the VRAM of a consumer GPU.
3. The inference workflow would be as follows: a user sends a query to the server. The server processes the input through the gating network, which determines the sequence of experts needed to generate the response. The server then breaks down the task and sends micro-jobs to the network nodes hosting the required experts. The nodes execute their experts and return the results, which the server assembles to form the final response.

This approach is the fundamental enabler that makes running a state-of-the-art AI model on a network of consumer hardware feasible. It transforms an intractable problem ("run a 100B parameter model on a PC") into a manageable logistical problem ("run a 7B parameter model on 16 different PCs and coordinate the results").

4.3. Federated Learning (FL): Privacy-Preserving Updates and Customization

For the AI model to evolve, improve, and adapt over time, it needs to be retrained with new data. Collecting user data on a central server would pose enormous privacy problems. **Federated Learning (FL)** is the solution to this dilemma.(13), (61)

In an FL system, the global model is sent to the users' devices. The model is trained locally on each device, using the user's local data, which never leaves the device. After local training, only the model weight updates (the gradients), not the data itself, are sent back to the server. The server securely aggregates these updates (often using additional cryptographic techniques to ensure anonymity) to create a new, improved version of the global model.(13), (62)

In our system, FL will be used to:

- **Improve the Global Model:** Periodically, FL rounds can be conducted to retrain both the gating network and the experts, improving the overall system performance.
- **Create Specialized Experts:** FL is particularly effective for handling heterogeneous and non-independently and identically distributed (non-IID) data, which is exactly the situation in a network of diverse user devices.(61), (63) This could allow for the creation of

specialized experts in certain domains or data types that emerge organically from the user community.

4.4. Optimization for Low Latency: Adapting the High-Throughput Model

The primary goal of AI inference is a fast response. Therefore, the architecture must be optimized for low latency, a goal often in conflict with the high-throughput optimization of traditional volunteer computing systems.

- **Intelligent Node Selection:** The server's scheduler must be sophisticated. When receiving an inference request, it must consider not only the computational capacity of the nodes but also their network latency relative to the server and, potentially, the end-user. Priority will be given to high-reputation nodes that are geographically closer to minimize data round-trip time.(11), (50)
- **Speculative Execution:** For the most critical tasks within an inference chain, the scheduler can send the same micro-task to several high-quality nodes simultaneously. The system will use the first correct result that arrives, canceling the duplicate tasks. This increases computational redundancy but reduces latency by mitigating the risk of relying on a single node that might be slow.
- **Aggressive Management of "Stragglers":** In a low-latency system, slow nodes are a significant problem. Very short timeouts will be set for each micro-task. If a node does not respond within this threshold (e.g., a few hundred milliseconds), the task will be instantly reassigned to another waiting node. The reputation system will heavily penalize nodes that are consistently slow or fail to complete tasks, ensuring that the active node pool is of high quality.

Section 5: The Economic Engine: A Sustainable Tokenomics Model

To transition from a volunteer computing model based on altruism to a commercial-grade infrastructure ecosystem, a robust economic engine is indispensable. This engine will be based on a native token designed to incentivize participation, facilitate transactions, and align the interests of all network participants in the long term.

5.1. Incentivizing Participation at Scale

The fundamental principle is that computational resources are not free. Users incur costs for electricity and hardware wear and tear. To attract and retain a large-scale, reliable, and high-performance compute supply, the network must compensate providers fairly and predictably.(64), (65), (66) A crypto-economic reward system, paid in the network's native token, transforms resource contribution from a voluntary activity into an economic opportunity, attracting a broader spectrum of participants, from individual enthusiasts to professional hardware operators.

5.2. Token Design: A Dual-Purpose Asset (Utility and Governance)

We propose the creation of a native token, provisionally named NEURO, which will serve two critical functions within the ecosystem.

Utility Functions

The token's utility is what generates organic and sustainable demand, beyond mere speculation.

