Gonka: Designing a Compute-Native Decentralized Economy

2020-05-20

Gonka is a decentralized AI infrastructure designed to optimize computational power specifically for AI model training and inference, offering a competitive alternative to traditional centralized cloud providers. Centralized systems are often expensive, monopolistic, and carry risks of censorship, while existing decentralized networks frequently waste resources on non-productive tasks like network security.

We introduce an innovative consensus mechanism that ensures nearly 100% of computational resources are used for meaningful AI tasks, maximizing efficiency and minimizing operational costs.

The system features key roles: Developers build and deploy AI applications using the network's distributed power; Hosts contribute computational resources to the network and are rewarded based on the amount and quality of resources they provide.

This collaboration allows the platform to offer AI services at significantly lower prices, making advanced AI technology more accessible to a broader audience.

This document explains the economic architecture that drives Gonka Network: how incentives are aligned, how resources are priced, and how the gonka (GNK) coin ensures sustainable growth.

It outlines the mechanisms that govern rewards, fees, and token issuance, with a focus on predictable Developer costs and long-term fairness for Hosts. Each component is designed to maximize real utility while minimizing complexity and unnecessary overhead.

In particular, this paper details:

1. How the dual-reward system balances early network bootstrapping with long-term efficiency;
2. The principles behind the gonka (GNK) coin issuance model and its adaptive supply mechanics;
3. The structure of Developer-facing costs and how these remain transparent over time;
4. Risk factors and how they are mitigated through incentive design;
5. Comparative benchmarks that highlight Gonka's competitive advantages over existing alternatives;
6. The governance and pricing mechanisms that enable flexibility as the network evolves.

Contents

1. The Incentive Model of Gonka Network

1.1. Dual-reward system

A dual-reward system analogous to the mechanisms observed in Bitcoin is proposed to incentivize participation within the network. Hosts who provide computational power are compensated through two primary avenues: transaction fees and a network-wide inflationary reward system.

Transaction fees are embedded within each transaction, similar to Bitcoin, where the end-user or Developer attaches a fee to execute specific tasks. This ensures that Hosts contributing computational resources are rewarded directly for their efforts.

In addition to transaction fees, Hosts receive rewards through an inflationary mechanism. Each block generated within the network issues a reward distributed proportionally among Hosts based on their computational contribution. This reward decreases over time, following a phased approach:

1. Initial subsidy phase: A substantial subsidy offsets the capital expenditure (CAPEX) associated with providing hardware in the network's early stages. This incentivizes early participation and ensures network growth. This subsidy will taper off as the network matures, transitioning the primary reward source to transaction fees, mirroring Bitcoin's diminishing block rewards.
2. Market efficiency: As subsidies decrease, market efficiency is expected to rise, driven by advancements in hardware performance. Just as the Bitcoin network saw the development of highly efficient ASICs, the network anticipates similar innovations, reducing the cost of computational power provision and ensuring sustainability as rewards shift entirely to transaction fees.

This dual-reward system also addresses the challenge of hardware utilization. In the network's launch phase, when hardware may be underutilized, initial subsidies distribute costs across fewer tasks, making CAPEX more manageable. As the network expands and hardware utilization increases, the need for subsidies diminishes, and transaction fees become the sole reward source. This progression allows the network to scale effectively while maintaining a competitive edge over centralized cloud providers, ultimately offering services at a fraction of the cost.

1.2. Transaction cost management

Transaction costs within the network are designed to be flexible and predictable, akin to Ethereum’s GAS model or the OpenAI API's pricing structure. Users specify a maximum cost they are willing to pay, ensuring that tasks are performed within this predefined limit.

If computational resources are exhausted before task completion, the transaction is canceled, and the spent tokens are not refunded, similar to Ethereum’s handling of exceeded GAS limits. The rewards generated from these transactions are distributed between the Host executing the task and the one verifying it, ensuring fair compensation across Hosts.

