NeuroShard: A Decentralized Architecture for Collective Intelligence

LZ

November, 2025

Abstract

We present **NeuroShard**, a decentralized architecture for building and operating Large Language Models (LLMs) as a **public good** rather than a corporate asset. Instead of training a frontier model inside a single data center and selling access to its outputs, NeuroShard turns a global network of heterogeneous devices, from gaming GPUs and laptops to data center servers, into one continuously evolving optimizer that collectively trains a shared model, **NeuroLLM**, from random initialization. At the core of the system is a **Dynamic Layer Pool** that decouples model architecture from physical infrastructure: model depth and capacity grow organically with available network resources, without fixed "model sizes" or synchronized upgrade phases.

To make this feasible in a trustless setting, we introduce **Proof of Neural Work (PoNW)**, where cryptographically verifiable neural network computation (forward and backward passes) serves directly as the network's consensus and reward mechanism. PoNW is complemented by **Byzantine-robust gradient aggregation** (Trimmed Mean, Krum, Coordinate-wise Median) to defend against poisoning, and by **Sharded Checkpointing** plus **Tensor Parallelism** to distribute model state and computation across many small devices. Together, these mechanisms form a protocol in which every useful gradient, every contributed sample, and every token generated pushes NeuroLLM forward - demonstrating that a competitive, community-owned LLM can be constructed and governed by its users, one gradient at a time.

Contents

1 Introduction

1.1 The Problem with Centralized AI

The current landscape of artificial intelligence is defined by a **small number of vertically integrated AI platforms** that own the models, the data, and the distribution channels. Large Language Models (LLMs) such as GPT-5, Claude, and LLaMA are developed, trained, and deployed by a handful of organizations that decide who is allowed to build on top of them and under what terms. This concentration of power is not just an economic issue; it is an architectural bottleneck that constrains what kinds of AI systems we can build.

- **Centralized Control:** Alignment, safety, and access policies are set unilaterally, with limited transparency or public oversight. A single TOS change can reshape what billions of users are allowed to ask or build.
- **Infrastructure Barriers:** Training frontier models requires tightly managed, billion-dollar clusters. This shuts out independent researchers, communities, and smaller organizations from ever creating models at comparable scale.
- **Compute Latency and Waste:** While hyperscale data centers consume enormous amounts of energy, the world's edge compute, consumer GPUs, idle CPUs, and phones, remains largely dark and unused for AI.
- **Data Privacy and Sovereignty:** Centralized training aggregates raw data into monolithic repositories, creating systemic privacy risk and making it difficult for communities or individuals to control how their data shapes the models that govern their lives.

1.2 NeuroShard: A Decentralized Solution

NeuroShard proposes a decentralized alternative: a protocol for **collaborative model creation** where the network, not any single operator, is the place where the model lives. Instead of distributing pre-trained weights from a centralized lab, NeuroShard coordinates the training of a new model, **NeuroLLM**, from scratch using the aggregated compute and data of participating nodes. Each node that joins strengthens both the model and the network, and each contribution is cryptographically accounted for.

- **Organic Scalability:** The model architecture is not fixed but elastic. Using a **Dynamic Layer Pool**, the model expands as new nodes join and contribute memory, decoupling model size from the limitations of any single device and allowing capacity to grow continuously rather than in discrete "7B/70B" jumps.
- **Verifiable Computation:** Through **Proof of Neural Work (PoNW)**, the network cryptographically validates that participants are performing useful training operations. Gradients are not just updates, they are on-chain proofs of work, solving the free-rider problem inherent in open distributed systems.
- **Byzantine Tolerance:** The training process is secured against malicious or irrational actors through robust statistics (e.g., Krum, Trimmed Mean, Coordinate-wise Median) and a fraud-proof slashing mechanism that makes poisoning attacks economically costly.
- **Economic Alignment:** The NEURO token aligns incentives, rewarding participants for verifiable contributions of compute (gradients), data, and uptime. In effect, the token is backed by the network's collective computational intelligence and the evolving capability of NeuroLLM itself.

1.3 Contributions

This paper presents the following contributions:

- **Dynamic Model Architecture:** A transformer implementation that scales depth and width based on real-time network topology via a distributed layer registry.
- **Decentralized Training Protocol:** A hybrid system combining gradient gossip and optional ring-allreduce primitives for efficient parameter synchronization over P2P networks.
- **Proof of Neural Work (PoNW):** A consensus algorithm that validates useful deep learning computations (forward/backward passes) as proof of work.
- **Byzantine-Robust Aggregation:** Implementation of statistical defenses including Trimmed Mean and Coordinate-wise Median to reject poisoned gradients.
- **Sharded State Management:** A distributed checkpointing system that shards model weights across the network, enabling the training of models larger than any single node's capacity.
- **Tensor Parallelism:** Integration of intra-layer model parallelism (column/row splitting) to distribute large weight matrices across consumer devices.

2 Related Work

2.1 Distributed Deep Learning

Scaling deep learning beyond a single device has historically followed three paradigms:

2.1.1 Data Parallelism and Federated Learning

Federated Learning (FL) [1] enables training across decentralized devices by keeping data local and aggregating gradients. However, FL assumes each client can hold the full model in memory and requires a central aggregator. NeuroShard extends FL by eliminating the central aggregator through gossip-based gradient sharing.

2.1.2 Pipeline Parallelism

Pipeline parallelism [2] partitions models vertically by layers. Systems like Petals [3] use this for distributed inference. NeuroShard goes further by enabling distributed *training*, not just inference.

2.1.3 Decentralized Training

Prior work on decentralized training (e.g., Hivemind [4]) focuses on coordinating training of existing model architectures. NeuroShard introduces a model designed from the ground up for decentralized training.

2.2 Decentralized AI Networks

2.2.1 Bittensor

Bittensor [5] creates a decentralized intelligence market where nodes are rewarded based on the quality of their outputs. However, Bittensor treats models as black boxes and focuses on output validation rather than collective model creation. NeuroShard creates the model itself through collective training.

2.2.2 Gensyn

Gensyn [6] focuses on verifiable compute for ML training. NeuroShard incorporates similar verification concepts through PoNW while also providing the model architecture and training protocol.

2.3 Key Differentiators

NeuroShard is unique in combining:

- A **custom model architecture** designed for decentralized training (NeuroLLM)
- **Gradient gossip** for decentralized SGD without central coordination
- **Training rewards** that economically incentivize compute and data contribution
- **From-scratch training** rather than distributing pre-existing models

3 NeuroLLM Architecture

3.1 Design Philosophy

NeuroLLM is not a fork or fine-tune of an existing model. It is a completely new transformer architecture designed with the following goals:

1. **Efficient Gradients:** Optimized for gradient compression and gossip transmission
2. **Stable Training:** Uses techniques that are robust to asynchronous, noisy gradient updates
3. **Scalable Architecture:** Can grow from 125M to 70B+ parameters as the network matures
4. **Privacy-Compatible:** Supports differential privacy in training data

3.2 Architecture Details

NeuroLLM uses a modern transformer architecture with the following components:

3.2.1 RMSNorm (Root Mean Square Normalization)

Unlike LayerNorm, RMSNorm [7] is more stable for distributed training and requires fewer parameters:

$$\text{RMSNorm}(x) = \frac{x}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 + \epsilon}} \cdot \gamma \tag{1}$$

3.2.2 Rotary Position Embeddings (RoPE)

RoPE [8] encodes positional information directly into attention computations, enabling:

- No fixed maximum sequence length
- Better extrapolation to longer contexts
- No learnable position parameters (reducing gradient noise)

3.2.3 Grouped Query Attention (GQA)

GQA [9] reduces memory usage by sharing key-value heads across multiple query heads:

- 12 query heads, 4 key-value heads (bootstrap phase)
- 3x reduction in KV cache size
- Minimal quality degradation

3.2.4 SwiGLU Activation

SwiGLU [10] provides better gradient flow than ReLU or GELU:

$$\text{SwiGLU}(x) = \text{Swish}(xW_{\text{gate}}) \otimes (xW_{\text{up}}) \tag{2}$$

3.3 Dynamic Scaling via Adaptive Architecture

Traditional distributed training relies on static model architectures (e.g., "GPT-3 175B has 96 layers × 12,288 dim"). NeuroShard introduces **Adaptive Architecture Scaling**, where both model width (hidden dimension) and depth (number of layers) adapt automatically to total network capacity.

3.3.1 The Scaling Algorithm

NeuroShard's architecture follows empirical scaling laws derived from GPT-3 [11] and Chinchilla [12]:

$$\text{Optimal Architecture} = f(\text{Total Network Memory}) \tag{3}$$

Where the function f balances width and depth according to:

$$\text{Width (hidden\_dim)} \propto M^{0.6} \tag{4}$$

$$\text{Depth (num\_layers)} \propto M^{0.4} \tag{5}$$

This ensures width grows faster than depth, which empirically produces more efficient models [13].

3.3.2 Automated Architecture Updates

The network automatically recalculates optimal architecture every 50 nodes:

1. **Calculate Total Capacity:** $M_{\text{total}} = \sum_{i=1}^{N} M_i$ where M_i is each node's memory
2. **Compute Optimal Architecture:** Use scaling formulas to determine best width × depth
3. **Check Improvement:** Only upgrade if new architecture is $\geq 1.3\times$ better
4. **Trigger Migration:** New nodes use new architecture; existing nodes migrate gradually

No Votes Required: Architecture scaling is fully automated based on network capacity, not governance votes.

3.4 Dynamic Scaling via Layer Pooling

The **Dynamic Layer Pool** manages layer assignment across nodes using the current architecture.

