ORIGINAL RESEARCH ARTICLE

Ankh AI: A Decentralized Protocol for Collaborative Artificial Intelligence

Coordinating Compute, Talent, and Data Through Blockchain-Based Incentives and Hybrid Verification Mechanisms

Serge Destin TAMPOLLA

4IR Advisor
Digital Transformation & Emerging Technologies
Fourth Industrial Revolution (4IR) Strategy

serge.destin@tampolla.com | www.tampolla.com

ABSTRACT

Background: The development of foundation models is increasingly concentrated among a handful of corporations controlling computational resources, proprietary datasets, and research talent. This centralization poses systemic risks including algorithmic censorship, opaque governance, and unequal value distribution.

Methods: We propose Ankh AI, a decentralized protocol that coordinates the three pillars of modern AI---compute, human expertise, and data---through a dedicated appchain (Cosmos SDK), hybrid verification (zero-knowledge proofs, optimistic verification, VRF-based sampling), Soulbound Token reputation systems, and programmable royalties. The protocol employs a native utility token ($COLLAB) with veTokenomics governance.

Results: The architecture achieves 3,000+ TPS with sub-second finality, supports distributed training via adapted DiLoCo protocols, and enables cryptographically verifiable inference through zk-SNARK circuits (ezKL). Economic modeling projects market capitalization scenarios ranging from $75M to 500M over five years under conservative-to-optimistic adoption assumptions.

Conclusion: Ankh AI demonstrates the technical feasibility and economic viability of decentralized AI model production, offering a transparent, censorship-resistant alternative to centralized solutions while maintaining competitive performance through collective intelligence.

Keywords: decentralized artificial intelligence; blockchain; zero-knowledge proofs; distributed machine learning; tokenomics; DAO governance; collaborative AI; Soulbound Tokens; proof of useful work

1. Introduction

The concentration of artificial intelligence capabilities among a small number of technology corporations has reached unprecedented levels. In 2025, the top three enterprise LLM API providers---Anthropic, OpenAI, and Google---collectively commanded **88%** of corporate generative AI spending [1]. OpenAI, despite market share erosion from 50% (2023) to 27% (late 2025), maintains a valuation of **$260 billion** and coordinates the Stargate infrastructure initiative valued at **$500 billion** [1]. The estimated training cost for next-generation models such as GPT-5 exceeds **$500 million** per run, with some models surpassing **$1 billion** [2,3]. Anthropic's CEO has confirmed that certain models in development exceed the billion-dollar training cost threshold [3].

This centralization creates three systemic vulnerabilities that this paper addresses. *Algorithmic opacity*: model behavior is shaped by unilateral corporate decisions regarding content filtering, response modification, and access restriction, without meaningful user or regulatory oversight. The content moderation policies of major LLM providers have been criticized for lacking transparency and consistency across different linguistic and cultural contexts. *Value extraction asymmetry*: researchers, engineers, and annotators contributing to model improvement receive no proportional share of generated value. A researcher whose architectural innovation reduces training costs by 15% receives neither attribution nor ongoing compensation, despite their contribution being embedded in every subsequent model deployment. *Barrier to entry*: the capital requirements for competitive model training have become prohibitive for independent researchers and smaller organizations [2].

The historical trajectory of AI development illustrates this concentration pattern. From the early days of academic research---where breakthroughs in neural networks emerged from university laboratories and were shared through open publications---the field has progressively shifted toward industrial research laboratories.

The transformer architecture [22], originally published by researchers at Google Brain in 2017, catalyzed the current wave of large language models but was subsequently deployed primarily within corporate frameworks. The compute requirements for training state-of-the-art models have increased by approximately **10x every 18 months** since 2018, far outpacing Moore's Law and effectively excluding all but the best-funded organizations from frontier research [2].

Decentralized approaches offer a structural alternative to this concentration. Blockchain technology enables trustless coordination, immutable contribution tracking, and programmatic value distribution without intermediaries [4]. The emergence of decentralized finance (DeFi) demonstrated that complex economic coordination---previously requiring centralized institutions---can be encoded in smart contracts and executed autonomously. Projects such as Bittensor (market cap $2.54B, 129 active subnets) [5], Gensyn (specialized ML compute verification) [6], and Ocean Protocol (compute-to-data architecture) [7] have demonstrated components of decentralized AI infrastructure. However, no existing protocol unifies the three pillars---compute, talent, and data---within a single coordinated framework.

This paper presents **Ankh AI**, a protocol designed to address this gap. Our contributions are: (i) a unified architectural framework coordinating compute providers, human experts, and data contributors through a dedicated appchain; (ii) a hybrid verification system combining zero-knowledge proofs, optimistic verification, and random sampling; (iii) a reputation mechanism based on non-transferable Soulbound Tokens; (iv) programmable perpetual royalties for intellectual contributions; (v) a comprehensive tokenomics model with veTokenomics governance. The remainder of this paper is organized as follows: Section 2 reviews related work; Section 3 presents the protocol architecture; Section 4 details the contribution and reward mechanisms; Section 5 describes the model lifecycle; Section 6 covers governance and tokenomics; Section 7 addresses security, privacy, and ethics; Section 8 provides discussion; and Section 9 concludes.

2. Related Work

2.1 Decentralized AI Protocols

Bittensor (TAO) employs a Proof-of-Intelligence consensus where miners are rewarded based on output quality assessed by validators [5]. Its dynamic governance (dTAO) allows capital allocation to specific subnets. Bittensor's subnet architecture has grown to 129 active subnets, each specializing in different AI tasks---from text generation to protein folding prediction. The token TAO, with a maximum supply of 21 million modeled after Bitcoin, underwent its first halving in December 2025, reducing daily issuance from 7,200 to 3,600 TAO [5]. Limitations include validator concentration (top 12 validators control 79% of stake) and absence of unified coordination across the three AI pillars [5]. The subnet model, while enabling specialization, fragments the ecosystem and complicates the user experience for contributors seeking to engage across multiple domains.