- **Medium of Payment:** Developers, companies, and end-users consuming the AI inference services will pay for them using NEURO. To simplify the user experience, the platform may allow payments in stablecoins (like USDC) or even fiat currency, which would be automatically converted to NEURO on the backend to settle transactions on the network. This ensures that all economic activity flows through the native token.(34), (65), (67)
- **Staking for Node Operators:** To participate as a compute provider, a user will need to *stake*, i.e., lock a minimum amount of NEURO as a form of collateral or bond. This mechanism serves several functions: (1) it acts as a deterrent against malicious behavior (a node attempting to cheat the system may have its *stake* "slashed"); (2) it aligns the incentives of providers with the long-term health of the network, as they have a direct economic interest in its success; and (3) it reduces the circulating supply of the token, which can have a positive impact on its value.(35), (36)
- **Transaction Fees:** Small fees in NEURO will be applied to network operations to prevent spam attacks and fund network security.

Governance Functions

As the network matures, its control will be progressively transferred to its user community through a DAO.

- **Voting Rights:** Holding NEURO tokens, especially those that are staked, will grant users the right to propose and vote on Protocol Improvement Proposals (PIPs).(36), (65), (67)
- **Scope of Governance:** Key decisions to be voted on by the DAO will include: the integration of new AI models into the network, adjustments to the fee and reward structure, the allocation of funds from the DAO treasury for development grants, marketing, and other ecosystem initiatives, and the election of members to technical committees.

5.3. Value Accrual and Sustainability: The Economic Flywheel

A successful tokenomics design must create a closed-loop system where value accrues to the token as network usage increases.

- **Revenue Stream and DAO Treasury:** A percentage of all payments for inference services (e.g., 5-10%) will be automatically directed to a treasury controlled by the DAO. These funds will be used to finance the ongoing development of the protocol, developer grants, marketing programs, and other initiatives that benefit the ecosystem.(38)
- **Burn Mechanism:** To directly link the token's value to network activity, a portion of the collected fees (e.g., 20% of the treasury's revenue) will be programmatically "burned," i.e., permanently removed from circulation. This deflationary mechanism means that as the network processes more inferences and generates more revenue, the total supply of NEURO decreases, increasing scarcity and potentially the value of the remaining tokens.(8)
- **Sustainable Reward Distribution:** Rewards for compute providers will be funded by a dual model. In the initial phase, inflationary token emissions (a predefined amount of new NEURO created per block or epoch) will be used to bootstrap the network and attract early providers. However, the long-term goal is to transition to a model where the majority of rewards are funded by the actual revenue generated by the network. This avoids the unsustainable model of many projects that rely solely on high inflation to pay rewards, which often leads to constant selling pressure and token devaluation.(7), (8)
- **Prevention of "Death Spirals":** The economic model must be designed to be robust against extreme market conditions. Reflexive feedback mechanisms that led to the

collapse of projects like Terra/LUNA, where the value of one asset depended algorithmically on another, creating systemic risk, will be avoided.(68), (69) The value of NEURO will not be based on an algorithmic peg but on the real demand for a tangible service: AI computation. Additionally, staking rewards will be kept at sustainable levels and backed by revenue, to avoid repeating the mistakes of platforms like Celsius, which offered unsustainable yields.(68)

5.4. DePIN Tokenomics Comparative Table

To contextualize the proposed model, it is useful to compare it with the approaches of other leading projects in the DePIN compute sector.

Feature	Akash Network (AKT)	Render Network (RNDR)	Bittensor (TAO)	Proposed Distributed AI (NEURO)
Primary Token Function	Utility (payment, PoS security staking), Governance (30), (67)	Utility (payment for rendering), Governance (34)	Utility (access to subnets, staking), Incentive (reward for intelligence) (35), (36)	Dual-Purpose: Utility (payment, collateral staking) and Governance (DAO)
Value Accrual Mechanism	Network fees, staking (participation in inflation) (38)	Burn-and-Mint Equilibrium Mechanism (37)	Fixed issuance with halvings (similar to Bitcoin), rewards based on informational value (70)	Network usage fees, burning a portion of fees, staking
Provider Incentive	Block rewards (inflation) and lease fees paid by users (65), (67)	RNDR tokens paid by creators for each completed	TAO emissions distributed to miners and validators based on their	Block rewards (initial phase) and a portion of inference revenue paid

		rendering job (33), (34)	performance and consensus (35), (71)	by users
Governance	On-chain via voting by AKT stakers (67)	On-chain governance via voting by RNDR holders (34)	Governance via a "Triumvirate" and voting by TAO stakers (70)	Progressively decentralized DAO, controlled by NEURO stakers, using a modular framework like Aragon