3.4.1 The Dynamic Layer Pool

The network maintains a DHT-based registry of model layers:

1. **Resource Advertisement:** Upon joining, a node announces its available VRAM and compute capacity.
2. **Layer Allocation:** The protocol assigns a specific range of transformer layers to the node. This assignment is dynamic; as nodes join or leave, layers are redistributed to ensure redundancy ($R \geq 2$ replicas per layer).
3. **Organic Growth:** The total depth of the model is a function of the aggregate network memory. There are no pre-set "sizes" (e.g., 7B, 70B); the model grows layer-by-layer as capacity permits.

$$\text{Model Depth} \approx \frac{\sum_{i=1}^{N} \text{AvailableMemory}_i \times \alpha}{\text{Memory per Layer}} \tag{6}$$

where α is a utilization factor (e.g., 0.8).

This architecture allows for **Pipeline Parallelism** by default: inference and training forward passes traverse the network, routing activations from one node's assigned layers to the next.

3.4.2 Contribution-Based Rewards

Nodes earn NEURO proportional to their contribution:

$$R_{\text{multiplier}} = 1.0 + \frac{\text{Layers Held}}{\text{Total Network Layers}} \times (1 + 0.1 \cdot \mathbb{1}_{\text{embed}} + 0.1 \cdot \mathbb{1}_{\text{head}}) \tag{7}$$

Where $\mathbb{1}_{\text{embed}}$ and $\mathbb{1}_{\text{head}}$ are indicator functions for holding the embedding layer and LM head respectively.

3.4.3 Example: Network Growth

Table 1: Dynamic Model Growth Example (Width × Depth)

Network State	**Total Memory**	**Architecture**	**Params**	**Comparable To**
1 node (2GB)	2GB	8L × 512H	50M	GPT-2 Small
10 nodes (4GB avg)	40GB	16L × 1024H	350M	GPT-2 Large
50 nodes (6GB avg)	300GB	24L × 2048H	2.7B	GPT-3 Small
100 nodes (8GB avg)	800GB	32L × 3072H	9.2B	GPT-3 Medium
500 nodes (8GB avg)	4TB	48L × 5120H	47B	LLaMA 2 70B
1000 nodes (8GB avg)	8TB	64L × 7168H	123B	GPT-4 class

Key Insight: The model grows in BOTH width (hidden dimension) and depth (layers), following empirical scaling laws from GPT-3 and Chinchilla research. This produces efficient, well-architected models at any scale. No fixed dimensions - purely capacity-driven.

3.5 Forward Pass

The NeuroLLM forward pass follows standard transformer architecture:

Algorithm 1 NeuroLLM Forward Pass

```
1:  procedure FORWARD($input\_ids$)
2:      $x \leftarrow \text{TokenEmbedding}(input\_ids)$
3:      for $layer \in [0, \text{num\_layers})$ do
4:          $x \leftarrow x + \text{Attention}(\text{RMSNorm}(x))$          ▷ Pre-norm
5:          $x \leftarrow x + \text{SwiGLU}(\text{RMSNorm}(x))$
6:      end for
7:      $x \leftarrow \text{RMSNorm}(x)$
8:      $logits \leftarrow \text{LinearProjection}(x)$
9:      return $logits$
10: end procedure
```

3.6 Memory Requirements - Dynamic Architecture

NeuroLLM's architecture adapts to network capacity. Each node contributes based on available memory AND current network size:

Table 2: Layer Capacity (varies by network architecture)

Available RAM	**Network: 10 nodes**	**Network: 100 nodes**	**Network: 1000 nodes**
	Architecture adapts: 16L×1024H → 32L×3072H → 64L×7168H		
1GB	2-3 layers (30M)	1-2 layers (55M)	1 layer (120M)
2GB	4-6 layers (60M)	3-4 layers (110M)	2-3 layers (240M)
4GB	10-12 layers (120M)	6-8 layers (220M)	4-6 layers (480M)
8GB	20-24 layers (250M)	12-16 layers (440M)	8-10 layers (960M)
16GB	Full model (350M)	24-32 layers (880M)	16-20 layers (1.9B)

Key Principle: Architecture scales with total network capacity. Early nodes hold entire small model. As network grows, model becomes wider and deeper, with layers distributed across nodes. Every device participates optimally.

4 Decentralized Training

4.1 Training Overview

NeuroLLM is trained through a decentralized process where:

1. Nodes contribute training data locally
2. Each node computes gradients on its local data
3. Gradients are compressed and shared via P2P gossip
4. A training coordinator aggregates gradients and applies updates
5. NEURO rewards are distributed to contributors

4.2 Genesis Dataset and Deterministic Sharding

To solve the "garbage in, garbage out" problem inherent in user-submitted data, NeuroShard employs a **Genesis Dataset Strategy**. Instead of users submitting arbitrary text, the training data is drawn from a cryptographically verified manifest of high-quality, open-source datasets (e.g., FineWeb, RedPajama).

4.2.1 The "Driver" Architecture

Data is introduced into the network by specific nodes known as **Drivers** (those holding Layer 0). The dataset is sharded deterministically:

$$\text{Assigned Shard ID} = \text{Hash}(\text{Node ID}) \pmod{\text{Total Shards}} \tag{8}$$

This deterministic assignment enables **Proof of Neural Work Verification**:

- A peer can verify a Driver's work by downloading the same shard and re-computing the first layer.
- If a Driver sends garbage or modified data, the cryptographic hash of their output will mismatch the honest computation.
- This forces Drivers to actually download and process the real training data to earn rewards.

4.2.2 The Role-Based Pipeline

The network self-organizes into three distinct roles based on hardware capacity **and stake**:

1. **Drivers (Layer 0):** High-bandwidth nodes that download Genesis Shards and initiate the forward pass. They attach ground-truth labels to the activation packet.
2. **Workers (Layers 1..N):** Standard nodes that receive activations, perform matrix multiplications, and forward to the next peer. They do not need to store the dataset.
3. **Validators (Last Layer + Stake):** High-compute nodes with **minimum 100 NEURO staked** that calculate the Loss between the final output and the labels provided by the Driver. They initiate the backward pass **and validate PoNW proofs from other nodes**.

This architecture ensures that massive datasets (terabytes) can be used to train the model without requiring every participant to download them. Only the Drivers pay the storage/bandwidth cost, for which they are compensated with a 1.2x reward multiplier.

Important: The Validator role in NeuroShard is a **hybrid** of Model Validation (computing loss on the last layer) and Consensus Validation (verifying PoNW proofs). This dual responsibility requires both computational capacity (memory for the LM head) and economic stake (100 NEURO minimum). See Section 7 for details.

4.3 Gradient Gossip Protocol

Nodes share gradients through a gossip protocol that follows the Driver-Worker-Validator pipeline:

Algorithm 2 Role-Based Gradient Gossip

```
1:  procedure ComputeAndShareGradients
2:      if Role = Driver then
3:          $batch, labels \leftarrow$ GenesisLoader.get_batch()
4:          $activations \leftarrow$ Forward($batch$)
5:          SendToPipeline($activations, labels$)
6:      else if Role = Worker then
7:          $input, labels \leftarrow$ ReceiveFromPeer()
8:          $output \leftarrow$ Forward($input$)
9:          SendToPipeline($output, labels$)
10:     else if Role = Validator then
11:         $final, labels \leftarrow$ ReceiveFromPeer()
12:         $loss \leftarrow$ CrossEntropy($final, labels$)
13:         Backward($loss$)
14:         GossipGradients()
15:     end if
16: end procedure
```

4.3.1 Gradient Compression

To minimize P2P bandwidth usage, NeuroShard implements a multi-stage compression pipeline for gradient updates:

1. **Top-K Sparsification:** Only the largest k% of gradient elements (by magnitude) are transmitted.
2. **INT8 Quantization:** Selected values are quantized to 8-bit integers.
3. **Entropy Coding:** The sparse, quantized stream is further compressed using zlib.
4. **Error Feedback:** Residuals (discarded values) are accumulated locally and added to the next step's gradient, ensuring convergence.

This pipeline achieves compression ratios of 50-100x with minimal impact on model convergence.

4.4 Training Coordinator

Each node runs a Training Coordinator that:

1. **Initiates Rounds:** Every 60 seconds, starts a new training round
2. **Collects Contributions:** Receives gradient contributions from peers
3. **Aggregates Gradients:** Uses weighted averaging based on batch size
4. **Applies Updates:** Performs gradient descent with clipping
5. **Distributes Rewards:** Calculates and distributes NEURO tokens

Algorithm 3 Training Round

```
procedure CompleteRound($contributions$)
    if $|contributions| <$ MIN_CONTRIBUTIONS then
        return                                          ▷ Not enough participants
    end if
    $aggregated \leftarrow$ WeightedAverage($contributions$)
    for $name, grad \in aggregated$ do
        $param \leftarrow$ model.get($name$)
        $param.grad \leftarrow grad$
    end for
    ClipGradNorm(1.0)
    optimizer.step()
    $rewards \leftarrow$ CalculateRewards($contributions$)
    DistributeNEURO($rewards$)
end procedure
```

4.5 Training Rewards

Training is the **dominant** reward activity in NeuroShard. Contributors earn NEURO tokens based on their role and contribution:

$$R_{\text{training}} = R_{\text{batch}} \times B_{\text{count}} \times M_{\text{role}} \times M_{\text{training}} \tag{9}$$

Where:

- $R_{\text{batch}} = 0.0005$ NEURO per training batch
- B_{count} = number of training batches completed
- $M_{\text{training}} = 1.1$ (10% training bonus)

Training earns approximately **300× more** than idle uptime, ensuring only active contributors are rewarded meaningfully.