Gensyn focuses specifically on ML compute verification through referred delegation and cryptographic proofs [6]. Innovations include NoLoCo (low-bandwidth distributed training), CheckFree (checkpoint-free fault tolerance), and RL Swarm (collaborative reinforcement learning) [6]. Gensyn does not natively address human talent coordination or model governance, focusing exclusively on the compute verification problem. This narrow scope, while enabling deep technical specialization, leaves the broader coordination challenge unsolved.

Ocean Protocol pioneered decentralized data economies through Compute-to-Data (C2D), where algorithms execute at data locations without data movement [7]. Data NFTs and Datatokens create liquid markets for data assets. Ocean withdrew from the ASI Alliance in October 2025, illustrating coordination challenges among protocols with partially divergent objectives [8]. The fragmentation of the ASI Alliance---originally conceived as a unified decentralized AI ecosystem--

-demonstrates the difficulty of coordinating independent projects with different governance structures and economic incentives.

Fetch.ai / ASI Alliance focuses on autonomous agents and decentralized coordination, combining Fetch.ai, SingularityNET, and CUDOS with a combined token capitalization of $450M [8]. The ASI Chain under development targets L1 blockchain for agentic AI. The alliance faces organizational complexity following Ocean's departure, with questions remaining about the effective integration of the three constituent projects.

Akash Network and **Render Network** provide decentralized GPU infrastructure. Akash achieved 80% GPU utilization in 2025 with 34,300+ active leases per quarter [9]; Render processed 68M+ frames [10]. Neither addresses model training coordination, intellectual contribution rewards, or inference marketplaces. Their role in the broader ecosystem is complementary: they provide the computational substrate upon which protocols like Ankh AI can build higher-level coordination mechanisms. Akash's GPU marketplace demonstrated that decentralized compute pricing can undercut centralized cloud providers by 50--80%, establishing the economic viability of the decentralized compute model [9].

Table 1. Comparative analysis of decentralized AI protocols

Dimension	Ankh AI	Bittensor	Gensyn	Ocean
Scope	3 pillars unified	AI marketplace	ML compute	Data economy
Verification	ZK + Opt. + Sampling	Validator eval.	Probabilistic	N/A
Reputation	SBT on-chain	Implicit score	Basic	N/A
Royalties	Programmable, perpetual	No	No	Datatokens
Governance	Federated DAO +ve	dTAO	Elected council	Classic DAO
Supply cap	1B $COLLAB	21M TAO	Unspecified	1.41B OCEAN

2.2 Zero-Knowledge Machine Learning

Zero-knowledge proofs enable verification of ML inference without revealing model parameters or input data. The ezKL library converts ONNX models to ZK-SNARK compatible circuits, generating proofs in minutes with second-scale verification [11]. Modulus Labs demonstrated on-chain verification of models up to **18 million parameters** [12]. This capability is essential for trustless inference: users can cryptographically verify that a provider executed the promised model (e.g., GPT-4 rather than GPT-3) without accessing proprietary weights [12].

Recent advances in zkML have reduced proof generation time by orders of magnitude. The development of specialized proving systems---including Groth16, PLONK, and STARKs---has enabled practical deployment of ZK proofs for neural network inference [12]. Lookup arguments and sum-check protocols have further optimized the arithmetic circuit representation of common ML operations such as matrix multiplication and activation functions. The field is rapidly evolving, with new techniques emerging that promise to extend verifiable inference to models with billions of parameters within the next 2--3 years.

2.3 Tokenomics and Governance Mechanisms

Curve Finance's veTokenomics established the precedent of time-weighted voting power through vote-escrowed tokens [13]. Locking duration determines governance influence, aligning long-term stakeholder interests with protocol health. The mathematical elegance of this approach---where voting power decays linearly with remaining lock time---has inspired numerous subsequent protocols. Bittensor's dTAO extends this to subnet-level capital allocation [5]. These mechanisms inform Ankh AI's veCOLLAB design.

The academic literature on mechanism design for decentralized systems provides theoretical foundations for Ankh AI's approach. Vickrey-Clarke-Groves (VCG) mechanisms, though computationally expensive for large-scale systems, inform the design of the reverse auction system for compute allocation. The concept of proof-of-useful-work, distinct from Bitcoin's proof-of-work, aligns economic incentives with productive computation [6]. Research on decentralized reputation systems demonstrates that non-transferable credentials (Soulbound Tokens) provide stronger anti-Sybil guarantees than transferable tokens [15].

3. Protocol Architecture

3.1 System Overview

The Ankh AI protocol comprises five interconnected layers. The *Blockchain Coordination Layer* (Cosmos SDK appchain) provides transaction finality, smart contract execution, and on-chain governance with 3,000+ TPS and sub-3-second finality. The *Smart Contract Layer* implements six core contracts: ComputeRegistry, TaskScheduler, DataMarketplace, RoyaltyEngine, ReputationSystem, and GovernanceCore. The *Verification Layer* combines ZK proofs for inference, optimistic verification for training, and VRF-based sampling for audit. The *Storage Layer* uses IPFS, Filecoin, and Arweave with TEE-based encryption. The *Application Layer* provides APIs, SDKs, and user interfaces.