The greatest risk for a DePIN project is often not technical failure but economic collapse due to flawed tokenomics design. A successful model must go beyond speculation and create a self-reinforcing value cycle. The token's utility must be indispensable to the network's operation. In this case, every AI query generates a tangible fee. By channeling a portion of this fee into a burn or buyback mechanism, a direct link is created between platform usage and token scarcity. This mechanism generates a positive feedback loop: increased AI usage leads to more fees, which increases token scarcity and value. A more valuable token translates into more attractive incentives for compute providers, which in turn improves the quality and quantity of the infrastructure supply. Better infrastructure attracts more users, completing the cycle. This design aligns the incentives of all participants—developers, providers, users, and investors—towards the common goal of maximizing the real and useful application of the network.

Section 6: Ensuring Trust and Privacy in a Trustless Environment

In a system where computations are performed on the computers of thousands of strangers, trust cannot be assumed; it must be cryptographically and economically designed into the protocol itself. The privacy of user data and the integrity of model results are paramount.

6.1. Privacy of User and Model Data

- **Privacy by Design with Federated Learning (FL):** As mentioned earlier, FL is the cornerstone of the privacy strategy for model training and improvement. By ensuring that raw training data never leaves the users' devices, the principle of data minimization is met, and privacy risks are significantly mitigated.(13), (72), (73)
- **Secure Multi-Party Computation (SMPC) for Private Inference:** For the use case of inference, where a user sends a query and receives a response, FL is not applicable. Here, the privacy of the query itself must be considered. **Secure Multi-Party Computation (SMPC)** is a set of cryptographic techniques that allow multiple parties (in this case, the network nodes) to jointly compute a function on their private inputs (the model weights and the user's query) without revealing those inputs to each other.(14), (74), (75) However, this security comes at a cost. SMPC protocols require multiple rounds of communication and significant computational overhead, which drastically increases the latency and cost of inference.(76), (77), (78)
- **Proposal of a Tiered Service Model:** Given the tension between performance and privacy, the most pragmatic solution is to offer different service levels.
  1. **Standard Tier:** Fast and low-cost inference. Privacy is based on the anonymization of requests and the trust that individual nodes only see a small fraction of the overall task (a single expert processing one token), making it difficult to reconstruct the full query.
  2. **High-Security Tier (Premium):** Inference using SMPC. This service would be considerably slower and more expensive but would offer end-to-end cryptographic privacy guarantees. It would be aimed at high-sensitivity use cases, such as the analysis of medical, financial, or legal data, where clients are willing to pay a premium for confidentiality.(79), (80), (81)
- **Protection of Model Intellectual Property:** To prevent nodes from stealing the weights of the "experts" they run, several strategies can be employed. The models can be distributed in obfuscated formats. Additionally, execution could take place within Trusted Execution Environments (TEEs) like Intel SGX or AMD SEV, if the node's hardware supports it, which isolates the computation from the host operating system.(72), (82) Frequent rotation of experts and verification of the client software's integrity can also mitigate this risk.

6.2. Reputation and Node Selection System

The integrity of the network depends on the ability to identify and reward honest and efficient actors while penalizing malicious or underperforming ones.