1.3. Long-term sustainability

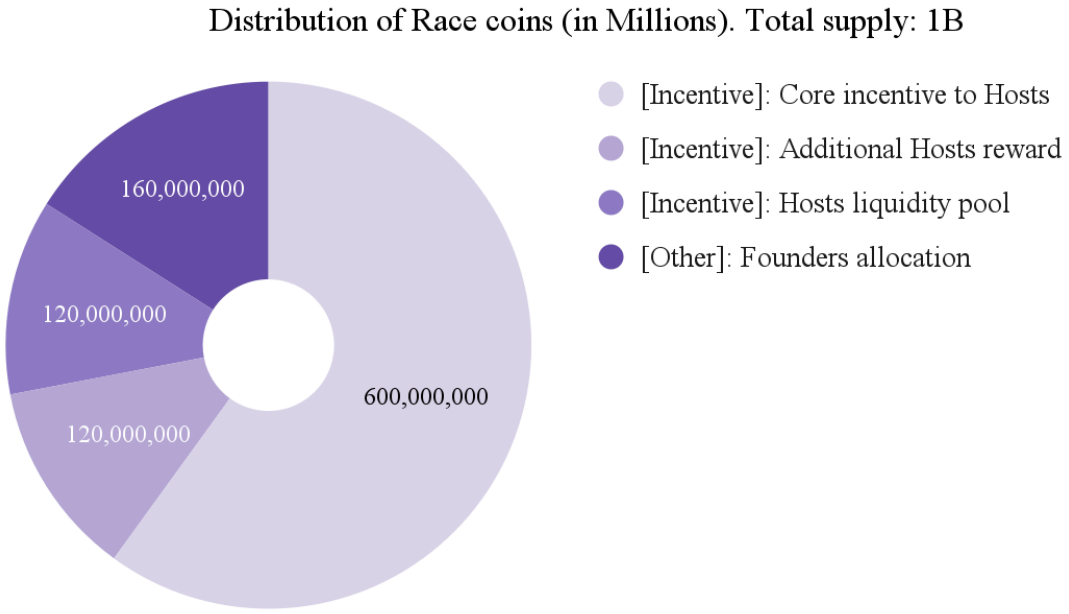
The long-term strategy focuses on attracting Hosts willing to dedicate their hardware to the network, positioning the project as a cost-competitive alternative to traditional cloud providers. Leveraging decentralization, the network is designed to offer significantly lower prices for computational tasks, making it an attractive option for Developers and businesses seeking cost-effective solutions.

In conclusion, the tokenomics strategy is crafted to ensure the network's sustainability, support early adopters, and maintain competitive pricing. Balancing initial subsidies with a gradual transition to transaction fees establishes a robust ecosystem that rewards Hosts, drives hardware innovation, and positions the network as a leader in decentralized AI computation.

2. Tokenomics

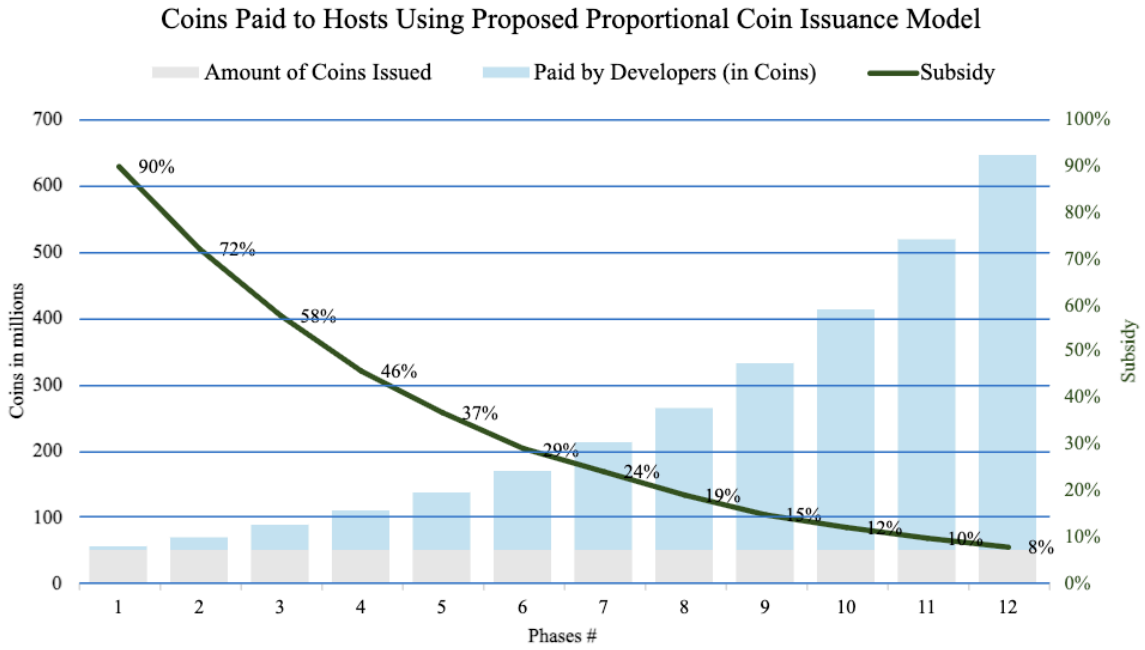
The total supply of gonka (GNK) coins is fixed at 1 billion, allocated to incentivize Hosts, support network development, and ensure fair compensation for contributors. Below is the detailed breakdown of the allocation. The distribution is divided into two primary categories:

1. **Incentives to Hosts** (840 million gonka (GNK) coins, 84% of total supply).
2. **Other: Founders allocation** (160 million gonka (GNK) coins, 16% of total supply) is a portion reserved for the founding team as allocation to the founders for their ownership stake.



Incentives to Hosts consist of the following:

1. **Core incentive to Hosts** (600 million gonka (GNK) coins, 60% of total supply) is designed to reward Hosts for contributing computational power to the decentralized network. Similar to Bitcoin mining, Hosts directly receive newly minted coins proportional to their computational contribution.