Role multipliers M_{role}:

- **Drivers (Layer 0):** 1.2x (compensates for data bandwidth, initiates forward pass)
- **Validators (Last Layer):** 1.3x (compensates for loss computation, initiates backward pass, proof validation)
- **Workers (Middle Layers):** 1.0x base + 5% per layer held (max 100% bonus)

Example: A node holding 25 layers, actively training at 60 batches/minute:

- Base: $0.0005 \times 60 = 0.03$ NEURO/min
- Layer bonus: $1.0 + (25 \times 0.05) = 2.25$x
- Training bonus: 1.1x
- Total: $0.03 \times 2.25 \times 1.1 = 0.074$ NEURO/min ($\sim$107 NEURO/day)

This incentivizes nodes with higher capacity (RAM/Bandwidth) to take on the critical "Driver" and "Validator" roles, ensuring the network has enough throughput to process the massive Genesis Dataset.

5 System Architecture

5.1 High-Level Overview

NeuroShard consists of three main components:

1. **NeuroNode:** The core node that runs NeuroLLM, handles training and inference
2. **P2P Network:** Kademlia DHT for peer discovery, gossip for gradient sharing
3. **NEURO Economy:** Token system for rewards and payments

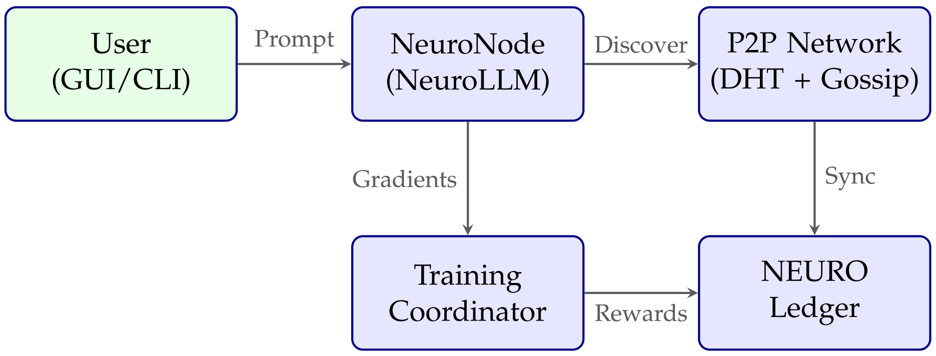

Figure 1: NeuroShard Architecture: Users interact with NeuroNode, which participates in P2P network for peer discovery and gradient gossip. Training Coordinator manages distributed training, and NEURO Ledger tracks token balances.

5.2 DynamicNeuroNode

The DynamicNeuroNode is the core component that:

- **Detects Memory:** Automatically detects available system memory
- **Registers with Network:** Announces capacity and receives layer assignments
- **Loads Assigned Layers:** Only loads the layers it's responsible for
- **Handles Inference:** Routes requests through the network pipeline
- **Participates in Training:** Computes gradients for its layers, shares via gossip
- **Earns NEURO:** Rewards proportional to layers held and compute contributed

```python
def create_dynamic_node(
    node_token: str,
    port: int = 8000,
    tracker_url: str = "https://neuroshard.com/api/tracker",
    available_memory_mb: Optional[float] = None  # Auto-detect if not
        provided
) -> DynamicNeuroNode:
    # Derive node ID from token
    node_id = sha256(node_token).hexdigest()

    # Create node (auto-detects memory)
    node = DynamicNeuroNode(
        node_id=node_id,
        port=port,
        available_memory_mb=available_memory_mb
    )

    # Start node:
    # 1. Register with network
    # 2. Get layer assignments based on memory
    # 3. Load only assigned layers
    node.start()

    # Node is now ready:
    # - Holding N layers based on memory
    # - Contributing to distributed model
    # - Earning NEURO proportional to contribution

    return node
```

Listing 1: DynamicNeuroNode Initialization

5.3 P2P Network Layer

5.3.1 Peer Discovery

NeuroShard uses a hybrid discovery mechanism:

1. **Tracker Bootstrap:** Initial peer discovery via central tracker
2. **Kademlia DHT:** Decentralized discovery after bootstrap
3. **Gossip Protocol:** Peer list exchange for redundancy

5.3.2 Gradient Gossip

Gradients are shared via gossip:

1. Node computes local gradients
2. Compresses gradients (Top-K + INT8 + Zlib)
3. Gossips to $k = 3$ random peers
4. Receiving peers validate and aggregate

5.4 Inference Protocol

Users can generate text through NeuroLLM:

Algorithm 4 Text Generation

```
1: procedure GENERATE($prompt, max\_tokens$)
2:     $tokens \leftarrow$ Tokenize$(prompt)$
3:     $input\_ids \leftarrow$ ToTensor$(tokens)$
4:     for $i \in [0, max\_tokens)$ do
5:         $logits \leftarrow$ NeuroLLM.forward$(input\_ids)$
6:         $next\_token \leftarrow$ Sample$(logits[-1])$
7:         $input\_ids \leftarrow$ Concat$(input\_ids, next\_token)$
8:         if $next\_token =$ EOS then
9:             break
10:        end if
11:    end for
12:    return Detokenize$(input\_ids)$
13: end procedure
```

Note: In the bootstrap phase, output quality will be low (the model is untrained). Quality improves as the network trains the model.

6 Proof of Neural Work (PoNW)

6.1 Overview

Proof of Neural Work (PoNW) is the foundational consensus mechanism of NeuroShard. Unlike traditional blockchain consensus:

- **Proof of Work (PoW):** Rewards arbitrary computation (SHA-256 hashing) that produces no useful output.
- **Proof of Stake (PoS):** Rewards capital lockup without requiring any computation.
- **Proof of Neural Work (PoNW):** Rewards *useful neural network computation* that directly improves NeuroLLM.

The key insight is that **training a neural network is inherently verifiable work**. A gradient computed on real data will, when aggregated with other gradients, reduce the model's loss. Fake or random gradients will not. This creates a natural proof mechanism.

6.2 PoNW as the Model Builder

PoNW is not just a reward mechanism - it is the **engine that builds NeuroLLM**. Every NEURO token minted represents a real contribution to the model's intelligence:

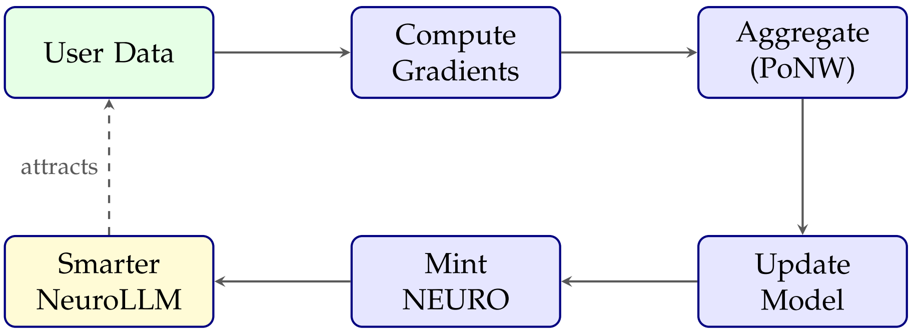

Figure 2: PoNW Cycle: User data → Gradients → Aggregation → Model Update → NEURO Rewards → Smarter Model → More Users

This creates a **virtuous cycle**:

1. Users contribute data and compute
2. PoNW validates contributions and mints NEURO
3. Model improves with each training round
4. Better model attracts more users
5. More users = more data = faster improvement

6.3 PoNW Components

PoNW validates three types of work:

1. **Proof of Inference:** Tokens processed during text generation. Each forward pass through NeuroLLM is counted.
2. **Proof of Training:** Gradient contributions to model improvement. The gradient's impact on loss reduction is measured.
3. **Proof of Data:** Training samples contributed. Data quality is assessed by the resulting gradient quality.

6.4 The Training-Mining Equivalence

In NeuroShard, **training IS mining**. Traditional cryptocurrency mining wastes energy on arbitrary computation. NeuroShard redirects that energy into building collective intelligence:

Table 3: Mining Comparison

Aspect	**Bitcoin (PoW)**	**NeuroShard (PoNW)**
Work Type	SHA-256 hashing	Neural network training
Output	Block hash	Model improvement
Useful?	No (arbitrary)	Yes (builds AI)
Verification	Hash check	Loss reduction
Energy Use	Wasted	Productive

Every NEURO token represents real intelligence added to NeuroLLM. The token's value is backed by the model's capability.