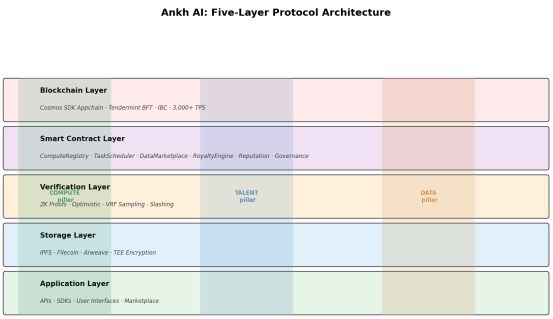

Figure 1. *The five-layer Ankh AI protocol architecture showing the compute, talent, and data pillars spanning all layers*

3.2 Blockchain Infrastructure

We select a dedicated appchain over Ethereum L2 for sovereignty and cost efficiency. The Cosmos SDK enables parameter customization (block size, fees, consensus timing) critical for a protocol generating thousands of micro-task transactions per second. Tendermint BFT provides instant finality with a 2-second target block time. IBC protocol ensures native interoperability with Cosmos Hub, Osmosis, and Akash; secure bridges extend connectivity to Ethereum and EVM-compatible chains.

The choice of a dedicated appchain reflects a fundamental trade-off in blockchain design. While Ethereum L2s benefit from the security guarantees of the Ethereum mainnet, they are subject to the congestion and fee volatility of the underlying chain. For a protocol whose core function involves coordinating thousands of micro-transactions per second---task assignments, proof submissions, reward distributions---the cost and latency of Ethereum mainnet interactions would be prohibitive. The Cosmos ecosystem, with its mature IBC interoperability protocol and growing suite of DeFi and infrastructure protocols, provides an optimal balance of sovereignty, performance, and connectivity.

Table 2. Blockchain infrastructure comparison

Parameter	Ankh AI (Appchain)	Ethereum L2	Parachain
Throughput (TPS)	3,000+	500--2,000	1,000+
Finality	~2 seconds	~12 seconds	~12 seconds
Tx fees	<$0.001	$0.01 - -1	<$0.01
Sovereignty	Full	L1-dependent	Shared
Smart contract lang.	CosmWasm + EVM	Solidity	Ink! (Rust)

3.3 Core Smart Contracts

ComputeRegistry maintains the canonical registry of compute providers. Node registration requires staking a minimum of 1,000 $COLLAB. The contract records hardware specifications (GPU type, VRAM, bandwidth), performance metrics (verified FLOPS, uptime, latency), and active status. Slashing penalizes malicious behavior: invalid computation proofs, confirmed fraud, or unjustified extended downtime. The economic security model requires that the cost of attacking the network (acquiring sufficient stake to control consensus) exceeds the potential gains from such an attack.

```
function registerNode(
    bytes32 hardwareFingerprint,
    uint256 stakeAmount,
    NodeSpec calldata specs
) external {
      require(stakeAmount >= MIN_STAKE, "Insufficient
stake");
      require(specs.vramGB >= MIN_VRAM, "Insufficient
VRAM");
    nodes[msg.sender] = Node({
        stake: stakeAmount,
        specs: specs,
        status: NodeStatus.ACTIVE,
        reputationScore: BASE_REPUTATION,
        registeredAt: block.timestamp
    });
    totalStaked += stakeAmount;
                    emit     NodeRegistered(msg.sender,
hardwareFingerprint);
}
```

TaskScheduler orchestrates training and inference task distribution. It maintains a priority queue of pending tasks, selects appropriate nodes based on task requirements and node capabilities, tracks execution in real time, and triggers payments upon result validation. For distributed training, TaskScheduler implements an adapted DiLoCo protocol [14], dividing training into asynchronous compute islands with periodic gradient synchronization. The scheduling algorithm optimizes for a weighted combination of expected execution time, cost efficiency, and fault tolerance. The scheduler must balance the competing objectives of minimizing latency (favoring high-performance nodes) and maximizing decentralization (favoring geographic and provider diversity).

DataMarketplace manages the complete data asset lifecycle. Each dataset is represented as an extended ERC-721 Data NFT recording ownership, provenance, and license terms. Datatokens (ERC-20) enable fractional access and liquidity. The contract integrates Compute-to-Data: training algorithms execute at data locations rather than moving data, preserving confidentiality [7]. The marketplace supports multiple pricing models including fixed-price sales, Dutch auctions, and subscription-based access. The provenance tracking ensures that all data used in model training can be traced back to its original source, a critical requirement for regulatory compliance and bias auditing.

RoyaltyEngine automates revenue distribution from deployed models. Each model records its "contribution recipe"---the intellectual contributions, datasets, and optimizations enabling its creation. When inference fees are paid, the engine decomposes the amount according to programmed percentages and distributes payments to each contributor in real time. This mechanism ensures that contributors to a model's creation continue to benefit from its success throughout its operational lifetime, addressing the fundamental asymmetry in current AI economics where contributors receive one-time compensation while their work generates ongoing value.

ReputationSystem manages Soulbound Token issuance, updates, and revocation. Each SBT is non-transferable and bound to a specific address. The system computes a composite score from on-chain metrics: accepted contributions, peer evaluations, tenure, and technical specialization. This score determines governance voting weight, access to high-value tasks, and reward multipliers [15]. The non-transferability property is critical: it prevents the emergence of secondary markets for reputation, which would undermine the integrity of the system by enabling wealthy participants to purchase influence rather than earning it through contribution.

GovernanceCore implements proposal, voting, and execution mechanisms. Supported proposal types include protocol improvement proposals (PIPs), budget allocation, research priority selection, and emergency decisions. The system incorporates timelock delays for critical modifications, quadratic voting for allocation decisions, and adaptive quorums scaled by decision type. The emergency mechanism, controlled by a Security Council, enables rapid response to critical vulnerabilities while requiring subsequent DAO ratification.