To effectively distribute these incentives, the Proportional Coin Issuance Model is introduced. Under the structured token-distribution mechanism, the issuance of gonka (GNK) coins occurs gradually across predefined phases. Each phase issues a fixed batch of 50 million coins while progressively reducing the network-funded subsidy for Hosts. Specifically, subsidies begin at 90% in Phase 1 and systematically decrease down to 8% by Phase 12, following a defined 1/5-life

reduction model. Although the exact timing between phases remains undetermined, the issuance strategy is to align each phase roughly with the network's annual growth milestone (~1-year interval).

To ensure long-term credibility, the entire issuance model is permanently encoded on-chain at the time of mainnet deployment. While parameters may be refined during the pre-launch (testnet) phase, no centralized party, including the founding team, will have the authority to unilaterally alter these rules after mainnet launch. Any future modification would require a decentralized governance process.

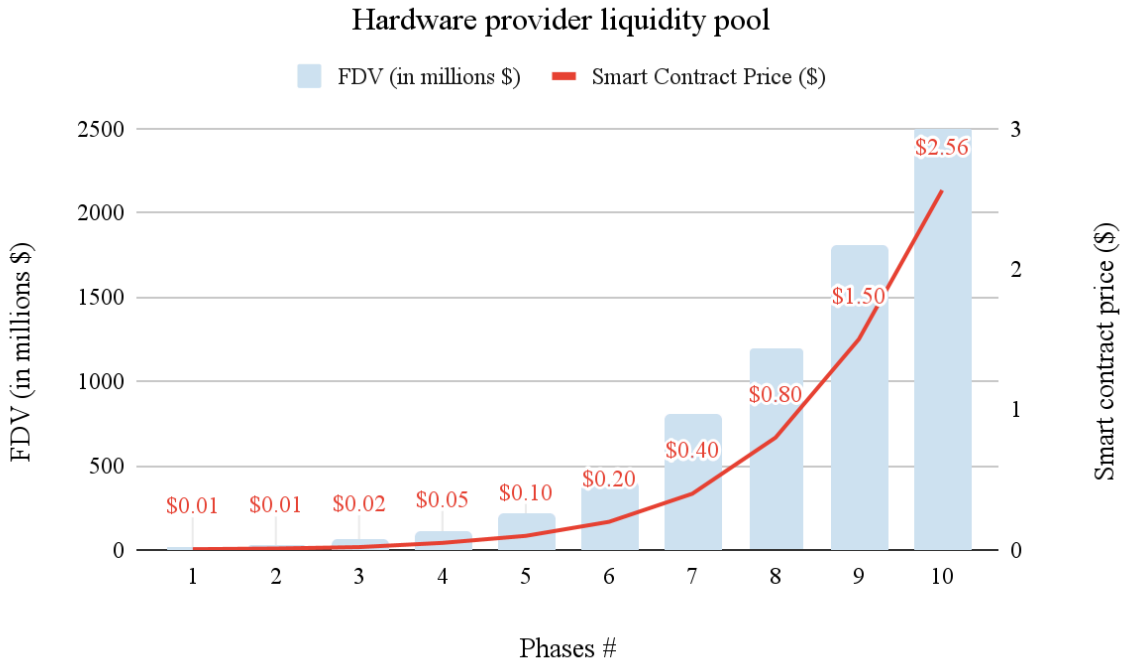
This Proportional Coin Issuance Model is specifically designed to balance incentives for early-stage adoption with long-term scalability of the network. By progressively aligning token issuance with network growth, the model ensures fair compensation for Hosts.

In the network's early stages, demand for inference service may be limited due to slower adoption rates by Developers utilizing AI models (Developers). Without incentive, Hosts might find it economically challenging to contribute computational power consistently while demands is low. To address this, **the subsidy mechanism** acts as a stabilizing force, guaranteeing fair compensation on all stages of network development. This ensures the network maintains adequate computational capacity, thereby supporting early adopters and core functionality. As adoption accelerates and Developer demand for inference increases, Hosts begin earning more market-driven rewards. At this stage, the need for network-funded subsidies diminishes. The gradual subsidy reduction ensures a smooth transition from subsidy reliance toward self-sustaining, market-driven reward system.

2. **Additional Host Reward** (120 million gonka (GNK) coins, 12% of total supply) is targeted at recognizing and additionally rewarding early, high-performing Hosts. Specifically, this 12% allocation is divided among the first top Hosts who each committed at least 1,000 A100 GPUs and sustained their contribution for 4 years. This structured reward mechanism ensures long-term stability and incentivizes sustained participation from high-performance Hosts. By gradually distributing the reward over four years, the system prevents short-term exploitation and encourages Hosts to maintain their contribution consistently. Additionally, the mechanism allows for dynamic replacement - if a top Host exits, the next eligible Host inherits the remaining rewards, ensuring continuity without disrupting the network.

3. **Host liquidity pool** (120 million gonka (GNK) coins, 12% of total supply) is designed to provide early liquidity for Hosts before gonka (GNK) coin is adopted by exchanges and P2P trading volume reaches levels comparable to similar projects. This pool will be distributed through a preprogrammed sale mechanism that allows Developers and token holders to buy gonka (GNK) coins at any time.