6.5 Reward Function

The total NEURO reward for a Proof of Neural Work is calculated as:

$$R_{\text{total}} = \left(R_{\text{uptime}} + R_{\text{inference}} + R_{\text{training}} + R_{\text{data}}\right) \cdot M_{\text{role}} \cdot M_{\text{stake}} \cdot M_{\text{training}} \tag{10}$$

Where the base rewards are:

- $R_{\text{uptime}} = 0.0001 \cdot \frac{T_{\text{seconds}}}{60}$ NEURO (uptime, minimal)
- $R_{\text{inference}} = 0.1 \cdot \frac{T_{\text{tokens}}}{10^6} \cdot S_{\text{role}}$ NEURO (role-weighted inference)
- $R_{\text{training}} = 0.0005 \cdot B_{\text{batches}}$ NEURO (training batches, dominant)
- $R_{\text{data}} = 0.00001 \cdot D_{\text{samples}}$ NEURO (data serving)

And the multipliers are:

- $M_{\text{role}} = 1.0 + \text{LayerBonus} + \text{DriverBonus} + \text{ValidatorBonus}$
- $M_{\text{stake}} = 1.0 + 0.1 \cdot \log_2\left(1 + \frac{S_{\text{stake}}}{1000}\right)$ (diminishing returns, see Section 7.4)
- $M_{\text{training}} = 1.1$ if actively training, else 1.0

The role share S_{role} for inference is:

- Driver (has embedding): $S_{\text{role}} = 0.15$
- Worker (middle layers): $S_{\text{role}} = 0.70$
- Validator (has LM head): $S_{\text{role}} = 0.15$

6.6 Proof Generation

Nodes generate PoNW proofs every 60 seconds:

```python
def get_ponw_proof(self) -> Dict:
    proof = {
        "node_id": self.node_id,
        "timestamp": time.time(),
        "tokens_processed": self.total_tokens_processed,
        "training_rounds": self.total_training_rounds,
        "layers_held": len(self.my_layer_ids),
        "reward_multiplier": self.calculate_reward_multiplier(),
    }

    # Sign the proof
    proof_string = f"{self.node_id}:{proof['timestamp']}:{proof['
        layers_held']}"
    proof["signature"] = sha256(proof_string).hexdigest()

    return proof
```

Listing 2: PoNW Proof Generation

6.7 Proof Validation

Receiving nodes validate proofs by checking:

1. **Signature:** Cryptographically valid
2. **Timestamp:** Within 5-minute window (prevents replay)
3. **Non-duplicate:** Not already processed
4. **Plausibility:** Token counts are reasonable for the time period

6.8 Gradient Verification

The core innovation of PoNW is **gradient verification**. When a node submits a gradient contribution, it can be verified through:

6.8.1 Statistical Verification

Valid gradients have predictable statistical properties:

- **Distribution:** Real gradients follow a roughly Gaussian distribution
- **Magnitude:** Gradient norms fall within expected ranges based on batch size
- **Sparsity:** After Top-K compression, valid gradients have specific sparsity patterns

6.8.2 Loss-Based Verification

The ultimate test of a gradient's validity is whether it reduces loss:

$$\mathcal{L}(w - \eta \cdot g) < \mathcal{L}(w) \tag{11}$$

Where w is the current weights, η is the learning rate, and g is the submitted gradient. If this inequality holds on a validation set, the gradient is valid.

6.8.3 Cross-Validation

Multiple nodes computing gradients on similar data should produce similar gradients. Outliers are flagged:

$$\text{score}(g_i) = \frac{\|g_i - \bar{g}\|}{\sigma_g} \tag{12}$$

Gradients with score > 3 (more than 3 standard deviations from mean) are rejected.

6.9 Attack Prevention

PoNW prevents common attacks through a multi-layered security model:

- **Freeloading:** Cannot earn rewards without running the model and computing real gradients
- **Inflation:** Token counts are cross-validated against actual computation time
- **Sybil:** Staking requirement (1000 NEURO) makes fake nodes expensive
- **Gradient Poisoning:** Robust aggregation and statistical verification reject malicious gradients
- **Replay Attacks:** Timestamp windows and signature uniqueness prevent proof reuse

6.10 Cryptographic Security Model

Every PoNW proof is cryptographically secured using **ECDSA (secp256k1)**:

1. **BIP39 Wallet:** Users generate a 12-word mnemonic seed phrase (similar to MetaMask). The node token is derived as $\text{token} = \text{BIP39\_seed}(\text{mnemonic})[:32]$. Private keys are never stored—only the user controls their wallet.
2. **Key Derivation:** $\text{private\_key} = \text{SHA256}(\text{node\_token})$ (32 bytes)
3. **Public Key:** $\text{public\_key} = \text{ECDSA.derive}(\text{private\_key})$ (33 bytes compressed)
4. **Node Identity:** $\text{node\_id} = \text{SHA256}(\text{public\_key})[:32]$ (deterministic from public key)
5. **Proof Signature:** $\text{sig} = \text{ECDSA.sign}(\text{private\_key}, \text{canonical\_payload})$
6. **Trustless Verification:** Anyone can verify signatures using only the public key
7. **Timestamp Freshness:** Proofs must be < 5 minutes old
8. **Replay Prevention:** Each signature stored in `proof_history`, can never be reused
9. **Rate Limiting:** Max 120 proofs/hour, max 1M tokens/minute per node

Why ECDSA over HMAC? HMAC requires sharing the secret token to verify signatures. ECDSA allows **trustless verification** - any node can verify a proof's authenticity using only the signer's public key, without needing their secret token. This is essential for decentralized consensus.

The canonical payload format ensures deterministic verification:

```
payload = f"{node_id}:{proof_type}:{timestamp}:{nonce}:
           {uptime}:{tokens}:{batches}:{samples}:
           {layers}:{has_embed}:{has_head}"
```

Transparency Guarantee: There is NO admin backdoor. The ONLY way to get NEURO is to run a node, do real work, create a signed proof, and pass ALL verification checks. Even the project creators must run nodes and earn like everyone else.

7 Hybrid Validator System

NeuroShard introduces a novel **Hybrid Validator** model that combines the best aspects of Proof of Work (computational contribution) and Proof of Stake (economic commitment) into a unified validation system. This hybrid approach solves critical problems that neither mechanism can address alone.

7.1 The Problem with Pure Approaches

7.1.1 Pure Proof of Work Limitations

In a pure PoW system for neural networks:

- **No Economic Stake:** Malicious nodes can submit poisoned gradients with no financial risk
- **Sybil Vulnerability:** Attackers can spin up many nodes cheaply
- **No Accountability:** Bad actors can simply restart with new identities

7.1.2 Pure Proof of Stake Limitations

In a pure PoS system:

- **Rich-Get-Richer:** Large stakeholders dominate rewards exponentially
- **No Work Required:** Validators can earn without contributing compute
- **Centralization:** Wealth concentration leads to validator monopolies

7.2 The Hybrid Solution

NeuroShard's Hybrid Validator combines **Model Validation** (layer-based work) with **Consensus Validation** (stake-based verification):

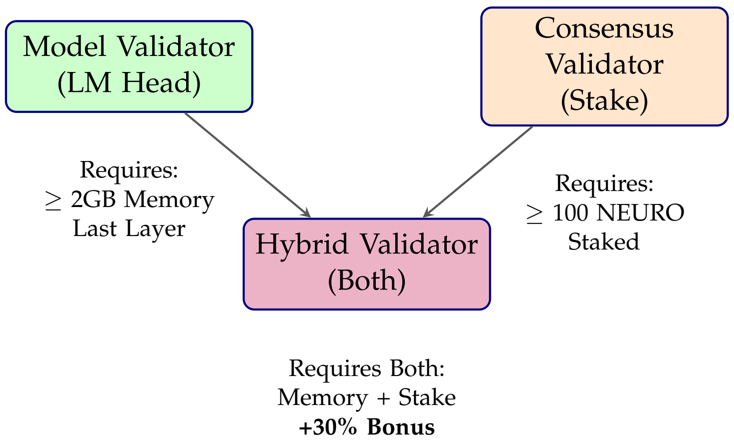
Model Validator (LM Head)
Consensus Validator (Stake)
Requires: ≥ 2GB Memory Last Layer
Requires: ≥ 100 NEURO Staked
Hybrid Validator (Both)
Requires Both: Memory + Stake
+30% Bonus

Figure 3: Hybrid Validator combines Model Validation (compute) with Consensus Validation (stake)

7.2.1 Validator Requirements

To become a Hybrid Validator, a node must meet **both** requirements:

Table 4: Hybrid Validator Requirements

Requirement	Type	Purpose
$\geq$ 2GB Memory	Compute	Hold LM Head layer
$\geq$ 100 NEURO Staked	Economic	Skin in the game
Last Layer Assignment	Network	Model validation role

7.2.2 Dual Responsibilities

Hybrid Validators perform two critical functions:

1. **Model Validation:** Compute loss on the final layer, initiate backward pass, verify gradient quality
2. **Consensus Validation:** Verify PoNW proofs from other nodes, cast stake-weighted votes, participate in consensus

7.3 Stake-Weighted Proof Validation

When a node submits a PoNW proof, Hybrid Validators verify it through a stake-weighted voting system:

Algorithm 5 Stake-Weighted Proof Validation

```
1:  procedure ValidateProof($proof$)
2:      $validators \leftarrow$ SelectValidators($proof, n = 3$)        ▷ Stake-weighted random
3:      for $v \in validators$ do
4:          $vote_v \leftarrow v$.ValidateLocally($proof$)        ▷ Check plausibility
5:          GossipVote($proof, vote_v, v.stake$)
6:      end for
7:      $valid\_stake \leftarrow \sum_{v:vote_v=\text{true}} v.stake$
8:      $total\_stake \leftarrow \sum_v v.stake$
9:      if $\frac{valid\_stake}{total\_stake} \geq 0.66$ then        ▷ 66% threshold
10:         ProcessProof($proof$)        ▷ Credit rewards
11:     else
12:         RejectProof($proof$)
13:     end if
14: end procedure
```

7.3.1 Validator Selection

Validators are selected using a **stake-weighted random** algorithm that balances fairness with security:

$$\text{Score}_i = \text{Stake}_i \times (1 - r) + \text{Random}_i \times r \times \text{MaxStake} \tag{13}$$

Where $r = 0.3$ (30% randomness factor). This ensures:

- Higher stake = higher chance of selection (security)
- Randomness prevents stake monopolies (fairness)
- Small stakers still get validation opportunities (decentralization)

7.4 Diminishing Stake Returns

To prevent the "rich-get-richer" problem inherent in pure PoS systems, NeuroShard implements **logarithmic diminishing returns** on stake multipliers:

$$M_{\text{stake}} = 1.0 + 0.1 \times \log_2 \left(1 + \frac{S_{\text{stake}}}{1000}\right) \tag{14}$$

Table 5: Diminishing Returns vs Linear Staking

Stake (NEURO)	Linear (Old)	Diminishing (New)	Reduction
1,000	1.10x	1.10x	0%
2,000	1.20x	1.16x	3%
5,000	1.50x	1.26x	16%
10,000	2.00x	1.35x	33%
50,000	6.00x	1.56x	74%
100,000	11.00x	1.66x	85%

Key Insight: Under the old linear system, a whale with 100,000 NEURO would earn 11x rewards. Under diminishing returns, they earn only 1.66x. This dramatically reduces wealth concentration while still rewarding commitment.

