Whitepaper

Ailo Network: Decentralized AI Through Distributed Inference & Federated Learning

1. Abstract

Ailo Network introduces a revolutionary approach to AI by combining and directly in web browsers. Using Transformers.js and WebSocket coordination, Ailo enables multiple browsers to work together running AI models up to , far beyond what any single device could handle. This creates a truly decentralized AI infrastructure where users can both chat with AI models and contribute computing power to train them, earning rewards.

2. Introduction

2.1 The Problem

Modern Large Language Models require massive computational resources that only large corporations can afford. Running models like GPT-4, Llama 70B, or Grok 600B requires specialized datacenter infrastructure costing millions of dollars. This creates a dangerous centralization of AI power.

2.2 The Solution

Ailo Network democratizes AI access through two revolutionary approaches:

- Run smaller models (up to 500M parameters) entirely in your browser using Transformers.js
- Multiple browsers coordinate via WebSocket to run massive models (1.8B to 600B parameters) by sharing the computational load
- Train models across thousands of browsers while keeping data private, earning AiloCoin rewards

2.3 Current Capabilities

As of December 2024, Ailo Network supports:

- Solo models: Qwen 0.5B, GPT-2, DistilGPT-2 (instant, single browser)
- Distributed: Qwen 1.8B (4 nodes coordinated)
- Python CUDA Client for NVIDIA GPUs (10x-50x faster)
- Roadmap: Llama 70B (64 nodes), Llama 405B (256 nodes), Grok 600B (512 nodes)

2.1 Key Innovations

- Supports both Browser-Based (WebGL) and Native (CUDA) clients
- Privacy-preserving gradient averaging without sharing raw data

- Cryptocurrency mining that produces useful AI work
- Efficient handling of billion-parameter models in limited memory

3. Architecture

3.1 Neural Network Model

Ailo employs a Transformer decoder-only architecture, similar to GPT-2, implemented entirely in JavaScript. The model specifications are:

Parameter	Value
Total Parameters	~840 million (target 1.2B)
Transformer Layers	24
Embedding Dimension	1600
FFN Hidden Dimension	6400
Attention Heads	25
Vocabulary Size	50,257 (GPT-2)
Maximum Context	2048 tokens

3.2 Federated Learning Protocol

The training process follows the Federated Averaging (FedAvg) algorithm:

1. Clients download the current global model M_t

2. Each client k trains locally on data D_k for E epochs
3. Clients submit gradients □L_k to the coordinator
4. Coordinator computes weighted average of gradients
5. New global model $M_{t+1} = M_t - \eta \cdot \Sigma(w_k \cdot$ □$L_k)$

$$M_{t+1} = M_t - \eta \cdot \Sigma_k \, (n_k/n) \cdot \quad L_k$$

3.3 Weight Sharding

To handle 3.2 GB of model weights in browser memory constraints, Ailo implements a VirtualTensor system that splits weights into 35 binary chunks. Chunks are cached in IndexedDB after first download, enabling instant loading on subsequent visits.

4. Proof-of-Contribution (PoC)

Traditional Proof-of-Work cryptocurrencies waste enormous computational resources on hash puzzles with no practical output. Ailo introduces Proof-of-Contribution, where mining work directly improves the AI model.

4.1 Contribution Validation

The coordinator validates each contribution by:

- Verifying loss reduction on held-out validation set
- Checking gradient magnitude for statistical plausibility
- Validating digital signature from user wallet

- Rate limiting to prevent Sybil attacks

4.2 Reward Formula

Contributors receive AiloCoin proportional to their contribution quality:

$$Reward = BaseRate \times BatchSize \times (1 + LossBonus)$$

Where LossBonus increases for contributions that significantly reduce model loss.