   Buyer can transfer USDT, Ethereum, or Bitcoin to a designated account on the network, automatically minting gonka (GNK) coins via the native bridge. The price follows a bonding curve, meaning it increases exponentially as more coins are minted. Meanwhile, Hosts will be able to send their mined gonka (GNK) coins to a dedicated smart contract on Gonka network and receive USDT, Ethereum, or Bitcoin from the pool at the latest achieved price, ensuring early liquidity.



For example, price changes can be as follows: The coins will be released in phases. The release of the next phase is contingent on the complete sale of the preceding phase. The first phase releases 12 million coins at a starting price of $0.005. Once all coins in the first phase are sold, the second phase is released at a higher price, increasing to $0.01, followed by subsequent phases with progressively higher prices. This "sell-out-before-next-phase" approach incentivizes early investment by allowing early Hosts to acquire gonka (GNK) coins at the lowest price point. As each phase sells out, the scarcity of available coins drives demand, rewarding Hosts who act early.

3. Reasoning behind the Proportional Coin Issuance Model

The token issuance model is designed to achieve two primary objectives: fostering network growth during its early stages and ensuring long-term sustainability. Drawing on principles from Proof-of-Work systems, the model integrates utility-token mechanics to create a predictable and efficient framework. This chapter explains the principles behind token issuance, the logic guiding its design, and the safeguards that prevent exploitation.

Tokens play a fundamental role in supporting the network by subsidizing early growth and ensuring long-term collaboration between two key Hosts: Hosts and Developers.

In the network's initial phase, when the number of inference requests is not sufficiently high, the network will be subsidized for incentivising participation:

- Subsidies motivate Hosts to contribute the computational power required to support network operations.
- Subsidies reduce costs, enabling Developers to deploy and run AI services on the network at significantly lower expenses compared to centralized alternatives.

These incentives help establish a robust foundation, encouraging adoption and ensuring sufficient activity to drive the network's growth. While Hosts compete to deliver services at the lowest effective cost (calculated as operational expenses minus subsidies), Developers benefit from predictable costs, even as subsidies decline, allowing them to plan operations and scale effectively.

Gonka (GNK) coin's issuance model is built on the best practices from the market while adding its own innovations for scalability and efficiency. Inspired by Bitcoin and Ethereum, gonka (GNK) coin's issuance model is based on a hybrid approach that adapts dynamically to the network's demand and usage. To

understand this approach, it's important to understand how Bitcoin and Etherium have structured their token distribution to incentivize participation and ensure long-term sustainability. In the beginning, more tokens are usually issued to incentivize participation:

- In **Bitcoin**, about 12.51% of Bitcoin's total supply was distributed in its first year (50 BTC per block). This reward undergoes a halving approximately every four years, reducing the issuance rate by half. The first halving in 2012 reduced the reward to 25 BTC, followed by 12.5 BTC in 2016, 6.25 BTC in 2020, and 3.125 BTC in 2024. This process continues until the total supply of 21 million BTC is distributed. During this issuance period, Hosts are incentivized by two mechanisms:
    - Transaction Fees: Paid by users for transactions.
    - Block Rewards: In the form of newly minted BTC.

  Together, these incentives ensure sufficient motivation for Hosts to participate actively in securing the network. Gradually, all tokens are distributed, and it is assumed that no additional investment will be required, as transaction fees will be sufficient for the network's operations to continue without the need for additional injections.
- **Ethereum** followed a different model. Initially, the annual issuance rate under Proof-of-Work was approximately 6.7% of the total supply. With the transition to Proof-of-Stake through "The Merge," the issuance rate dropped significantly to around 0.52%. Although Ethereum's approach differs from Bitcoin's time-based halving, both systems gradually reduce issuance as the network matures.

3.1. Gonka's dynamic supply-driven halving

Unlike Bitcoin's time-bound halving or Ethereum's rate adjustments, Gonka adopts a Hybrid Proportional Coin issuance model with a minimum independent Coin issuance (set at 5% of the total supply for the first year and will decrease by half each subsequent year) ties directly to network activity.

The total supply of the gonka (GNK) coin is 1,000,000,000. Subsidies will be issued by batches (1 batch = 50M gonka (GNK) coins, equivalent to 5% of the total supply). Each time 5% of the total token supply is issued, the subsidy decreases by 20%. This balances incentives with network demand, ensuring token issuance aligns with usage and avoids unnecessary inflation.

These issuance parameters (including batch size, total supply, and subsidy decrease rate) are permanently encoded at the time of mainnet deployment. While these values may be adjusted during the pre-launch (testnet) phase, they become immutable once the network goes live. After deployment, neither the founding team nor any centralized entity can unilaterally modify them. Any future adjustments would require a decentralized on-chain governance process, subject to majority approval.