- **Leader Election for Critical Tasks:** For certain tasks that require closer coordination, such as aggregating results from multiple experts for a single response, the scheduler can use a **leader election** algorithm. A node with an exceptionally high reputation would be temporarily designated as a "leader" to coordinate a small group of worker nodes, ensuring the task is completed efficiently and correctly.(55), (83)

- **Multi-Factor Reputation Metrics:** A node's reputation score will not be a single metric but a composite score based on:
  - **Honesty and Accuracy:** A node's history of providing results that match the network consensus.
  - **Timeliness and Performance:** The speed at which a node completes assigned tasks compared to the expected benchmarks for its hardware.
  - **Network Contribution:** Factors such as uptime and the amount of NEURO staked.
  - This reputation system, based on verifiable historical behavior, allows the scheduler to make informed decisions about who to assign work to.(53), (54)
- **Slashing Mechanism:** The staking of NEURO is not just an entry barrier but also an economic security mechanism. If the validation system detects that a node is consistently submitting incorrect results or attempting to manipulate the system, a portion of its NEURO stake will be "slashed" and transferred to the DAO treasury or burned. This risk of direct financial loss is the strongest deterrent against malicious behavior.

There is an inherent and fundamental tension between decentralization, performance (low latency), and strong cryptographic privacy. It is impossible to optimize all three simultaneously. Real-time inference demands minimal latency (84), (85), while robust privacy techniques like SMPC introduce significant latency due to their multiple communication rounds and computational complexity.(76), (78), (86) A "one-size-fits-all" approach is destined to fail: a system that imposes SMPC on all transactions would be too slow and expensive for general use, while a system that completely ignores inference privacy would be unacceptable for sensitive applications in sectors like healthcare or finance.(80), (81) The most viable strategy, therefore, is to offer a tiered service model. This allows users to select the appropriate balance point in this trilemma based on their specific needs, segmenting the market and maximizing the network's utility and reach.

Section 7: Governance, Legal, and Regulatory Strategy

Technological innovation alone is insufficient for long-term success. A solid governance structure and a proactive legal strategy are indispensable for navigating the complex and often uncertain regulatory landscape of decentralized technologies.

7.1. Decentralized Governance: The DAO

The ultimate goal is for the network to be owned and governed by its community. A Decentralized Autonomous Organization (DAO) is the vehicle to achieve this.

- **Modular DAO Frameworks:** Instead of building a governance system from scratch, which is complex and risky, it is recommended to use established and audited DAO frameworks. Options like **Aragon** and **TributeDAO** (based on the popular and secure MolochDAO framework) offer a modular approach.(9), (87), (88) This allows the community to assemble a custom governance structure, like Lego blocks, with modules for different types of voting, treasury management, membership control, and dispute resolution.(10), (89)
- **Progressive Decentralization:** The transition to fully decentralized governance should be a gradual process. Initially, the founding team will make key decisions to guide the protocol's development. As the network matures and token distribution widens, decision-making power will be progressively transferred to token holders through the DAO. This "progressive decentralization" approach ensures stability in the critical early stages of the project.

7.2. The Critical Need for a "Legal Wrapper"

Operating a DAO without a recognized legal entity is an extremely risky proposition. In many jurisdictions, especially in the United States, courts have shown a tendency to classify unstructured DAOs as **"general partnerships."** This classification has a devastating consequence: it imposes **personal, joint and several, and unlimited liability** on all DAO members for the organization's debts, obligations, and legal actions.(90), (91), (92) This means that any token holder, even a passive one, could be personally sued for the entirety of the DAO's liabilities.(16), (93)

To mitigate this existential risk, it is imperative to create a "legal wrapper." The most promising solution is the formation of a **DAO LLC**. Jurisdictions like Wyoming, Tennessee, and Utah in the U.S. have enacted laws that recognize DAOs as a type of Limited Liability Company (LLC).(15), (94), (95) This structure offers the best of both worlds:

- **Limited Liability Protection:** Like a traditional LLC, it protects members (token holders) from personal liability for the DAO's debts. Their liability is limited to their investment.
- **Algorithmic Governance:** The laws are designed to recognize governance through smart contracts and token voting, aligning the legal structure with the DAO's decentralized operation.(16)

7.3. Regulatory Compliance Considerations

Once the legal entity is established, the DAO must operate within the existing legal and regulatory framework.