7.5 Validator Economics

7.5.1 Validator Rewards

Hybrid Validators earn from multiple sources:

Table 6: Validator Reward Sources

Source	**Description**	**Rate**
Model Validation	LM Head computation	+30% bonus
Proof Validation	Per proof validated	0.001 NEURO
Stake Multiplier	Diminishing returns	up to 1.66x
Training Bonus	If actively training	+10%

7.5.2 Validator Slashing

Validators who vote against consensus are **slashed at 2x the normal rate**:

$$\text{Slash}_{\text{validator}} = \text{Slash}_{\text{base}} \times 2.0 = 20\ \text{NEURO} \tag{15}$$

The slashed amount is burned, creating deflationary pressure. This ensures:

- Validators have strong incentive to vote honestly
- Bad validators lose their stake (can no longer validate)
- The cost of attacking consensus scales with stake

7.6 Why Hybrid Validators Work

The Hybrid Validator model solves the fundamental tension between decentralization and security:

1. **Work Dominates:** You cannot earn significant rewards without actual computation (training, inference). Stake only provides a multiplier, not base rewards.
2. **Stake Provides Accountability:** Validators have economic skin in the game. Cheating costs real money.
3. **Diminishing Returns Prevent Monopolies:** Whales cannot dominate through pure capital. A node with 100x more stake only gets 1.5x more rewards.
4. **Low Barrier to Entry:** Only 100 NEURO required to become a validator. This is earnable in $\sim$1 week of active participation.
5. **Random Selection Ensures Fairness:** 30% randomness means small validators still get opportunities to validate and earn fees.

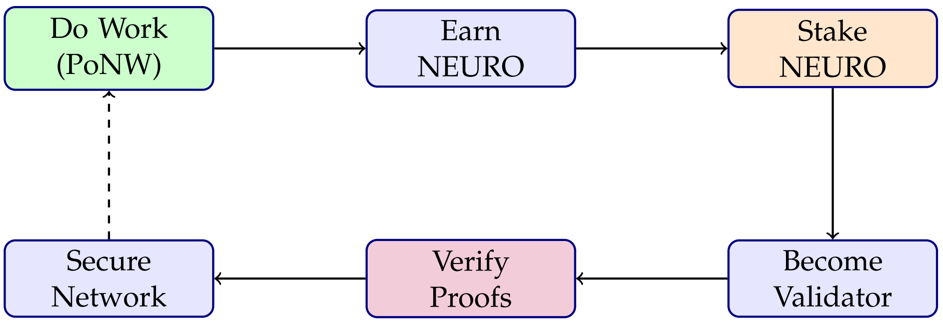

Figure 4: The Hybrid Validator Virtuous Cycle: Work $\rightarrow$ Earn $\rightarrow$ Stake $\rightarrow$ Validate $\rightarrow$ Secure $\rightarrow$ More Work

The Result: A system where validators are both computationally invested (they run the model) and economically invested (they stake tokens). This dual commitment creates the strongest possible incentive alignment for honest behavior.

8 Robustness and Anti-Poisoning

8.1 The Model Poisoning Threat

In a decentralized training environment, malicious actors may attempt to "poison" the model by submitting gradients that degrade performance or inject hidden backdoors (triggers). NeuroShard employs a multi-layered defense strategy.

8.2 Robust Aggregation

To secure the global model against adversarial attacks ("poisoning"), the Training Coordinator employs **Robust Aggregation** algorithms rather than simple averaging.

- **Trimmed Mean:** For each parameter coordinate, the highest and lowest k% of values are discarded before averaging. This neutralizes outliers.
- **Coordinate-wise Median:** The median value of gradients is used, which theoretically tolerates up to 50% malicious participants.
- **Krum/Bulyan:** Advanced selection algorithms that identify the gradient vector closest to the spatial center of the distribution.

8.3 The Fraud Proof System

To address the "Lazy Verifier" problem and ensure economic security, NeuroShard implements a challenge-response system:

- **Evidence Submission:** If a node detects a statistically anomalous gradient (e.g., $> 5\sigma$ deviation), it generates a cryptographic **Fraud Proof**.
- **Verification:** The proof, containing the signed malicious gradient and statistical context, is broadcast to the network.
- **Slashing:** Upon verification, the malicious node's staked NEURO is slashed.

9 Governance and The NeuroDAO

NeuroLLM is not a static artifact; it is a living system that must evolve. Governance is handled by the **NeuroDAO**.

9.1 What NeuroDAO Does NOT Control

Unlike traditional systems, NeuroDAO does **not** vote on:

- **Model Size:** This is determined automatically by network capacity
- **Phase Transitions:** There are no phases - the model grows organically
- **Who Can Participate:** Anyone with compute can join

9.2 What NeuroDAO Controls

NEURO token holders vote on:

- **Reward Rates:** Adjusting the $R_{compute}$ and R_{data} multipliers to balance supply and demand.
- **Slashing Conditions:** Defining what constitutes malicious behavior and the penalties.
- **Protocol Parameters:** Minimum stake, gradient compression ratio, training round duration.

9.3 Model Architecture Governance

The community decides on the "brain structure" of NeuroLLM:

- **Tokenizer Updates:** Voting to add new tokens to the vocabulary (e.g., for new languages or coding frameworks).
- **Architecture Changes:** Proposals to switch from RoPE to ALiBi, or introduce Mixture-of-Experts layers, are submitted as Neuro Improvement Proposals (NIPs).
- **Layer Architecture:** Changes to attention mechanism, FFN structure, or normalization.

10 NEURO Token Economics

10.1 Token Utility

The NEURO token serves four functions:

1. **Reward Currency:** Earned by contributing compute and data
2. **Payment Currency:** Spent to access NeuroLLM inference
3. **Staking Security:** Staked to earn reward multipliers (with diminishing returns)
4. **Validator Eligibility:** Minimum 100 NEURO stake required to become a Hybrid Validator

10.2 Genesis Block and Transparency

NeuroShard's ledger is initialized with a **Genesis Block** that guarantees zero pre-mine:

- **Zero Pre-Mine:** The ledger starts with `total_minted = 0.0`
- **No Founder Allocation:** There are no pre-allocated tokens for founders or investors
- **No ICO:** All NEURO must be earned through verified Proof of Neural Work
- **Audit Trail:** Every token minted is traceable to a specific PoNW proof

The Genesis Block is recorded in the `proof_history` table with signature `GENESIS_BLOCK` and reward amount `0.0`. Anyone can verify this by querying:

```
SELECT * FROM global_stats WHERE id = 1;
-- Returns: total_minted=0.0, total_burned=0.0 (at genesis)
```

10.3 Deflationary Mechanics

To ensure long-term value accrual, NeuroShard implements a **Burn Mechanism**:

- **Fee Burn:** 5% of all spending (inference, transfers) is permanently burned
- **Fraud Slashing:** 50% of slashed stakes from malicious nodes are burned (50% to whistle-blower)
- **Validator Slashing:** 100% of validator slashes are burned (validators slashed at 2x rate for voting against consensus)
- **Burn Address:** `BURN_0x0000...` - tokens sent here are irrecoverable

As network usage grows, the supply of NEURO decreases, rewarding long-term holders and contributors.

10.4 Earning NEURO

Users earn NEURO through a hierarchical reward structure that prioritizes **Training** as the core value:

Table 7: NEURO Earning Mechanisms (Reward Hierarchy)

Activity	Description	Rate	Daily (Active)
Training	Contributing gradients	0.0005 NEURO/batch	∼43 NEURO
Inference	Processing tokens	Market-based (pure supply/demand)	Variable
Data	Serving training shards	0.00001 NEURO/sample	Variable
Uptime	Running a node (idle)	0.0001 NEURO/min	∼0.14 NEURO

Note on Dynamic Pricing: Inference rewards use a **pure market** where price is determined by supply and demand. When the model is worthless (bootstrap), demand is zero and price approaches zero. As the model improves and attracts users, demand rises naturally and price increases. The market self-regulates: high prices attract more nodes to inference, while low prices push nodes back to training. Quality emerges from the demand signal itself.

10.4.1 Privacy-Preserving Distributed Inference Marketplace

NeuroShard implements a **request-response marketplace** that orchestrates distributed inference while preserving user privacy.

Architecture:

Inference in NeuroShard is inherently distributed across the pipeline:

1. **User** submits request to marketplace (metadata only - NO prompt!)
2. **Driver node** (Layer 0) receives encrypted prompt directly from user
3. **Worker nodes** (Layers 1-N) process activations (never see prompt)
4. **Validator node** (LM Head) generates output, returns to user

Privacy Guarantee: The prompt is NEVER stored in the marketplace. Users send encrypted prompts directly to their chosen driver node via a private channel. Worker nodes only process activations (meaningless vector representations), ensuring they cannot reconstruct the original prompt. This preserves privacy while enabling distributed computation.