#	gonka (GNK) coins	1/5-life reduction of subsidies	Fees paid by Developers in gonka (GNK) coins	Total paid to Hosts in gonka (GNK) coins
1	50,000,000	90.00%	5,555,556	55,555,556
2	50,000,000	72.00%	19,444,444	69,444,444
3	50,000,000	57.60%	36,805,556	86,805,556
4	50,000,000	46.08%	58,506,944	108,506,944
5	50,000,000	36.86%	85,633,681	135,633,681
6	50,000,000	29.49%	119,542,101	169,542,101
7	50,000,000	23.59%	161,927,626	211,927,626
8	50,000,000	18.87%	214,909,532	264,909,532
9	50,000,000	15.10%	281,136,915	331,136,915
10	50,000,000	12.08%	363,921,144	413,921,144

11	50,000,000	9.66%	467,401,430	517,401,430
12	50,000,000	7.73%	596,751,788	646,751,788
	600,000,000			

3.2. Minimum independent Coin issuance

Hybrid Proportional Coin Issuance Model ensures that a minimum number of new coins are created each year, regardless of demand.

How it works:

- At the beginning of each year, the system sets a minimum daily issuance rate, which is calculated based on a percentage of the total supply. This ensures that a certain number of coins enter circulation every day, regardless of demand.
    - In the first year, this minimum is 5% of the total supply. 5% of 1 billion = 50 million coins. Daily issuance: 50,000,000 ÷ 365 = 136,986 coins per day
    - In the second year, it reduces to 2.5% of 1 billion = 25 million coins. Daily issuance: 25,000,000 ÷ 365 = 68,493 coins per day.
    - In the third year, it further drops to 1.25% of 1 billion = 12.5 million coins. Daily issuance: 12,500,000 ÷ 365 = 34,247 coins per day
    - Etc.
- Hosts earn rewards based on the amount of work they have done. However, if demand is so low that the set minimum coin issuance is not fully utilized, this minimum issuance will still be distributed among those who have contributed work, in proportion to their efforts.

The tokenomics model provides clear and predictable costs for Developers, making it easier for them to plan long-term participation:

- **Early stages**. Developers can expect computational costs to be up to ten times cheaper than market rates due to the 90% subsidy provided during the initial phases. For example, if a Developer spends X tokens on a request, Hosts may receive up to 9X in rewards as part of the subsidy.
- **Gradual adjustment.** As subsidies decline with each halving, Developers experience steady and predictable cost increases. This transparency ensures that Developers can budget for growth without being affected by sudden or unpredictable changes.

To prevent exploitation of the subsidy structure, the model incorporates randomized task distribution. Tasks are distributed randomly across Hosts, ensuring no Host can disproportionately benefit from artificially generated requests. Developers have no control over which provider processes their tasks.

Also, as subsidies decrease, market efficiency is expected to rise, driven by advancements in hardware performance. Just as the Bitcoin network saw the development of highly efficient ASICs, the network anticipates similar innovations, reducing the cost of computational power provision and ensuring sustainability as rewards shift entirely to transaction fees.

4. Hypothetical* Competitive Analysis

**This model is a hypothesis, not a promise, designed to explore the potential of balancing scalability and fairness in a decentralized network. If the gonka (GNK) coin network achieves adoption rates comparable to competitors like Bittensor (Year 3) or OpenAI (Year 6), Hosts stand to benefit significantly. However, these outcomes depend on network growth and are subject to market dynamics.*

Year 3 Comparison with Bittensor

Competitor Analysis:

Scenario 1 - Bittensor by year 3

Market Cap (in millions)		FDV (in millions)	
Bittensor	gonka (GNK) coin	Bittensor	gonka (GNK) coin
$3,100	$411	$8,820	$894
	-655% Undervalued		-888% Undervalued

Difference in Market Cap: $3.1B (Bittensor) vs. $411M (gonka (GNK) coin).
Gonka (GNK) coin is hypothetically undervalued by 655%.

Difference in FDV: $8.8B (Bittensor) vs. $894M (gonka (GNK) coin).
Significant room for gonka (GNK) coin valuation growth.

Data inputs and assumptions for this analysis:

- **Bittensor metrics (December 2024)**
  - Approximate request volume (in millions AI-tokens): 7.8M
  - Market Cap: $3.1B
  - FDV: $8.8B
  - Market Price per 1M AI-tokens (2024): $10
- **Gonka assumptions (used in calculations below)**
  - Approximate discount expected by Developers compared to market price: 57.6%
  - Fees paid by Developers in gonka (GNK) coins (Year 3, not cumulative): 36.8M gonka (GNK) coins
  - gonka (GNK) coins issued by Year 3:3 batches x 50M + 30M (¼ of Host liquidity pool) + 90M (First three years of additional Host reward) + 160M Founders allocation = 430M

Based on these assumptions, the following mathematical calculations are performed:

1. **Adjusted market price per 1M tokens**
   Market price per 1M tokens (2024) — Discount $= \$10 - 57.6\% = \$4.2$

2. **Cost per 1M requests in gonka (GNK) coins to serve request volume similar to Bittensor**

$$\frac{Fees\ Paid\ by\ Developers\ (Year\ 3)}{Request\ Volume\ (M\ Tokens)} = \frac{36.8M}{7.8M} \approx 4.7\ gonka\ (GNK)\ coins\ per\ 1M\ requests$$