- **Data Protection (GDPR):** If the network provides services to users in the European Union, it will be subject to the General Data Protection Regulation (GDPR). The DAO, through its legal entity, would be considered a "data controller." Compliance requires adherence to principles such as data minimization, transparency, and privacy by design.(96), (97), (98) The use of Federated Learning, where user data is not centrally collected, is a very strong compliance measure and a significant competitive advantage.(99)
- **Intellectual Property (IP):** This is an emerging and complex legal area for AI.
  1. **IP of the Trained AI Model:** Who owns the AI model that is continuously improved by the community? The strategy most aligned with the spirit of decentralization is to license the model under a permissive open-source license (like MIT or Apache 2.0). The DAO, as the governing entity, would oversee the official versions of the model.
  2. **IP of Generated Outputs:** The current law in most jurisdictions is ambiguous but tends not to grant copyright protection to works generated purely by AI without substantial creative human intervention.(100), (101), (102) The platform's policy must be explicit: the rights to the output generated by a query belong to the user who initiated it. This encourages commercial use and adoption.
- **Securities Regulations:** There is a significant regulatory risk that the NEURO token could be classified as a security by regulators like the SEC in the U.S. To mitigate this risk, the token must be designed to pass the "Howey Test," demonstrating that its primary purpose is **utility** within the ecosystem (used to pay for services, staking, etc.) rather than being a passive investment in a common enterprise with expectations of profits derived from the efforts of others. Despite careful design, this risk persists and will require ongoing, specialized legal advice.(90)

Section 8: Implementation Roadmap and Strategic Recommendations

An ambitious vision requires a pragmatic, phased execution plan. The following roadmap outlines a path from concept to a public, sustainable network, along with strategic recommendations to maximize the chances of success.

8.1. Phased Roadmap

Development will be divided into three main phases, each with clear objectives, technical milestones, and defined success metrics.

Phase	Estimated Duration	Key Objectives	Main Technical Milestones	Success KPIs
Phase 1: Foundational Network (Alpha)	6 months	Validate the core mechanics of task distribution and execution. Build the initial community.	- Development of the orchestration server (Scheduler, Validator). - Creation of the lightweight, cross-platform client with ONNX Runtime. - Implementation of the basic reputation system. - Launch of a closed testnet with a simple model (e.g., image classification).	- 1,000 concurrent active nodes. - Task completion rate >99%. - Discord/Twitter community with >5,000 members.
Phase 2: MoE and Low-Latency Integration (Beta)	9 months	Demonstrate the feasibility of large-scale, low-latency LLM inference.	- Implementation of the Mixture of Experts (MoE) AI model. - Integration of	- Average inference response time < 5 seconds for standard queries. - Successful

			libp2p and optimization of the scheduler for low latency. - Launch of an incentivized testnet with test tokens. - Development of a developer API.	deployment of an MoE model with >50B parameters. - >10,000 active nodes. - >10 developer projects building on the testnet.
Phase 3: Token and DAO Launch (Public)	12 months onwards	Create a sustainable economy and transfer governance to the community.	- Token Generation Event (TGE) and mainnet launch. - Enablement of staking and payment mechanisms with the NEURO token. - Implementation of the DAO framework (e.g., Aragon) and holding the first governance votes. - Establishment of the DAO LLC legal entity.	- Network transaction volume generating sustainable revenue for the treasury. - >50% of the eligible token supply participating in staking and/or governance. - Continuous growth in the number of nodes and API users.