3.4 Hybrid Verification Layer

The integrity of models produced by Ankh AI relies on a three-tier verification architecture, with each tier optimized for a specific computation type. This hybrid approach addresses the fundamental challenge that no single verification mechanism is

optimal for all operations in the AI pipeline: inference requires fast, low-latency verification; training requires cost-effective verification of massive computations; and ongoing audit requires statistical sampling.

ZK Verification for Inference. The ezKL library converts ONNX-exported models to ZK-SNARK compatible circuits [11]. The four-step process---model export, circuit conversion, proof generation, and public verification---completes in minutes with cryptographic guarantees. This solves the fundamental trust problem: users can verify that providers executed the promised model without accessing proprietary parameters [12]. The current limitation is model size: ezKL handles models up to approximately 18 million parameters on-chain. For larger models, the proof generation is performed off-chain with the resulting proof verified on-chain, maintaining the trustless property while accommodating larger architectures.

The ezKL workflow begins with model export to the ONNX format, a standard representation that captures the computational graph of neural network operations. The ezKL compiler then converts this graph into an arithmetic circuit suitable for ZK proof generation. The proving step generates a succinct proof that the model was executed correctly on the given inputs, producing the claimed outputs. Finally, the verification step---which can be performed by any party---confirms the proof's validity in constant time regardless of model size. This property is crucial for on-chain verification, where computational resources are limited.

Optimistic Verification for Training. Full ZK verification of foundation model training is prohibitively expensive---generating proofs for billions of operations would take days and cost thousands of dollars. Ankh AI adopts optimistic verification: training results (aggregated gradients, updated weights) are submitted and accepted by default. During a 24-hour challenge window, any validator may submit a fraud proof. Confirmed fraud triggers slashing of the malicious node and rewards the challenger. This approach, inspired by optimistic rollups in the Ethereum ecosystem, trades immediate finality for dramatic cost reduction while maintaining cryptographic security guarantees.

The economic design of the optimistic verification system ensures that rational actors have no incentive to submit fraudulent results. The slashing penalty for confirmed fraud exceeds the potential gain from a successful deception, and the reward for valid challenges ensures that validators have strong incentives to monitor submitted results. The 24-hour challenge window balances the need for timely finality with the practical constraints of generating and verifying fraud proofs.

VRF-Based Sampling. Verification committees are randomly selected from validator nodes using cryptographically verifiable random functions (VRF) [16]. This selection guarantees impartiality: no node can predict or influence its selection. Committees audit a statistically significant sample of computations, providing independent validation complementing ZK proofs and optimistic verification. The sampling rate is calibrated to achieve a target false negative rate (undetected fraud) below 0.1% while minimizing verification overhead.

The VRF mechanism operates as follows: each validator registers a public key during node registration. When a verification committee needs to be selected, the protocol generates a random seed from the blockchain's entropy (derived from block hashes and other unpredictable sources). Each validator applies the VRF to this seed using their private key, producing a pseudorandom output and a proof of correct computation. Validators with the highest VRF outputs are selected for the committee. The proof ensures that the selection was truly random and not manipulated.

4. Contribution and Reward Mechanisms

4.1 Compute Pillar

Ankh AI accommodates four compute provider categories. *Professional data centers* deploy H100/H200/Blackwell GPU clusters with highest

density and availability guarantees. These providers form the backbone of the training infrastructure, offering the sustained high-performance compute required for foundation model training. *Decentralized cloud providers* (Akash, Render) aggregate resources under unified interfaces, providing elastic capacity that scales with demand. *Individual contributors* (RTX 4090/5090 gaming GPUs) provide underutilized capacity; clustered RTX 4090s reduce inference costs by **75%** versus H100s with minimal batch-processing performance loss [9]. *Edge devices* (IoT, mobile NPUs) contribute to lightweight tasks such as model validation, data preprocessing, and small-batch inference.

Proof of Useful Work. Unlike Bitcoin's Proof of Work, where computation solves cryptographic puzzles with no practical utility, Ankh AI's Proof of Useful Work directs every GPU cycle toward model training or inference. The fundamental metric is *verified FLOPS*: floating-point operations effectively executed and cryptographically verified. Complementary quality-of-service indicators include uptime (target >99%), response latency, and output fidelity. The verification process ensures that reported compute contributions correspond to actual productive work, preventing the "hashrate inflation" problem that has plagued some decentralized compute networks.

Phase-Dependent Rewards. Training-phase providers compete via reverse auction: the protocol publishes compute volume requests and providers bid competitively. The auction mechanism ensures market-driven pricing that reflects real-time supply and demand conditions. The reverse auction design is particularly suited to decentralized compute markets because it allows providers with diverse cost structures (professional data centers with low marginal costs vs individual GPU owners with higher opportunity costs) to find their optimal price points.

Inference providers receive **70%** of user fees, payable in stablecoins or $COLLAB (with 5% COLLAB at a premium. Fine-tuning providers earn a share of specialized model fees plus quality bonuses for benchmark achievement. This three-phase reward structure aligns incentives across the entire model lifecycle, from initial training through deployment to specialization.

4.2 Human Talent Pillar

Five contributor categories are recognized with adapted mechanisms. *Researchers* design model architectures, optimization methods, and alignment techniques. Their contributions are evaluated through peer review and empirical validation on benchmark tasks. *ML Engineers* implement and optimize models for distributed infrastructure, ensuring efficient execution across heterogeneous hardware. *MLOps/Infrastructure Engineers* manage deployment and scalability, maintaining the health and performance of the compute network. *Annotators and Domain Experts* provide RLHF alignment data and specialized knowledge [17]. The quality of their annotations directly impacts model safety and utility. *Auditors* evaluate security, bias, and ethical compliance, providing independent validation of model behavior.