3. **Price per gonka (GNK) coin (based on calculations above)**

$$\frac{Market\ price\ per\ 1M\ tokens\ (2024)\ reduced\ by\ the\ expected\ discount}{gonka\ (GNK)\ coins\ per\ 1M\ requests} = \frac{\$4.2}{4.7} \approx \$0.8936\ per\ gonka\ (GNK)\ coin$$

4. **Market valuation calculation**
   $Price\ per\ gonka\ (GNK)\ coin \times Total\ Coins\ issued\ by\ Year\ 3$
   $\$0.8936 \times 460,000,000 \approx \$411M$

5. **Fully Diluted Valuation (FDV) calculation**

$\$0.8936 \times 1,000,000,000 \approx \$894M$

The calculations highlight that gonka (GNK) coin could deliver comparable network functionality with significantly lower operational costs.

<u>Gonka network in Year 6 vs. Major centralized market player</u>

Scenario 2 - OpenAI by year 6				
Market Cap (in millions)			FDV (in millions)	
OpenAI	gonka (GNK) coin		OpenAI	gonka (GNK) coin
$157,000	$14,000		$157,000	$21,875
	-1024% Undervalued			-619% Undervalued

Difference in Market Cap: $157B (OpenAI) vs. $14B (gonka (GNK) coin).
Gonka (GNK) coin undervalued by 1024%

Difference in FDV: $157B (OpenAI) vs. $21.8B (gonka (GNK) coin).
Illustrates network's scalability potential.

Data inputs and assumptions for this analysis:

- **OpenAI Metrics (December 2024)**
  - Approximate request volume (in millions AI-tokens): 370M
  - Market Cap and FDV: $157B
  - Market price per 1M AI-tokens (2024): $10
- **Gonka assumptions (used in calculations below)**
  - Approximate discount expected by Developers in Year 6: 30%
  - Fees paid by Developers in gonka (GNK) coins (Year 6, not cumulative): 119M
  - Gonka (GNK) coins issued by Year 6: 6 batches x 50M + 60M (½ of Host liquidity pool) + 120M (Additional Host reward) + 160M Founders allocation = 640M

Based on these assumptions, the following mathematical calculations are performed:

1. **Adjusted market price per 1M tokens**
   Market price per 1M tokens (2024) — Discount $= \$10 - 30\% = \$7$

2. **Cost per 1M requests in gonka (GNK) coins to serve request volume similar to OpenAI**

$$\frac{Fees\ Paid\ by\ Developers\ (Year\ 6)}{Request\ Volume\ (M\ Tokens)} = \frac{119M}{370M} \approx 0.32\ gonka\ (GNK)\ coins\ per\ 1M\ requests$$

3. **Price per gonka (GNK) coin (based on calculations above)**

$$\frac{Market\ price\ per\ 1M\ tokens\ (2024)\ reduced\ by\ the\ expected\ discount}{gonka\ (GNK)\ coins\ per\ 1M\ requests} = \frac{\$7}{0.32} = \$21.875\ per\ gonka\ (GNK)\ coin$$

4. **Market valuation calculation**

$$Price\ per\ gonka\ (GNK)\ coin \times Total\ Coins\ Issued\ by\ Year\ 6 = \$21.875 \times 640,000,000 \approx \$14,000M$$

5. **Fully diluted valuation (FDV) calculation**

$$Price\ per\ gonka\ (GNK)\ coin \times Total\ Supply\ of\ gonka\ (GNK)\ coins = 21.875 \times 1,000,000,000 = 21,875M$$

The calculations highlight that gonka (GNK) coin could deliver comparable network functionality with significantly lower operational costs due to its decentralized approach.

5. Financial benefits for Developers and Hosts

5.1. For Hosts

- Early growth and high reward potential:
    a. Gonka (GNK) coins are in their early lifecycle, with significant network growth ahead.
    b. Hosts benefit from low valuation, accumulating more coins before demand scales.
    c. As inference requests grow, rewards will increase, tied directly to network adoption.
    d. When gonka (GNK) coins' FDV reaches $8.8B (Bittensor's level) by Year 3, early Hosts could see a 100x increase in earned coin value.
- Our Tokenomics is developed with hardware providers interests at its core:
    a. Direct token rewards for computational contribution:
        i. 84% of total gonka (GNK) coins (840M) allocated for Host incentives.
        ii. Compensation is based on actual computational work, ensuring fair rewards.
    b. Early liquidity support:
        i. 120M gonka (GNK) coins set aside for an early liquidity pool.
        ii. Hosts can exchange mined tokens for stable assets (USDT).
    c. 120M gonka (GNK) coins set aside as an additional reward for the TOP Hosts committing large-scale GPU power.
    d. Gradual subsidy reductions prevent over-inflation and ensure steady demand for gonka (GNK) coins.
- Optimal use of computational power and Hosts contribution to AI future:
    a. Gonka doesn't use Hosts' computational power (and electricity) in the absence of tasks. Computational power can be used on other platforms while there is nothing to compute on Gonka.
    b. Mining of gonka (GNK) coins is a way for individual hardware owners to contribute to and benefit from AI computational markets.
    c. gonka (GNK) coins help to diversify revenue streams with less speculative, therefore more stable demand for compute power.
    d. GPU crypto mining clusters have compute power comparable with the top AI companies and cloud providers. But a lot of crypto resources are used for token security vs useful computation. gonka (GNK) coins are optimised for computational outcome (not compromising security) the majority of computational power contributes directly to AI future.