8.2. Strategic Recommendations and Critical Success Factors

- **Build a Strong Community from Day Zero:** The success of a decentralized project is synonymous with the success of its community. It is crucial to invest in building a vibrant and engaged community from the beginning. This involves using proven Web3 strategies such as maintaining active channels on Discord and Telegram, holding "Ask Me Anything" (AMA) sessions with the founding team, publishing educational content on blogs and social media, and organizing hackathons and grant programs to encourage building on the platform.(103), (104) The community is not just a marketing channel; it is the source of compute providers, developers, validators, and future governors of the network.
- **Obsessive Focus on Developer Experience (DevEx):** The long-term success of the network will not depend on end-users massively installing the client, but on developers integrating the distributed AI into their own applications. Therefore, creating a first-class developer experience is fundamental. This includes providing clear and comprehensive documentation, intuitive APIs and SDKs in popular languages (Python, JavaScript), and tools that simplify interaction with the network. The goal is to make using the distributed AI as easy, or even easier, than using an API from a centralized cloud provider.
- **Radical Transparency as a Default Policy:** In an ecosystem that operates on the principle of "don't trust, verify," trust is built through transparency. All key network metrics (number of nodes, tasks processed, revenue generated), governance decisions, meeting minutes, and financial reports from the DAO treasury must be public and easily accessible. This openness fosters accountability and strengthens the community's trust in the project.
- **Continuous and Proactive Legal Counsel:** The regulatory environment for cryptocurrencies, AI, and DAOs is fluid and often hostile. Attempting to navigate this landscape without expert legal advice is a recipe for disaster. It is imperative to hire a law firm specializing in blockchain technology and securities law from the beginning to correctly structure the legal entity, design the tokenomics to minimize regulatory risk, and stay up-to-date with legislative and jurisprudential changes globally.

NEURO Tokenomics and Utility Design Table

Category	Detail	Strategic Rationale
Supply and Distribution	**Maximum Supply:** Fixed (e.g., 1,000,000,000 NEURO)	Predictable scarcity fosters trust and long-term value retention.(8)

	Allocation:	
	- Community and Ecosystem: 40%	The largest portion is allocated to incentivize network participants, ensuring healthy and decentralized growth.(105)
	- DAO Treasury: 20%	Provides funds for future development, grants, and operations, controlled by the community.
	- Founding Team and Advisors: 20%	Rewards initial contributors, aligning their interests with long-term success.
	- Early Investors: 15%	Capital for initial development.
	- Liquidity and Marketing: 5%	Ensures healthy markets at launch and funds initial awareness.
	Vesting Schedule: Team and Investors: 4 years with a 1-year cliff.	Prevents massive sell-offs by insiders after launch, demonstrating long-term commitment.(8), (106)
Utility Mechanisms	**Staking for Nodes:** Minimum NEURO requirement to operate a node.	Ensures network security (economic collateral) and reduces circulating supply.(36), (65)
	Service Payments: All inference fees are paid or settled in NEURO.	Creates constant and fundamental demand for the token, directly linked to network usage.(34), (67)

	Governance: 1 staked token = 1 vote.	Empowers the community and decentralizes decision-making about the protocol's future.(107)
Economic Model	**Issuance Rate:** Controlled inflation in the early years for node rewards, decreasing over time.	Bootstraps the network's supply when fee revenues are low, then reduces to prevent devaluation.(8)
	Burn Mechanism: A % of network fees is permanently burned.	Creates a deflationary mechanism that rewards long-term holders and increases scarcity as usage grows.(69)
	Treasury Revenue: A % of network fees funds the DAO treasury.	Creates a sustainable business model for the protocol, allowing it to self-fund its own growth and development.(38)

Section 9: References

1.(https://github.com/BOINC/boinc/wiki/BOINC-overview)
2.(https://en.wikipedia.org/wiki/BOINC_client%E2%80%93server_technology)
3.(https://docs.libp2p.io/concepts/introduction/overview/)
4.(https://en.wikipedia.org/wiki/Mixture_of_experts)
5. Applying Mixture of Experts in LLM Architectures
6.(https://tangem.com/en/blog/post/depin-crypto/)
7.(https://speedrunethereum.com/guides/sustainable-erc20-supply-models)
8.(https://tas.co.in/blockchain/tokenomics-design-hidden-flaws-that-sink-new-crypto-projects-in-2025/)
9.(https://www.rapidinnovation.io/post/dao-tools-comparison-aragon-vs-daostack-vs-colony)
10.(https://ideausher.com/blog/what-is-aragon-dao/)
11.(https://www.researchgate.net/publication/251898211_An_Optimized_Algorithm_of_P2P_Network_Routing_base_on_Geographic_Position)