Pure Market Pricing (No Artificial Caps):

$$P_{\text{market}} = P_{\text{base}} \times (1+u)^2 \tag{16}$$

Where:

- $P_{\text{base}} = 0.0001$ NEURO (starting price)
- $u =$ demand rate/supply rate (utilization)
- NO minimum or maximum caps - market finds true value

Request-Response Matching:

To prevent timing attacks, prices are **locked at submission time**:

1. User submits request $\rightarrow$ Price locked at current market rate
2. Driver claims request $\rightarrow$ Starts distributed pipeline
3. All nodes submit proofs $\rightarrow$ Each rewarded by role
4. Request marked complete $\rightarrow$ All participants paid at locked price

This ensures users always pay the price they saw at submission, even if market price spikes during processing.

Distributed Reward Distribution:

Multiple nodes participate in each inference request. The locked price determines the total reward pool, distributed by role:

- **Driver (Layer 0):** 15% of pool (processes embedding, sees prompt)
- **Workers (Layers 1-N):** 70% of pool (divided among participants)
- **Validator (LM Head):** 15% of pool (generates output)

Each node submits its own signed PoNW proof linking to the same `request_id`. The ledger validates all proofs and distributes rewards accordingly.

Market Dynamics:

- **Worthless model:** No demand $\rightarrow$ Price ≈ 0 (nodes focus on training)

- **Improving model:** Rising demand → Price increases naturally
- **Viral demand:** Price spikes → Attracts more inference capacity
- **Equilibrium:** Market self-regulates where training profit $\gg$ inference profit

Key Properties:

1. **Privacy:** Workers never see prompts, only activations
2. **Decentralization:** Uses existing pipeline parallelism architecture
3. **Fair pricing:** Price locked at submission (no timing attacks)
4. **Quality signal:** Demand IS quality (no manual adjustments needed)
5. **Self-regulation:** Market finds equilibrium automatically

This marketplace enables NeuroShard to provide inference services while maintaining its core principles of privacy, decentralization, and fair compensation.

10.4.2 Multipliers and Bonuses

Table 8: Reward Multipliers

Multiplier	Condition	Bonus
Staking Bonus	Diminishing: $\log_2(1 + S/1000)$	up to +66%
Training Bonus	Actively training	+10% rewards
Driver Bonus	Holding embedding layer	+20% rewards
Validator Bonus	Holding LM head + 100 NEURO stake	+30% rewards
Layer Bonus	Per layer held	+5% (max 100%)
Validation Fee	Per proof validated	0.001 NEURO

Note: The Validator Bonus increased from 20% to 30% to compensate for the additional stake requirement and proof validation responsibilities.

10.5 Spending NEURO

Users spend NEURO for:

- **Inference:** 1 NEURO per 1M tokens generated
- **Priority:** Higher payment for faster processing
- **Transfers:** Send NEURO to other users

10.6 Economic Equilibrium

The system creates a sustainable economy:

- **Active Contributors:** Earn more than they spend
- **Passive Users:** Must purchase NEURO to use the service
- **Network Effect:** More users → more training → better model → more users

11 Security Considerations

11.1 Threat Model

NeuroShard assumes:

- Up to 49% of nodes may be malicious
- Attackers are economically rational
- Cryptographic primitives are secure

11.2 Data Privacy

Training data is protected through:

- **Local Processing:** Raw data never leaves the node
- **Differential Privacy:** Token-level noise injection
- **Gradient Compression:** Only sparse gradients are shared

11.3 Model Integrity

NeuroLLM integrity is maintained through:

- **Gradient Validation:** Anomalous gradients are rejected
- **Checkpoint Hashing:** Model state is cryptographically verified
- **Consensus:** Updates require majority agreement

11.4 Sybil Resistance

Sybil attacks are prevented through multiple layers:

- **Validator Stake:** 100 NEURO minimum to become a Hybrid Validator (verify proofs)
- **PoNW Requirement:** Must actually run the model to earn base rewards (stake only provides multiplier)
- **Diminishing Returns:** Creating 10 nodes with 100 NEURO each yields less than 1 node with 1000 NEURO
- **Slashing Risk:** Malicious validators lose their stake (2x penalty)
- **ECDSA Identity:** Node IDs derived from public keys prevent identity spoofing

12 Checkpoint System and State Sharding

12.1 Sharded Checkpoints

Unlike centralized systems where the full model state exists on a single disk, NeuroShard utilizes a **Sharded Checkpoint** system. The global model state is partitioned, with each node responsible for persisting only the weights of its assigned layers.

- **Distributed Storage:** No single node needs to store the entire multi-terabyte model file.

- **Redundancy:** Each shard is replicated across multiple nodes (based on layer assignment redundancy).
- **Merkle Integrity:** A root Merkle hash ensures the consistency of the global state across all shards without requiring full assembly.

12.2 Checkpoint Structure

Each node maintains a local shard:

```
shard = {
    "shard_id": "hash_of_node_id",
    "version": 1234,
    "layer_ids": [0, 1, 2, ...],      # Specific layers held by this
        node
    "state_dict": {...},               # Weights for ONLY these layers
    "optimizer_state": {...},          # Optimizer state for these layers
    "model_hash": "sha256...",         # Hash of this shard
    "timestamp": 1712345678
}
```

Listing 3: Sharded Checkpoint Structure

This approach allows the network to persist models that are significantly larger than the storage capacity of any individual participant.

12.3 Checkpoint Synchronization

New nodes joining the network synchronize via:

1. **Discovery:** Query DHT for nodes with latest checkpoint version
2. **Download:** Fetch checkpoint from multiple peers (parallel chunks)
3. **Verification:** Verify hash matches consensus
4. **Load:** Initialize NeuroLLM from checkpoint

12.4 Model Evolution Over Time

NeuroLLM evolves through two mechanisms:

12.4.1 Continuous Training

Every 60 seconds, a training round completes:

- Gradients from all contributing nodes are aggregated
- Model weights are updated
- New checkpoint is created
- NEURO rewards are distributed

12.4.2 Organic Growth

As more nodes join, the model automatically grows:

1. New node joins with available memory
2. Network assigns new layers to the node
3. Model capacity increases
4. Training continues on larger architecture

No votes required. The model grows as naturally as the network grows. A node with 16GB RAM joining the network immediately increases the model's capacity by ∼156 layers.

12.4.3 Architecture Upgrades

The community can vote on architectural improvements:

- Adding new attention mechanisms
- Expanding vocabulary (new tokens)
- Introducing specialized experts (MoE)

12.5 The Intelligence Ledger

Every improvement to NeuroLLM is recorded in an **Intelligence Ledger**:

Table 9: Intelligence Ledger Entry

Field	**Description**
Round ID	Unique training round identifier
Timestamp	When the round completed
Contributors	List of node IDs that contributed
Batch Size	Total samples processed
Loss Before	Model loss before update
Loss After	Model loss after update
NEURO Minted	Total tokens distributed
Checkpoint Hash	Hash of resulting model state

This creates an immutable record of how NeuroLLM was built - every contribution, every improvement, every NEURO token can be traced back to specific training rounds.

13 Implementation

13.1 Technology Stack

- **PyTorch:** Neural network framework
- **FastAPI:** HTTP API server
- **gRPC:** Inter-node communication
- **Tkinter:** Desktop GUI
- **Kademlia:** DHT implementation
- **PyInstaller:** Cross-platform packaging

13.2 API Endpoints

Table 10: NeuroNode API Endpoints

Endpoint	Method	Description
/generate_text	POST	Generate text from prompt
/contribute_data	POST	Add training data
/train_step	POST	Trigger training step
/training_status	GET	Get training metrics
/api/stats	GET	Node statistics
/api/ponw	GET	Get PoNW proof
/api/model_info	GET	Model information

13.3 Desktop Application

The NeuroShard GUI provides:

- One-click node startup
- Real-time training metrics
- NEURO balance display
- Training enable/disable toggle
- Automatic updates

14 Vision and Roadmap

NeuroShard's evolution is driven by network growth, not predetermined schedules. The following milestones represent capabilities that emerge naturally as participation increases.

14.1 Foundation

The core infrastructure for decentralized AI:

- Dynamic layer pool for organic model scaling
- Gradient gossip protocol for decentralized training
- Proof of Neural Work consensus
- NEURO token economics
- Cross-platform node software (desktop, server)

14.2 Robust Training

Byzantine-tolerant training at scale:

- Trimmed mean and median gradient aggregation
- Fraud proof mechanism with slashing
- Anomaly detection for gradient poisoning
- 50x+ gradient compression
- **Network Target:** 100+ active nodes

14.3 Scale

Optimizing for global-scale throughput:

- **Mobile Optimization:** INT4 quantization and structural pruning for efficient edge device participation.
- **Hierarchical Aggregation:** Multi-tier parameter aggregation to minimize P2P latency in networks >10,000 nodes.
- **Developer SDK:** Comprehensive Python and JavaScript bindings for direct integration into third-party applications.
- **Network Target:** 1,000+ nodes, 1B+ parameters

14.4 Maturity

Global-scale collective intelligence:

- Multi-modal capabilities (vision, audio, code)
- Specialized domain fine-tuning
- NeuroDAO governance for protocol decisions
- Enterprise and institutional integrations
- **Network Target:** 10,000+ nodes, 100B+ parameters

Key Principle: These are not "phases" requiring coordinated upgrades. They are capabilities that emerge as the network grows. A single node joining the network immediately increases model capacity.