Proof of Expertise via Soulbound Tokens. Each contributor possesses an on-chain SBT portfolio recording accepted contributions, peer evaluations, participation regularity, and certified specializations [15]. SBTs are non-transferable, preventing reputation marketization. The composite reputation score determines governance weight, task access privileges, and reward multipliers. The mathematical formulation of the reputation score is:

$$R(u) = w_1 \cdot C(u) + w_2 \cdot E(u) + w_3 \cdot T(u) + w_4 \cdot S(u) \quad (1)$$

where $C(u)$ counts accepted contributions, $E(u)$ averages peer evaluations, $T(u)$ measures tenure, and $S(u)$ quantifies specialization depth. Weights w_i sum to unity and are calibrated through protocol governance.

Bounty and Royalty System. Specific tasks are published as bounties with community-elected winners receiving $COLLAB rewards. Perpetual royalties provide ongoing income: a researcher whose optimization technique is integrated into a deployed model receives **0.1%** of all inference fees for the model's operational lifetime. This creates a powerful incentive for fundamental research: a

single breakthrough can generate passive income for years. Research grants fund high-risk, high-potential exploratory projects that may not yield immediate results but could catalyze future breakthroughs. Streaming payments via protocols like Superfluid enable continuous compensation for ongoing contributions [18].

4.3 Data Pillar

Five data categories are accepted: raw corpora (web crawls, archives); curated datasets (cleaned, deduplicated, quality-filtered); alignment data (RLHF preference pairs); specialized domain data (medical, legal, financial); and synthetic data (model-generated augmentation). The diversity of accepted data types reflects the multifaceted nature of modern AI training pipelines, where different types of data serve different purposes in the model development process.

Dynamic Data Valuation. Data value is determined by a three-factor oracle:

$$V(D) = \alpha \cdot \Delta P(D) + \beta \cdot R(D) + \gamma \cdot Dem(D) \quad (2)$$

where $\alpha + \beta + \gamma = 1$; $\Delta P(D)$ measures perplexity improvement after dataset integration; $R(D)$ evaluates corpus rarity via cosine dissimilarity with existing datasets; and $Dem(D)$ quantifies expressed demand from specialized model consumers. Uniqueness verification combines cryptographic hashing with vector embedding cosine similarity. Quality verification employs a proxy model to estimate impact on full-scale models. This multi-factor approach addresses the challenge of pricing data in the absence of established markets: rather than relying on subjective human valuation, the oracle derives value from measurable impact on model performance.

5. Model Lifecycle Phases

5.1 Foundation Model Training

The initial training phase produces a competitive foundation model serving as the base for all subsequent applications. The first target---**Ankh-GPT-7B** (7 billion parameters)---balances capability with decentralized deployment efficiency. The estimated budget of **$5--10 million** spans 3--6 months, covering compute allocation, contributor rewards, security audits, and operational reserves [2]. For comparison, GPT-4 training exceeded $100 million, though modern efficient architectures (e.g., DeepSeek R1) have demonstrated competitive performance at one-tenth the compute cost [3].

Distributed training coordination employs the **Decoupled DiLoCo** protocol [14], dividing global training into independent compute islands (typically regional data centers or clusters) that train local model copies. Islands synchronize gradients via lightweight gossip rather than bandwidth-intensive all-reduce operations. This approach naturally isolates local failures: if one island becomes unavailable, others continue uninterrupted [14]. The synchronization frequency is adaptively tuned based on network conditions and convergence metrics. Early experiments with DiLoCo demonstrated that the approach can achieve comparable convergence to centralized training with only a modest increase in total training steps (typically 10--20%).

Table 3. Training reward allocation

Category	Allocation	Description
Compute	45--50%	GPU providers, data centers, individuals
Talent & R&D	20--25%	Researchers, ML engineers, MLOps
Data	15--20%	Dataset providers, annotators
Treasury & Audit	5--10%	DAO reserves, security audits
Community Reserve	5%	Bounties, grants, incentives

5.2 Deployment and Inference

Once trained and validated, the foundation model deploys to the decentralized inference network. On-chain load balancing dynamically selects optimal nodes based on estimated latency, current load, and requested service tier. Three service tiers are available: *Standard* (moderate latency, cost-effective for consumer applications); *Low-Latency* (sub-second responses for real-time applications); and *Confidential* (TEE-isolated execution protecting both model and input data) [19].

The inference API is designed to be compatible with existing OpenAI API specifications, minimizing migration friction for developers currently using centralized services. This compatibility strategy is critical for adoption: by offering a drop-in replacement with lower costs and stronger privacy guarantees, Ankh AI can capture market share from existing providers without requiring users to redesign their applications.

Inference fee distribution is programmed into RoyaltyEngine and executes atomically:

Table 4. Inference fee distribution

Destination	Percentage
Compute provider	70%
Contributor royalties	15%
DAO treasury	10%
Token burn (deflationary)	5%

The **5% burn mechanism** creates structural deflationary pressure: increased network usage mechanically reduces circulating supply, supporting token value---analogous to Ethereum's EIP-1559 mechanism. Under moderate adoption scenarios, the annual burn could reach 5--10 million $COLLAB by year three, representing a significant deflationary force.

5.3 Fine-Tuning and Specialization

The fine-tuning pipeline follows five stages: community proposal (domain, datasets, architecture, budget); financing (self-funded, DAO allocation, or crowdfunding); specialized data contribution with dynamic valuation; distributed fine-tuning execution with full verification; and independent benchmark validation before deployment.