12.(https://github.com/BOINC/boinc/wiki/LowLatency)
13.(https://arxiv.org/html/2504.06457v1)
14.(https://www.microsoft.com/en-us/research/publication/secure-multi-party-computation-for-machine-learning-a-survey/)
15.(https://www.l4sb.com/blog/what-is-a-decentralized-autonomous-organization-dao-llc/)
16.(https://aurum.law/newsroom/DAO-3-0-ultimate-dao-legal-structuring-in-2025-and-beyond)
17. Hardware-based Heterogeneous Memory Management for Large Language Model Inference
18.(https://www.iaps.ai/research/are-consumer-gpus-a-problem-for-us-export-controls)
19.(https://arxiv.org/html/2503.11698v1)
20.(https://en.wikipedia.org/wiki/Folding@home)
21.(https://pmc.ncbi.nlm.nih.gov/articles/PMC10398258/)
22. About Folding@home
23.(https://setiathome.berkeley.edu/sah_papers/cacm.php)
24.(https://www.matec-conferences.org/articles/matecconf/pdf/2018/16/matecconf_mms2018_05009.pdf)
25.(https://pr.princeton.edu/pwb/99/0920/p/seti.shtml)
26.(https://github.com/NVIDIA/TensorRT-LLM)
27.(https://www.themoonlight.io/en/review/decentralized-physical-infrastructure-network-depin-challenges-and-opportunities)
28.(https://www.gate.com/learn/articles/2025-de-pin-market-outlook-and-trends/6556)
29.(https://www.coinex.network/en/academy/detail/3030-5-top-depin-projects-to-watch-out-for-in-2025)
30. Akash Network
31. Akash Network Price - Coinbase
32.(https://www.binance.com/en/square/post/6536083007786)
33.(https://www.coingecko.com/learn/what-is-render-network-rndr-crypto)
34.(https://www.gemini.com/cryptopedia/render-network-3d-rendering-software-render-token-rndr-token)
35.(https://www.gate.com/learn/course/de-pin-deep-dives-bittensor/tao-tokenomics)
36.(https://stealthex.io/blog/bittensor-price-prediction-can-tao-coin-reach-1000/)
37.(https://www.findas.org/tokenomics-review/coins/the-tokenomics-of-bittensor-tao/r/MGsxXzUYXLQ2b2zX8xLqaV)
38.(https://coinmarketcap.com/cmc-ai/akash-network/price-prediction/)
39. How AI Model Censorship Impacts Cybersecurity - Kindo
40.(https://freedomhouse.org/report/freedom-net/2023/repressive-power-artificial-intelligence)
41.(https://www.diadata.org/web3-ai-map/gensyn/)
42.(https://web3.bitget.com/en/dapp/gensyn-24546)
43. Volunteer Computing Notes
44.(https://onnxruntime.ai/docs/)
45.(https://github.com/microsoft/onnxruntime)