5.2. For Developers

- Lower AI computation costs:
    a. Initial subsidies (up to 90%) significantly reduce inference and training expenses. For example, if a Developer spends X tokens on a request, Hosts may receive up to 9X in rewards as part of the subsidy.
    b. Costs remain lower than centralized providers like AWS, Azure, and OpenAI.
- Predictable and gradual cost adjustments:
    a. Well-structured subsidy reductions allow for long-term financial planning.
    b. Avoids sudden price hikes seen in centralized cloud services.
- Developers who fund or contribute to model training earn a share of future inference revenue.
- Early participation rewards:
    a. Developers using the network in its early phases benefit from lower-cost computational resources.
    b. Increased token value over time can enhance early adopters' financial returns.
- Easy to use API with no DevOps (similar platform.openai.com)
- Mitigates risk of big centralized AI platforms, their censorship and anti-competitive practices.

6. Risks and Considerations

While the proposed issuance model offers significant earning potential for Hosts and investors, particularly in early stages, its success depends on achieving adoption milestones, managing token value stability, and fostering trust through transparent governance.

Hosts must carefully weigh the potential benefits of early participation against the following risks:

1. Market volatility risk - the hypothesis relies on achieving significant Developer adoption and sustaining demand for inference services. Fluctuations in the token's market value may reduce the real-world value of rewards earned by Hosts and unpredictable earnings may deter long-term investments in infrastructure. Transparent communication and governance can reduce speculation and foster token stability.
2. Regulatory uncertainty risk - tokens issued as rewards could face regulatory scrutiny, especially in jurisdictions with strict cryptocurrency laws. The model assumes compliance with decentralized network principles and thus, no classification as a security in the United States and other jurisdictions with similar approaches to crypto asset regulation. As such, all figures and growth projections are theoretical and should not be construed as investment promises. Legal challenges may impact token issuance and create unforeseen compliance burdens for providers. Adherence to decentralized principles and proactive monitoring of legal frameworks may, but is not guaranteed to, minimize risk exposure.
3. Subsidy reduction over time - gradual reduction in Developer subsidies (from 90% in Phase 1 to 8% in Phase 12) could discourage Developer participation and lower Developer engagement may reduce network usage, impacting proportional rewards for Hosts. Adjusting subsidy reduction rates based on adoption metrics could align reward structures with actual growth.
4. Technological and market competition risk - centralized providers or competing networks may develop more efficient systems, reducing demand for decentralized alternatives. Providers could see reduced utilization of their infrastructure, however, differentiation through decentralization, transparency, and competitive rewards is key and can minimize risk exposure.

7. AI inference pricing model

AI inference pricing mechanism is built around a standardized unit of compute, allowing Hosts to estimate computational expenses in a universally accepted format. The system dynamically adjusts pricing through a decentralized voting mechanism, ensuring fairness and adaptability based on real-time supply and demand.

7.1. Key concepts of the pricing model

1. Unit of compute is a standardized measure representing the computational effort required for AI operations (inference and model training). It’s used to create a uniform pricing structure.
2. Prices can be set in multiple denominations (different scales of the network’s native currency):
   - icoin
   - uicoin (micro-coins)
   - micoin (milli-coins)
   - nicoin (nano-coins)
3. Market-driven pricing is determined by how many blockchain coins correspond to one unit of compute. To achieve a fair market price:
   a. Each Host submits a vote indicating how many coins per unit of compute they propose.
   b. The weighted median of these submissions is determined at the end of each epoch (when validators are rotated).
   c. That weighted median becomes the common price per unit of compute for the next epoch.
4. If a Host never votes, the most recently used epoch price (or the genesis parameter for the very first epoch) is used as their default vote.

7.2. Automated API for Price Management

Hosts can interact with the pricing system using an API that allows them to:

- Set or update price proposals
- Check their current price vote

- Query the latest global unit of compute price and model pricing
- Register new AI models

7.3. Dynamic Pricing for AI Models

Each AI model consumes a specific number of unit of compute per processed token. The pricing model ensures transparency by:

1. Listing models with their unit of compute price per token.
2. Setting a price per token derived from the unit of compute price.
3. Allowing new models to be proposed and voted on before they are approved for the network.