Each validated specialized model is represented by a fractional ERC-1155 NFT encoding collective ownership. Shares distribute to fine-tuning contributors: data providers, compute operators, and architects. Ownership confers proportional revenue rights. A secondary market enables share trading, creating liquidity for specialized model investments. This tokenization mechanism transforms AI models into tradeable assets, enabling new forms of investment and speculation while ensuring that value flows to contributors rather than intermediaries.

6. Governance and Tokenomics

6.1 DAO Structure

The Ankh AI DAO adopts a federated structure balancing broad participation with technical expertise. The *General Assembly* (all veCOLLAB holders) votes on strategic decisions: protocol parameter changes, major budget allocations, and council elections. Three *Specialized Councils* provide targeted expertise: Technical Council (blockchain architects, AI researchers, security engineers); Economic Council (tokenomics analysts, financial strategists); and Ethics Council (philosophers, legal experts, civil society representatives). Four *Sub-DAOs* manage operational domains (Data, Compute, Models, Ecosystem) with significant operational autonomy.

This federated structure addresses a fundamental challenge in decentralized governance: the tension between democratic participation and technical competence. Pure token-weighted voting can lead to decisions that are popular but technically unsound, while expert-only governance undermines the decentralization principle. The council system provides a middle ground, ensuring that technical decisions are informed by expertise while ultimate authority remains with the token-holding community.

6.2 veTokenomics

Vote-escrowed $COLLAB (veCOLLAB) is obtained by locking tokens for 1 week to 4 years. Voting power follows:

$$veCOLLAB = COLLAB_{locked} \times \frac{t_{remaining}}{t_{max}} \quad (3)$$

where $t_{max} = 4$ years. This linear formula ensures that $100\ COLLAB locked for 4 years equals the voting power of 400$ COLLAB locked for 1 year. Power decays linearly as unlock approaches, incentivizing regular re-locking for maximum influence [13].

Adaptive quorums scale with decision significance: minor technical proposals require 10% participation and simple majority; major changes (architecture, >5% treasury) require 25% participation and 60% supermajority; governance rule modifications require 40% participation and two-thirds majority. Quadratic voting attenuates whale influence for allocation decisions. A 48-hour timelock provides reaction time for controversial decisions. The combination of these mechanisms---time-weighted voting, adaptive quorums, quadratic allocation voting, and timelock delays---creates a robust governance framework that resists both capture by wealthy actors and paralysis by excessive caution.

6.3 Token Distribution and Value Mechanisms

Total $COLLAB supply is capped at **1 billion tokens**, immutably fixed. Initial distribution: Community/Ecosystem 40%; Early Investors 20% (4-year vesting, 12-month cliff); Founding Team 15% (5-year vesting, 18-month cliff); DAO Treasury 15%; Strategic Partners 5% (3-year vesting, 6-month cliff); Airdrops/Liquidity 5%. The long vesting periods for the founding team (5 years with 18-month cliff) and early investors (4 years with 12-month cliff) align their interests with the protocol's long-term success and prevent early sell pressure.

Post-TGE emission follows a decreasing curve: 8% annual rate in year 1, declining 20% each year until reaching a 1% floor. The emission schedule is:

Table 6. $COLLAB emission schedule

Year	Annual Rate	New $COLLAB (M)	Cumulative (M)
1	8.0%	80	80
2	6.4%	61	141
3	5.1%	44	185
4	4.1%	33	218
5	3.3%	26	244
6+	1.0% (floor)	~10	---

This declining issuance guarantees attractive early contributor rewards while ensuring long-term sustainability. The 5% burn on inference fees creates a direct correlation between network usage and supply reduction. Staking rewards (8--15% APR) incentivize supply lock-up, further reducing circulating pressure. The combined effect of declining issuance and increasing burn creates a supply dynamic that becomes progressively more deflationary as network adoption grows.

Value capture mechanisms. The $COLLAB token derives value from multiple demand drivers (1) staking requirement for compute providers (minimum 1, COLLAB per node); (2) payment medium for inference services (with 5% discount vs stablecoins); (3) governance rights through veCOLLAB conversion; (4) fee accrual via the 10% protocol fee directed to treasury; and (5) deflationary pressure from the 5% burn mechanism. These multiple demand drivers reduce correlation with broader crypto market cycles and anchor token value to protocol usage.

Table 5. Five-year capitalization projections ($M)

Scenario	Year 1	Year 2	Year 3	Year 4	Year 5
Conservative	5	12	25	45	75
Moderate	8	20	50	100	200
Optimistic	15	40	100	250	500

7. Security, Privacy, and Ethics

7.1 Protocol Security

Attack vectors and mitigations include: *Sybil attacks* (mitigated by mandatory staking, non-transferable SBTs, and on-chain behavior analysis); *training data poisoning* (detected through random sampling, with economic slashing of fraudulent providers; diversity requirements ensure no single source exceeds 10% of training corpus); *inference evasion* (filtered by real-time security checks and RLHF alignment); and *economic attacks* (rendered prohibitively expensive by supply caps, long vesting periods, and veTokenomics) [15,16].

All core smart contracts undergo third-party audit (Trail of Bits, OpenZeppelin, CertiK) before mainnet deployment. A permanent bug bounty program rewards researchers from 1,000 to 500,000 $COLLAB based on severity. An emergency update mechanism, controlled by a 7-member Security Council multi-sig, enables critical vulnerability patching with mandatory DAO ratification within 7 days. The security model follows defense-in-depth principles, with multiple

independent layers of protection rather than reliance on any single mechanism.