46.(https://onnxruntime.ai/docs/execution-providers/)
47.(https://en.wikipedia.org/wiki/Llama.cpp)
48.(https://jan.ai/docs/llama-cpp)
49.(https://docs.libp2p.io/)
50.(https://taylorandfrancis.com/knowledge/Engineering_and_technology/Computer_science/Geographic_routing/)
51.(https://www.mdpi.com/2079-9292/10/12/1484)
52.(https://arxiv.org/html/2412.08367v1)
53.(https://www.mdpi.com/2079-9292/14/2/303)
54.(https://core.ac.uk/download/pdf/215700792.pdf)
55.(https://aws.amazon.com/builders-library/leader-election-in-distributed-systems/)
56.(https://engineering.fb.com/2024/08/05/data-center-engineering/roce-network-distributed-ai-training-at-scale/)
57.(https://www.alibabacloud.com/tech-news/a/ai/gssh8sok30-accelerate-ai-model-training-on-gpu-clusters)
58. Hardware-based Heterogeneous Memory Management for Large Language Model Inference
59.(https://openreview.net/forum?id=iTVReq7BtX)
60.(https://arxiv.org/abs/2501.05313)
61. Principles and Components of Federated Learning Architectures
62.(https://arxiv.org/pdf/2101.02373)
63. Principles and Components of Federated Learning Architectures
64.(https://depinhub.io/projects/gensyn)
65.(https://akash.network/token/)
66. Gensyn Litepaper
67.(https://tokenomist.ai/akash-network)
68.(https://www.certik.com/resources/blog/tokenomics-failures-in-2022)
69.(https://tas.co.in/blockchain/tokenomics-design-hidden-flaws-that-sink-new-crypto-projects-in-2025/)
70.(https://www.findas.org/tokenomics-review/coins/the-tokenomics-of-bittensor-tao/r/MGsxXzUYXLQ2b2zX8xLqaV)
71.(https://www.chaincatcher.com/en/article/2161622)
72. A survey of human-in-the-loop for machine learning
73.(https://www.telusdigital.com/glossary/human-in-the-loop)
74.(https://techcommunity.microsoft.com/blog/healthcareandlifesciencesblog/leverage-secure-multi-party-computation-smpc-for-machine-learning-inference-in-r/4057703)
75.(https://www.researchgate.net/publication/379843467_Secure_Multi-Party_Computation_for_Machine_Learning_A_Survey)
76.(https://www.researchgate.net/publication/393501784_Secure_Multi-Party_Computation_Applications_in_collaborative_machine_learning_without_exposing_raw_data)
77.(https://www.researchradicals.com/index.php/rr/article/download/94/88/171)
78.(https://drops.dagstuhl.de/entities/document/10.4230/OASIcs.NG-RES.2025.2)
79. High Performance Privacy Preserving AI - Now Publishers

80.(https://scopicsoftware.com/blog/privacy-preserving-ai/)

81.(https://ai.jmir.org/2025/1/e60847)

82.(https://www.wipo.int/en/web/frontier-technologies/artificial-intelligence/index)

83.(https://www.geeksforgeeks.org/blogs/top-interesting-blockchain-project-ideas-for-beginners/)

84.(https://www.redhat.com/en/topics/edge-computing/latency-sensitive-applications)

85.(https://aerospike.com/blog/real-time-ai-latency-cost-reduction/)

86. Influences of network latency manipulation

87.(https://tributedao.com/docs/intro/overview-and-benefits/)

88.(https://www.gemini.com/cryptopedia/dao-crypto-decentralized-governance-blockchain-governance)

89.(https://www.rapidinnovation.io/post/the-future-of-depin-predictions-and-trends-for-2025-and-beyond)

90.(https://www.corporatedirect.com/blog/five-reasons-why-we-dont-recommend-dao-llcs)

91.(https://www.dechert.com/knowledge/onpoint/2023/4/federal-court-holds-dao-members-can-be-treated-as-general-partne.html)

92.(https://arxiv.org/abs/2408.04717)

93.(https://www.stinson.com/newsroom-publications-decentralized-autonomous-organization-laws-across-the-us)

94.(https://www.heritage.org/government-regulation/commentary/the-legal-status-decentralized-autonomous-organizations-do-daos)

95.(https://www.stinson.com/newsroom-publications-decentralized-autonomous-organization-laws-across-the-us)

96.(https://gdpr.eu/what-is-gdpr/)

97.(https://commission.europa.eu/law/law-topic/data-protection/data-protection-explained_en)

98.(https://www.paloaltonetworks.com/cyberpedia/gdpr-compliance)

99.(https://www.mdpi.com/2076-3417/14/2/675)

100.(https://www.dentons.com/ru/insights/articles/2025/january/28/ai-and-intellectual-property-rights)

101. Intellectual Property Considerations for AI Companies

102.(https://www.financemagnates.com/thought-leadership/ai-has-a-copyright-problem-do-decentralized-networks-have-a-solution/)

103.(https://blog.ethermail.io/how-to-build-an-engaging-web3-community-in-2025)

104.(https://tokenminds.co/blog/web3-marketing/community-building)

105.(https://docs.chainbase.com/introduction/tokenomics)

106.(https://www.magna.so/blog-posts/token-management-and-distribution-in-web3-and-crypto-latest-trends)

107.(https://shardeum.org/blog/what-is-tokenomics/)