4.3 Distributed Inference Rewards

In addition to training rewards, users earn AiloCoin by contributing computational power for distributed AI inference:

- Nodes that join distributed inference rooms process requests collaboratively
- per successful inference processed
- Qwen 1.8B (4 nodes), Llama-2 7B (12 nodes), Llama-2 70B (120 nodes)
- All participating nodes receive rewards, excluding the requestor

This creates a self-sustaining network where users are incentivized to keep their browsers active and contribute to the collective intelligence.

4.4 Token Economics

This design choice enables perpetual rewards for network contributors without arbitrary caps that would eventually halt incentives.

Minting Mechanism

New tokens are minted

to the Ailo Network:

- ~0.1 ALC per verified gradient submission
- Up to +0.05 ALC for fast processing (<5 seconds)
- 0.05 ALC per distributed inference processed

Economic Sustainability

Token issuance is algorithmically regulated and balanced by:

- 0.01 ALC fee on transfers to prevent spam
- A portion of fees is permanently burned, reducing circulating supply
- Only valid, verified contributions receive rewards
- Anti-Sybil protections prevent reward farming

This creates a balanced economy where token supply grows proportionally to actual network utility and value creation.

4.5 Transaction Verification & Rollback

Ailo Network implements a

where nodes collaborate to detect and reverse fraudulent transactions:

Client-Side Verification

Each node downloads and verifies the blockchain locally using IndexedDB storage. Verification checks include:

- Transaction type and amount validity
- Wallet address format and consistency
- Timestamp integrity (not future-dated)
- Transfer logic (sender ≠ recipient)

Node-Based Confirmation System

Ailo uses a mechanism where multiple independent nodes must verify a transaction before it is considered confirmed:

- 10 node verifications required for a transaction to be spendable
- Up to 10 confirmations are tracked per transaction
- When users view transactions in their wallet, their client automatically submits a verification
- Each node/wallet can only verify a transaction once, preventing Sybil attacks
- Transactions from NETWORK (mining, training rewards) are auto-confirmed and immediately spendable

$$\textit{Confirmation Status} = \textit{"PENDING" if confirmations} < 10\textit{, else "CONFIRMED"}$$

$$\textit{System TX (from NETWORK)} = \textit{Always "CONFIRMED"}$$

Consensus-Based Rollback

When a node detects a suspicious transaction, it reports to the network:

- If 51% of active nodes report the same transaction, consensus is reached
- A REVERSAL transaction is created, returning funds to the original sender
- Transactions are marked as REPORTED → DISPUTED → REVERSED
- System minting (training/inference rewards) cannot be rolled back

5. Privacy and Security

Ailo prioritizes user privacy through several mechanisms:

- Raw data never leaves the user's device
- Only aggregated updates contain user contributions
- Only a unique username is required to participate (no email needed)

6. Roadmap

Phase	Status	Description
Phase 1	□ Completed (Dec 2025)	Network Launch & Browser AI Chat (Qwen 0.5B, GPT-2)
Phase 2	□ Completed (Dec 2025)	Distributed Inference - Qwen 1.8B across 4 nodes
Phase 3	□ Q1 2026	Llama-2 70B distributed across 120 nodes

Phase 4	□ Q2 2026	Llama 3.1 405B distributed across 256 nodes
Phase 5	□ 2026+	Global scale: 512+ nodes, 600B+ parameter models

7. Founder

Ailo Network is the result of passion for decentralized technology and artificial intelligence.

Riccardo Sparacino

FOUNDER & LEAD SOFTWARE ENGINEER

Senior Software Engineer specializing in AI Agents, Parallel Computing, and Full Stack Development. Riccardo leads the architectural vision, bringing extensive experience from Onlinestore, Vodafone, and AiLo to create a truly distributed AI infrastructure.

8. Conclusion

Ailo Network represents a paradigm shift in AI development: from centralized corporate control to distributed community ownership. By aligning economic incentives (AiloCoin) with useful work (AI training), we create a sustainable ecosystem where everyone benefits from the collective intelligence.

Join the revolution. Train the future.

Whitepaper v1.0 | December 2025 | Ailo Network Team