7.2 Data Privacy

Trusted Execution Environments (Intel SGX, AMD SEV-SNP, AWS Nitro Enclaves, NVIDIA Confidential Computing) create hardware-isolated enclaves where data processes in plaintext without exposure to host OS, hypervisor, or cloud provider [19]. Federated learning enables model training at data locations, with only aggregated gradients shared centrally [20]. Zero-knowledge proofs allow computation verification without revealing model parameters or inputs [12]. Differential privacy adds calibrated statistical noise to protect individual identities in datasets [20].

The combination of these techniques provides a comprehensive privacy toolkit. For highly sensitive data (medical records, financial transactions), TEEs provide the strongest guarantees by isolating computation at the hardware level. For moderately sensitive data, federated learning with differential privacy offers a lighter-weight alternative. For inference verification, ZK proofs enable public validation without disclosure. The protocol selects the appropriate technique based on data sensitivity, computational constraints, and regulatory requirements.

7.3 AI Alignment and Safety

Distributed RLHF leverages geographically and culturally diverse annotators, reducing cultural and ideological biases prevalent in centrally annotated models [17]. Community red-teaming before major deployments incentivizes proactive vulnerability discovery through economic rewards for validated findings. A decentralized emergency halt mechanism requires 75% super-quorum or Security Council consensus, preventing both paralysis and abusive suspension. All deployed models publish a "bias passport" documenting training data provenance, annotator demographics, benchmark evaluations, and known limitations.

The bias passport represents a significant advance in AI transparency. Unlike current practices where model documentation is often incomplete or opaque, the on-chain bias passport provides an immutable, publicly verifiable record of a model's training history and known limitations. This enables users to make informed decisions about model suitability for their specific use cases and provides regulators with the information needed for compliance assessment.

8. Discussion

The Ankh AI protocol addresses a fundamental tension in contemporary AI development: the incompatibility between the collaborative, open nature of scientific progress and the proprietary, centralized model of industrial AI production. By encoding contribution tracking, value distribution, and governance into immutable smart contracts, the protocol creates institutional infrastructure for collective intelligence production.

Several design decisions warrant critical examination. The hybrid verification approach trades cryptographic completeness for economic practicality: ZK proofs guarantee inference integrity but remain computationally expensive for training-scale verification. The optimistic layer introduces a 24-hour challenge window that delays final reward distribution---an acceptable trade-off given the cost reduction versus full ZK verification. The SBT-based reputation system, while preventing reputation marketization, creates permanent records that may disadvantage contributors with early mistakes; mechanisms for reputation rehabilitation through sustained positive contribution are under development.

The tokenomics model's sustainability depends critically on network effects: compute providers require model demand; models require compute supply; data providers require active training. This chicken-and-egg problem is addressed through substantial initial funding to bootstrap both market sides simultaneously, with enhanced incentives for early contributors accepting nascent-network risk. The declining emission curve (8% to 1% over 5+ years) ensures that reward sustainability improves as the protocol matures. Historical precedents from other decentralized networks (e.g., Bitcoin's early mining rewards, Ethereum's initial issuance)

suggest that generous early incentives are essential for achieving the critical mass needed for self-sustaining network effects.

Regulatory considerations span token classification (utility vs. security), AI regulation (EU AI Act compliance) [21], and data intellectual property. The AI Act's transparency and traceability requirements are natively satisfied by on-chain provenance records. Programmable data licenses encode usage rights in smart contracts, addressing copyright concerns through explicit permission frameworks. The protocol's design anticipates regulatory evolution by building compliance mechanisms into the core architecture rather than retrofitting them as afterthoughts.

Compared to existing decentralized AI projects, Ankh AI's primary differentiator is the unified coordination of all three production pillars. While Bittensor excels at incentivizing model outputs, Gensyn at compute verification, and Ocean Protocol at data markets, none provides the integrated pipeline from raw resources to deployed model that Ankh AI offers. This integration creates powerful network effects: compute attracts developers, developers attract data providers, improved data enhances models, and better models generate more inference demand that rewards all contributors. The resulting flywheel effect, once initiated, can drive exponential growth in network value and capability.

The limitations of the current proposal should be acknowledged. The DiLoCo-based distributed training, while promising, has not been validated at the scale of billion-parameter models on heterogeneous, globally distributed infrastructure. Early experiments suggest a 10--20% increase in training steps compared to centralized training, but the impact on wall-clock time depends heavily on network conditions and synchronization frequency. The ZK proof generation for large models remains computationally expensive, potentially adding significant latency to the inference pipeline. Current ezKL implementations require several minutes to generate proofs for models with tens of millions of parameters---acceptable for batch inference but problematic for real-time applications.

The reputation system's effectiveness depends on the quality and integrity of peer evaluations, which may be subject to collusion or bias. Mitigating these risks requires careful design of evaluation criteria, multi-source validation, and ongoing monitoring for anomalous patterns. The chicken-and-egg problem of network bootstrapping remains a significant challenge: compute providers need model demand to justify participation, while model development needs compute supply to proceed. Addressing this requires substantial initial capital commitment and creative incentive design.

From a broader perspective, Ankh AI contributes to an emerging research agenda at the intersection of distributed systems, cryptography, and artificial intelligence. The protocol demonstrates that blockchain technology can serve not merely as a financial infrastructure but as a coordination mechanism for complex collaborative production. This insight extends beyond AI to other domains where decentralized coordination of specialized resources could replace centralized intermediaries.