7.4. Decentralized Model Registration Process

1. Developers submit a model proposal with its estimated number of units of compute required for each token of that model.
2. The governance system, powered by the CosmoSDK, ensures that:
   a. A deposit is made to prevent spam.
   b. The proposal is reviewed by Hosts.
   c. Votes are cast to approve or reject the model.
3. Once approved, the model is listed in the pricing structure and available for computational tasks.

8. Hosts' compensation model for training new Large Language Models

The decentralized AI infrastructure is built on a foundational commitment to open-source development and accessibility. Advancing AI should not be controlled by centralized entities only. Instead, it must remain collaborative, transparent, and widely available to Developers and users worldwide. To uphold this commitment, the network guarantees that all models trained using its resources will remain open-source, ensuring that AI innovation remains decentralized and equitable. The network actively commits to continuously contributing to open-source AI development, setting it apart from centralized entities like Meta, which do not provide such guarantees. This commitment fosters motivation among contributors and ensures that innovation remains in the hands of a global community rather than a select few.

However, training new Large Language Models is one of the most computationally intensive processes in AI. Fair and sustainable compensation for Hosts is essential to keeping the decentralized AI network competitive and encouraging long-term participation.

To achieve this, the network adopts a Network-Financed approach, where 20% of all inference revenue (including transaction fees and mining rewards) will be allocated to fund the training of new models. This allocation will specifically cover the cost of the unit-of-compute used during training, ensuring that resources are efficiently utilized to support model development. Additionally, a portion of this 20% can be allocated as grants, awarded through a voting process, to support the most promising and iinovative traning procedures proposed by community. This mechanism ensures that the network continues to support large-scale AI advancements, even after initial subsidies phase out. Moreover, as the network grows, this approach allows for the collection of substantial funding, making it feasible to train massive AI models that compete with centralized alternatives.

The network will implement a consensus-based governance mechanism to maintain flexibility in adjusting the revenue allocation percentage. If new models significantly boost network adoption, consensus may choose to maintain or increase the 20% allocation. If the allocation impacts network competitiveness, consensus can reduce the percentage, ensuring a balance between affordability and innovation. This mechanism will only become available after Year 5, ensuring the network has demonstrated sufficient model training capabilities before modifications are allowed.

The decision on how to conduct the next training procedure follows a clear yet flexible process to ensure the most effective training approaches receive support. Rather than focusing solely on which model to train, the emphasis is on refining the training methodology. Contributors, including Hosts and Developers, can propose code changes (such as pull requests in a repository) that introduce new training strategies. These proposals undergo thorough discussion, with extensive presentations and debates to assess their feasibility and effectiveness.

Once a training approach gains consensus, it is first approved as a code update. Following this, the same Host can formally propose a full training experiment, detailing the parameters, dataset, and allocated funding. This proposal then enters a governance voting phase, where it is likely to undergo multiple iterations, including rejections and modifications before reaching final approval. The network ensures that the best AI research and development teams worldwide can propose and experiment with promising training techniques, even if they lack their own hardware. This ensures that innovative research is not limited by computational constraints while maintaining full transparency (everyone can still access and evaluate R&D results, fostering continuous innovation). Priority is given to training experiments that demonstrate strong potential to maximize impact, ensuring resources are directed toward AI advancements that offer the greatest value. This iterative process fosters innovation, aligns the network's growth with user needs, and ensures that training methodologies continue to evolve in the most effective way possible.

Why the Decentralized AI Training Fund? Various approaches were considered for financing LLM training. The Decentralized AI Training Fund was chosen due to its sustainability, fairness, and ability to encourage meaningful participation from Hosts.

Financing Model	Advantages	Challenges
Decentralized AI Training Fund (chosen)	● Ensures stable, long-term funding; ● Predictable rewards for Hosts; ● Scalable as network revenue grows	● Requires governance to adjust revenue allocation
New Coin Issuance	● Immediate incentives for providers; Allows early growth	● May lead to inflation and reduced token value; Unclear incentive alignment
Shareholding Structure	● Hosts receive compensation through additional fees paid by Developers, along with a proportional share of the network reward. ● Hosts receive a stake in the model and are compensated whenever it is used. ● Creates long-term incentives to support training.	● Developers using trained models may pay an additional fee. ● Additional usage fees may be at a disadvantage against free alternatives like LLaMa (funded by Meta). ● Network may subsidize the cost, though this could result in slightly lower earnings for Hosts. ● Technical challenges related to the rights for the model weights, which could lead to security threats and potential attacks targeting model ownership.

The Decentralized AI Training Fund ensures long-term funding for LLM training, balancing sustainability with competitive pricing. By allocating 20% of inference revenue to model training, the system secures billions in funding while allowing for governance-based adjustments in later years. This approach benefits

Hosts and Developers fostering a scalable, decentralized AI ecosystem that can compete with centralized cloud solutions while maintaining fairness and efficiency.

Host-financed training relies on direct funding from Hosts and Developers, creating a more demand-driven approach. While this model requires users to actively contribute to financing, potentially limiting participation, it remains a viable option for those who choose to fund training independently. Additionally, while there is a risk of model duplication and free redistribution, this approach offers flexibility for individuals or organizations willing to invest directly in model development.