15 Distributed Architecture and Parallelism

15.1 The Dynamic Approach

In NeuroShard, distributed inference is not a "future feature" - it is the **default architecture**. Each node holds only the layers it can support, and inference requests flow through the network:

- Node A (2GB): Holds layers 0-18, including embedding
- Node B (4GB): Holds layers 19-56
- Node C (8GB): Holds layers 57-134
- Node D (4GB): Holds layers 135-172, including LM head

An inference request flows: Input $\rightarrow$ Node A $\rightarrow$ Node B $\rightarrow$ Node C $\rightarrow$ Node D $\rightarrow$ Output. Note that while the logical flow is linear, the physical routing is handled by the Swarm Network Layer (Section 16), where each step is dynamically routed to a redundant pool of peers to ensure fault tolerance.

15.2 Advanced Parallelism

NeuroShard implements advanced parallelism strategies to support massive models on consumer hardware.

15.2.1 Tensor Parallelism

To handle layers that are too large for a single device's memory (e.g., a 8192-dimension projection matrix), NeuroShard implements **Tensor Parallelism**. Large weight matrices are split across multiple nodes:

- **Column Parallelism:** Dividing linear layers by output features.
- **Row Parallelism:** Dividing linear layers by input features.

This allows a single logical layer to span multiple physical devices, enabling the training of models with hidden dimensions far exceeding individual VRAM limits.

15.2.2 Pipeline Parallelism

As the default mode of operation, Pipeline Parallelism distributes the model vertically. Input activations flow through the network of nodes, each processing a subset of layers, similar to a packet moving through a network router chain. This naturally balances compute load across the heterogeneous cluster.

15.3 Hybrid Parallelism

For very large models, NeuroShard combines both approaches:

1. **Inter-layer (Pipeline):** Different nodes hold different layers
2. **Intra-layer (Tensor):** Large layers split across multiple nodes

This enables models of **any size** to run across the network, each node contributing what it can.

15.4 Training at Scale

Distributed training uses:

- **Gradient Accumulation:** Nodes accumulate gradients locally before sharing
- **Ring All-Reduce:** Efficient gradient synchronization across nodes
- **Asynchronous SGD:** Nodes can train at different speeds
- **Layer-Local Gradients:** Each node only computes gradients for its assigned layers

15.5 The Key Insight

Traditional distributed training requires homogeneous hardware and careful coordination. NeuroShard's dynamic layer pool allows **heterogeneous participation**:

- A Raspberry Pi with 1GB RAM contributes 9 layers
- A gaming PC with 32GB RAM contributes 312 layers
- A cloud server with 256GB RAM contributes 2,500 layers

All participate in the same model, all earn NEURO proportional to contribution.

16 Swarm Architecture: High-Latency Resilience

16.1 Design Philosophy: Throughput over Latency

The core insight driving NeuroShard's network design is: **"Compute is cheap; Bandwidth is expensive."** Residential internet connections (100Mbps-1Gbps) and node churn create fundamental challenges for distributed training. Rather than fighting these constraints, NeuroShard embraces them through an asynchronous, swarm-based architecture.

Goal: 95% GPU utilization, even if wall-clock convergence is 10% slower per epoch. We win by being *unstoppable*, not by being the fastest.

16.2 Swarm Network Layer (Fault-Tolerant Routing)

Traditional pipeline parallelism fails catastrophically when any node in the chain becomes unresponsive. NeuroShard replaces the brittle linear pipeline with a **multipath routing system**.

16.2.1 Key Directive

"If Node A hangs, the packet must automatically flow to Node B without crashing the run."

16.2.2 Multipath Routing

For each layer range, the DHT maintains a list of K potential peers (typically $K = 3$), sorted by a weighted score:

$$\text{Score}_i = w_{\text{latency}} \cdot \frac{\text{Latency}_i}{500\text{ms}} + w_{\text{queue}} \cdot \frac{\text{QueueDepth}_i}{100} \tag{17}$$

Where $w_{\text{latency}} = 0.4$ and $w_{\text{queue}} = 0.6$. Lower scores indicate better candidates.

16.2.3 Probabilistic Send with Failover

When sending activations:

1. Send to primary peer (lowest score)
2. If no ACK within 200ms, immediately re-route to secondary peer
3. Continue until success or all K candidates exhausted

This provides automatic failover without coordinator intervention.

16.2.4 Capacity Bitmask Heartbeats

Every 5 seconds, nodes broadcast a lightweight **Capacity Bitmask** ($\sim$20 bytes):

- Available memory (MB)
- Current queue depth
- Layer range held
- GPU utilization (%)
- Status flags (training/inference/accepting work)

This allows peers to know who is ready to receive work, enabling intelligent routing decisions.

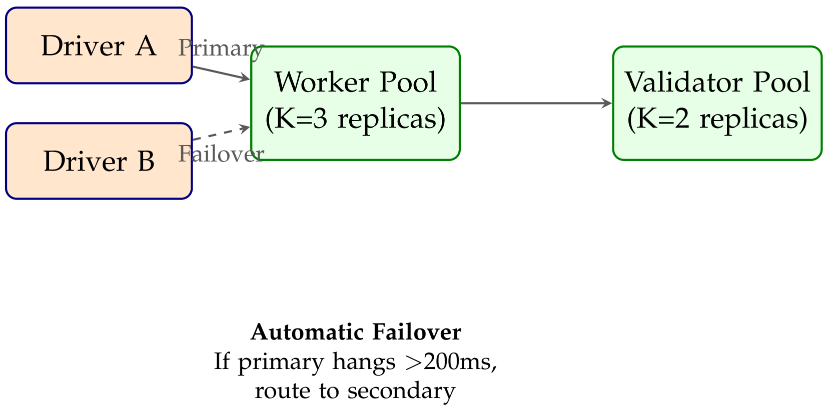

Figure 5: Swarm Routing: Multipath routing with automatic failover. Each layer range has K candidate peers.

16.3 Asynchronous Micro-Batch Engine

16.3.1 Key Directive

"Decouple Compute from Communicate. The forward pass must run 100 steps ahead of the backward pass if necessary."

The goal is to eliminate GPU starvation. The GPU should *never* wait for network packets.

16.3.2 Activation Buffering

Each node maintains two priority queues:

- **Inbound Buffer:** Incoming activations sorted by priority
- **Outbound Buffer:** Processed activations awaiting transmission

The GPU worker simply pops the next available item from Inbound, computes, and pushes to Outbound. Network I/O happens asynchronously.

$$\text{Buffer Fill Rate} = \frac{\text{Current Queue Size}}{\text{Max Queue Size}} \tag{18}$$

- Fill Rate < 0.1: Node is **Starved** (bad) — GPU idle
- Fill Rate > 0.9: Node is **Backpressured** (good) — GPU saturated

16.3.3 Interleaved 1F1B Schedule

Rather than strict synchronous execution, NeuroShard uses an **Interleaved 1F1B** (one-forward, one-backward) schedule:

```
Schedule (4 micro-batches):
F0 F1 F2 F3 B0 F4 B1 F5 B2 F6 B3 ...
```

Key insight: Start backward passes *before* all forwards complete. This overlaps backward compute with forward network latency, maximizing GPU utilization.

16.3.4 Dynamic Micro-Batch Sizing

Micro-batch size is dynamically tuned to find the "Goldilocks zone":

- Large enough to saturate the GPU
- Small enough to fit in network MTU ($\sim$64KB target)

$$\text{MicroBatchSize} = \min\left(\frac{\text{AvailableMemory} \times 0.1}{\text{ActivationSize}}, \frac{\text{64KB}}{\text{SeqLen} \times \text{HiddenDim} \times 4}\right) \tag{19}$$

16.4 Extreme Gradient Accumulation (DiLoCo-Style)

16.4.1 Key Directive

"Accumulate locally for as long as possible. Sync only when statistically necessary."

Inspired by DiLoCo [14], NeuroShard slashes bandwidth by 90%+ through lazy syncing.

16.4.2 Local Inner Loop

Instead of syncing gradients every step, each node:

1. Saves initial weights w_0
2. Trains independently for N steps (e.g., $N = 500$) using local optimizer
3. Computes pseudo-gradient: $\Delta w = w_0 - w_N$
4. Only then triggers network sync

16.4.3 Outer Optimizer

The aggregated pseudo-gradients are applied using an outer optimizer with Nesterov momentum:

$$m_{t+1} = \beta \cdot m_t + \Delta w_{\text{aggregated}} \tag{20}$$

$$w_{t+1} = w_t + \eta_{\text{outer}} \cdot (\beta \cdot m_{t+1} + \Delta w_{\text{aggregated}}) \tag{21}$$

Where $\beta = 0.9$ (momentum) and $\eta_{\text{outer}} = 0.7$ (outer learning rate).

16.4.4 Bandwidth Reduction

Table 11: Bandwidth Comparison: Standard vs DiLoCo-Style

Metric	Standard Gossip	DiLoCo (N=500)
Syncs per 1000 steps	1000	2
Bandwidth per step	100%	0.2%
Latency tolerance	Low	High
Straggler tolerance	Low	High

16.4.5 Enhanced Verification

Since syncs are less frequent, bad gradients are more damaging. NeuroShard implements enhanced verification:

1. Compute gradient on local "trusted" micro-batch
2. Compare distribution to submitted gradient:
    - Cosine similarity > 0.5 (direction alignment)
    - Magnitude ratio within $10\times$ (scale check)
    - Variance ratio within $100\times$ (distribution check)
3. Reject gradients that fail any check

16.5 Speculative Checkpointing

To enable fast crash recovery, each node maintains "hot" snapshots:

- **Frequency:** Every 2 minutes (background thread)
- **Contents:** Model weights, optimizer state, DiLoCo buffers
- **Retention:** Last 5 snapshots
- **Announcement:** Published to DHT for neighbor discovery

On crash, a replacement node fetches the hot snapshot from a neighbor rather than restarting the epoch. Recovery time drops from "full restart" to < 30 seconds.