9. Conclusion

This paper presented Ankh AI, a decentralized protocol unifying the three pillars of AI production---compute, human expertise, and data---within a blockchain-coordinated framework. The protocol's hybrid verification architecture (ZK proofs, optimistic verification, random sampling), SBT-based reputation system, programmable royalties, and veTokenomics governance collectively enable transparent, censorship-resistant, and equitably distributed AI model production.

The technical feasibility is supported by existing components: ezKL for ZK-ML verification [11], DiLoCo for distributed training [14], TEEs for confidential computation [19], and proven tokenomics mechanisms [13]. The economic viability is projected across three scenarios, with the moderate case projecting $200M capitalization by year five based on 2--5% capture of the decentralized inference market. The protocol's success will ultimately be measured by its ability to produce models competitive with centralized

alternatives while maintaining the transparency, resilience, and equity that decentralization enables.

Future work includes several critical directions. *Testnet implementation* of the complete verification pipeline will validate the theoretical design under realistic network conditions. *Empirical benchmarking* of distributed training efficiency versus centralized baselines is essential for demonstrating competitive performance---initial targets include measuring the convergence rate, wall-clock training time, and resource efficiency of DiLoCo-based training at the 7B and 70B parameter scales. *Reputation rehabilitation mechanisms* will address the concern that early mistakes permanently damage contributor standing, potentially through decay functions that reduce the weight of older negative evaluations over time. *Partnership development* with data center operators, academic institutions, and complementary blockchain protocols will expand the network's reach and capability.

Additional research directions include: extending ZK proof capabilities to larger models through recursive proof composition and hardware acceleration; developing formal verification tools for the core smart contracts to complement traditional security audits; and exploring the integration of formal verification techniques for AI model behavior to complement the empirical verification approaches described in this paper.

The long-term vision extends beyond technical achievement to the establishment of a new paradigm for AI development---one where the benefits of artificial intelligence are shared by those who create it, and where the direction of AI research is determined by collective intelligence rather than corporate strategy. In this vision, a researcher in Nairobi, a data annotator in Mumbai, and a GPU provider in Berlin can collaborate seamlessly, with their contributions tracked, verified, and rewarded through an open, transparent protocol. The democratization of AI production that Ankh AI proposes is not merely a technical achievement but a social one: it represents a shift from extraction to collaboration, from closed to open, and from centralized control to distributed governance.

References

SQ Magazine. (2026). OpenAI Statistics 2026. *sqmagazine.co.uk*.

Life Architect AI. (2026). GPT-5 (2025). *lifearchitect.ai*.

Fanatical Futurist. (2025). OpenAI GPT-5 is costing $500 Million per training run. *fanaticalfuturist.com*.

Nakamoto, S. (2008). Bitcoin: A Peer-to-Peer Electronic Cash System. *bitcoin.org*.

CoinMarketCap. (2026). Bittensor (TAO) --- Market Data and Analysis. *coinmarketcap.com*; CoinStats. (2026). Bittensor (TAO) Investment Analysis. *coinstats.app*.

Whales Market. (2025). What is Gensyn ($AI)? *whales.market*; Gensyn Documentation. (2026). Products & Research. *docs.gensyn.ai*.

Blockchain App Factory. (2025). How Ocean Protocol Marketed Decentralized Data for AI. *blockchainappfactory.com*.

CoinStats. (2026). Artificial Superintelligence Alliance (FET) --- Fundamental Analysis. *coinstats.app*.

Messari. (2026). State of Akash Q4 2025. *messari.io*; BlockEden. (2026). Decentralized Compute for AI. *blockeden.xyz*.

Render Network. (2026). Monthly Report --- December 2025. *rendernetwork.medium.com*.

EZKL Documentation. (2026). The EZKL System. *docs.ezkl.xyz*.

Kudelski Security. (2025). ZKML: Verifiable Machine Learning using Zero-Knowledge Proof. *kudelskisecurity.com*; arXiv. (2025). A Survey of Zero-Knowledge Proof Based Verifiable Machine Learning. *arxiv.org*.

Messari. (2025). Vote Escrow Explained. *messari.io*; CoinGecko. (2023). What are veTokens and Understanding veTokenomics. *coingecko.com*.

Google DeepMind. (2026). Decoupled DiLoCo: Resilient, Distributed AI Training at Scale. *deepmind.google*.

ACM Digital Library. (2025). Soulbound Token Applications: A Case Study in the Health Sector. *dl.acm.org*.

Chainlink. (2026). What Is a Verifiable Random Function (VRF)? *chain.link*.

ACM Digital Library. (2025). RLHF Deciphered: A Critical Analysis. *dl.acm.org*.

CryptoSlate. (2026). Superfluid --- Technology. *cryptoslate.com*.

AI21 Labs. (2025). What are Trusted Execution Environments (TEE)? *ai21.com*; Chainlink. (2026). What Is a Trusted Execution Environment (TEE)? *chain.link*.

ScienceDirect. (2025). Zero-knowledge machine learning models for blockchain peer-to-peer energy trading. *sciencedirect.com*.

EU AI Act. (2026). Official Updates and Compliance Information. *artificial-intelligence-act.com*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need.

Advances in Neural Information Processing Systems (NeurIPS), 30, 5998--6008.

Author: Serge Destin TAMPOLLA is a 4IR Advisor specializing in digital transformation and emerging technologies of the Fourth Industrial Revolution. His expertise spans blockchain architecture, artificial intelligence systems, decentralized economics, and strategic consulting for cutting-edge technology projects. Contact: serge.destin@tampolla.com | www.tampolla.com

Funding: This research received no external funding. **Conflicts of Interest:** The author declares no conflicts of interest. **Data Availability:** Protocol specifications and smart contract pseudocode are available in the appendices of the associated whitepaper.