16.6 Summary: Resilience Through Design

Table 12: Swarm Architecture Benefits

Challenge	Solution	Result
Node failure	Multipath routing	<200ms failover
GPU starvation	Activation buffering	95% utilization
High latency	DiLoCo accumulation	$500\times$ fewer syncs
Crash recovery	Hot snapshots	<30s recovery
Network variability	Capacity heartbeats	Intelligent routing

17 Philosophy: Why Decentralized AI Matters

17.1 The Concentration Problem

As of 2024, the ability to create frontier AI is concentrated in fewer than 10 organizations worldwide. These organizations:

- Control what the AI can and cannot say
- Decide who gets access and at what price
- Collect and monetize user interactions
- Can shut down access at any time

This concentration of power over intelligence is unprecedented in human history. NeuroShard exists to provide an alternative.

17.2 The NeuroShard Principles

17.2.1 Principle 1: Intelligence Should Be a Public Good

Just as the internet democratized access to information, NeuroShard aims to democratize access to *intelligence*. NeuroLLM is not a black-box API owned by a corporation; it is a public utility created, maintained, and improved by the community that uses it.

17.2.2 Principle 2: Contributors Should Own What They Build

Every gradient computed, every data sample contributed, every hour of compute donated -these contributions are recorded on the Intelligence Ledger. NEURO tokens represent a programmable claim on the collective intelligence being built; they are a way to say, in cryptographic form, "I helped create this model."

17.2.3 Principle 3: No Single Point of Control

NeuroLLM cannot be censored, shut down, or silently modified by any single entity. As long as the network exists, the model exists. This is not just a systems design choice; it is a political statement about who should steer the trajectory of powerful AI: a narrow set of firms, or the people who actually use and train it.

17.2.4 Principle 4: Start Simple, Grow Together

NeuroLLM starts as a 125M parameter model that produces mostly gibberish. This is intentional. We refuse to bootstrap decentralization from a centralized model trained behind closed doors, with unknown data, incentives, and filters. Instead, we start from random weights, in the open, and let the network grow the model into something powerful over time.

17.3 The Long-Term Vision

In 10 years, we envision:

- **NeuroLLM as a global brain:** Billions of parameters, trained on the collective knowledge of millions of contributors
- **NEURO as AI currency:** The standard token for AI services worldwide
- **Specialized variants:** Community-governed fine-tunes for medicine, law, science, art
- **True AI democracy:** Major decisions about AI development made by token holders, not corporations

18 Conclusion

NeuroShard introduces a fundamentally new approach to artificial intelligence: a model created, trained, and owned by the community. NeuroLLM represents the first truly decentralized large language model - one that:

- **Belongs to everyone:** No corporation controls NeuroLLM. Every contributor is a co-owner.
- **Grows with participation:** More users = more compute = more data = better model. The network effect is the engine of intelligence.

- **Rewards contributors:** NEURO tokens for compute and data. Your contribution is recorded forever on the Intelligence Ledger.
- **Resists censorship:** No single point of control. No kill switch. No terms of service that can change overnight.
- **Builds intelligence from scratch:** We don't redistribute corporate models. We create our own, from random weights, through collective effort.

The model starts "dumb" and grows intelligent through collective effort. This is not a limitation - it is the foundation of true decentralization. Every participant who contributes to NeuroLLM's training is a co-creator of a new form of collective intelligence.

Proof of Neural Work ensures that every NEURO token represents real intelligence added to the model. Unlike cryptocurrencies that waste energy on arbitrary computation, NeuroShard channels that energy into building something useful: a public AI that belongs to humanity.

The vision of NeuroShard is a world where AI is a public good, created by humanity, for humanity. We are not asking permission from corporations to build the future. We are building it ourselves, one gradient at a time.

Join us. Contribute compute. Contribute data. Earn NEURO. Own the future of intelligence.

19 Acknowledgments

We thank the open-source community for the foundational technologies that made NeuroShard possible, including PyTorch, the broader P2P research community, and the pioneers of decentralized AI.

References

[1] B. McMahan et al., "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*, 2017.

[2] Y. Huang et al., "GPipe: Efficient training of giant neural networks using pipeline parallelism," in *Advances in Neural Information Processing Systems*, 2019.

[3] A. Borzunov et al., "Petals: Collaborative Inference and Fine-tuning of Large Models," in *ACL*, 2023.

[4] M. Ryabinin et al., "Hivemind: a Library for Decentralized Deep Learning," in *NeurIPS*, 2020.

[5] Bittensor Network. https://bittensor.com

[6] Gensyn Protocol. https://docs.gensyn.ai/litepaper

[7] B. Zhang and R. Sennrich, "Root Mean Square Layer Normalization," *arXiv:1910.07467*, 2019.

[8] J. Su et al., "RoFormer: Enhanced Transformer with Rotary Position Embedding," *arXiv:2104.09864*, 2021.

[9] J. Ainslie et al., "GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints," *arXiv:2305.13245*, 2023.

[10] N. Shazeer, "GLU Variants Improve Transformer," *arXiv:2002.05202*, 2020.

[11] T. Brown et al., "Language Models are Few-Shot Learners," *Advances in Neural Information Processing Systems*, 2020.

[12] J. Hoffmann et al., "Training Compute-Optimal Large Language Models," *arXiv:2203.15556*, 2022.

[13] J. Kaplan et al., "Scaling Laws for Neural Language Models," *arXiv:2001.08361*, 2020.

[14] A. Douillard et al., "DiLoCo: Distributed Low-Communication Training of Language Models," *arXiv:2311.08105*, 2023.

[15] M. Ryabinin et al., "SWARM Parallelism: Training Large Models Can Be Surprisingly Communication-Efficient," *arXiv:2301.11913*, 2023.

A Code Examples

A.1 Starting a DynamicNeuroNode

```python
from neuroshard.core.dynamic_model import create_dynamic_node

# Create and start node (auto-detects memory)
node = create_dynamic_node(
    node_token="your-unique-token",
    port=8000
)

# Node automatically:
# 1. Detects available memory (e.g., 4GB)
# 2. Registers with network
# 3. Gets assigned layers (e.g., layers 0-38)
# 4. Loads only those layers (~350M params)
# 5. Starts earning NEURO (reward multiplier ~1.4x)

print(f"Assigned layers: {node.my_layer_ids}")
print(f"My params: {node.model.get_num_params() / 1e6:.1f}M")
print(f"Reward multiplier: {node.get_ponw_proof()['reward_multiplier
    ']:.2f}x")
```

Listing 4: Node Startup

A.2 Contributing Training Data

```python
# Contribute text data for training
node.contribute_training_data(
    text="Your training text here...",
    apply_dp=True  # Apply differential privacy
)

# Check buffer status
stats = node.data_manager.get_stats()
print(f"Buffer: {stats['buffer_size']} samples")
```

Listing 5: Data Contribution

A.3 Generating Text

```python
# Generate text (quality improves with training)
result = node.generate(
    prompt="Hello, world!",
    max_new_tokens=50,
    temperature=0.8
)

print(result)
# Early: "Hello, world! xkjf asd random gibberish..."
# Later: "Hello, world! How can I help you today?"
```

Listing 6: Text Generation

B Configuration Reference

B.1 NeuroLLM Dynamic Architecture Config

The architecture adapts to network capacity using scaling laws:

```python
def calculate_optimal_architecture(total_memory_mb):
    """
    Calculate optimal width x depth based on total network capacity.

    Examples:
      40GB network   -> 16L x 1024H  (~350M params)
      800GB network  -> 32L x 3072H  (~9.2B params)
      8TB network    -> 64L x 7168H  (~123B params)
    """
    # Apply empirical scaling laws
    # Width grows as memory^0.6, depth as memory^0.4
    params_budget = (total_memory_mb * 0.6) / 16

    hidden_dim = calculate_width(params_budget)
    num_layers = calculate_depth(params_budget)
    num_heads = hidden_dim // 64  # 64 dim per head
    num_kv_heads = num_heads // 3  # GQA ratio
    intermediate_dim = hidden_dim * 8 / 3  # SwiGLU

    return ModelArchitecture(
        hidden_dim=hidden_dim,
        num_layers=num_layers,
        num_heads=num_heads,
        num_kv_heads=num_kv_heads,
        intermediate_dim=intermediate_dim,
        vocab_size=32000,
        max_seq_len=2048
    )
```

Listing 7: Dynamic Architecture Calculation

B.2 Dynamic Layer Pool Configuration

```python
# Layer pool parameters
MIN_REPLICAS = 2                # Each layer replicated on 2+ nodes
```

```
HEARTBEAT_TIMEOUT = 120          # Seconds before layer reassignment
RECALC_INTERVAL_NODES = 50       # Recalc architecture every N nodes
MIN_UPGRADE_IMPROVEMENT = 1.3    # Only upgrade if 30%+ better

# Architecture-aware layer assignment
def calculate_layer_assignment(node_memory_mb, current_arch):
    """How many layers can this node hold with current architecture?"""
    memory_per_layer = estimate_memory_per_layer(current_arch)
    usable_memory = node_memory_mb * 0.6  # Safety margin
    return max(1, int(usable_memory / memory_per_layer))

# Reward calculation (unchanged)
def calculate_reward_multiplier(
    num_layers_held,
    total_network_layers,
    has_embedding,
    has_lm_head
):
    base = 1.0 + (num_layers_held / total_network_layers)
    if has_embedding: base *= 1.2  # Driver bonus
    if has_lm_head: base *= 1.3    # Validator bonus
    return base
```

Listing 8: Layer Pool with Architecture Scaling