Concept Paper

Intelligence Cubed: A Decentralized Modelverse for Democratizing AI

Jade Zheng [†] , Fernando Jia [*,†] , Florence Li [†] , Rebekah Jia , Tianqin Li [*,†]

Posted Date: 20 June 2025

doi: 10.20944/preprints202506.1717.v1

Keywords: decentralized AI; blockchain; machine learning models; modelverse; initial model offering (IMO); proof of intelligence (PoI); web3; AI democratization; model marketplace; tokenization; AI governance; create-to-earn; model composition; distributed computing; DePIN (Decentralized Physical Infrastructure Network)

Concept Paper

Intelligence Cubed: A Decentralized Modelverse for Democratizing AI

Jade Zheng [1,2,†], Fernando Jia [1,3,*,†], Florence Li [1,4,†], Rebekah Jia [1,5] and Tianqin Li [6,*,†]

[1] Intelligence Cubed
[2] Duke University
[3] University of California, Berkeley
[4] Stanford University
[5] Georgetown University
[6] Carnegie Mellon University
[*] Correspondence: fernando.jia@intelligencecubed.com (F.J.); tianqinl@cs.cmu.edu (T.L.)
[†] Equal Contribution.

Abstract

The rapid advancement of artificial intelligence (AI) has been largely characterized by centralized development and control, limiting accessibility and innovation. This paper introduces Intelligence Cubed (I Cubed), a decentralized, open-source "modelverse" designed to democratize the creation, distribution, and utilization of machine learning (ML) models. I-Cubed aims to establish a community-driven ecosystem where ML model developers and AI creators can collaborate, monetize their contributions, build reputation, and engage in novel co-creation using a spectrum of techniques from prompt-level conditioning to advanced model composition via task arithmetic. The platform leverages blockchain technology to ensure transparency, immutability, and fair governance. Key mechanisms such as Proof of Intelligence (PoI) are proposed to validate model originality and performance, while Initial Model Offerings (IMOs) facilitate early-stage funding and community engagement. By fostering a decentralized marketplace and integrating distributed compute resources, I-Cubed seeks to lower entry barriers for developers, provide users with access to a diverse range of specialized AI models, and collectively advance the pursuit of Artificial General Intelligence (AGI) through a more open and collaborative paradigm.

Keywords: ecentralized AI; blockchain; machine learning models; modelverse; initial model offering (IMO); proof of intelligence (PoI); web3; AI democratization; model marketplace; tokenization; AI governance; create-to-earn; model composition; distributed computing; DePIN (decentralized physical infrastructure network)

Contents

1. Preface

The AI market suffers from extreme centralization, with companies like OpenAI and others capturing a significant share of users. For instance, OpenAI's ChatGPT alone is estimated to hold approximately 60% [1] of the conversational AI user base as of early 2025, with over 400 million [2] weekly active users. When people think of AI, their instinct is to turn to ChatGPT or similar offerings from tech behemoths, overshadowing specialized models painstakingly developed by small teams or independent researchers for niche applications. As a founding team—having published research on deep neural architectures and built machine learning models ourselves—we have experienced this frustration firsthand: the market's bias toward established players often leaves innovative, tailored solutions unnoticed.

Training a competitive language model demands immense resources, creating a formidable barrier to entry. For example, OpenAI's GPT-4o, an advanced multimodal model, is rumored to have cost upwards of $100 million [3] to train, factoring in computational resources, data acquisition, and engineering efforts. Similarly, DALL-E 2, another Transformer-based model from OpenAI introduced by [4], boasts 12 billion parameters and was trained on over 400 million captioned images, requiring significant computational power. While OpenAI bore the costs of training DALL-E, they controversially decided against open-sourcing the model, meaning the code and architecture are not publicly available. Smaller models remain inaccessible to most, A 7B parameter LLM requires about 100,000 GPU hours Please check the refs part and revise them to make the all refs citation format display correct. Please ensure all the references so that they appear in numerical order. [5] ($150,000 on Nvidia A100). These high resource demands disproportionately favor well-funded organizations, sidelining independent innovators.

However, a shift began in January 2025 with the emergence of DeepSeek, a Chinese company that disrupted the AI landscape. DeepSeek unveiled their MoE model (DeepSeek V3 [6]) achieving GPT-4 level reasoning at approximately 5.5% [7] training cost. Their FlashMLA improves the performance of NVIDIA H800 AI chips by 8 times [8]. Inspired by this breakthrough, an explosion of efficient training techniques and products follows: Stanford's LIRE framework enables 26-minute LLM training [9] via dynamic architecture search. A proliferation of derivatives and adaptations fostered by DeepSeek's open-sourcing of its models, with communities rapidly building on its frameworks [10]. The rise of edge AI products, spurred by DeepSeek's efficient models that can be deployed on consumer hardware, enabling sophisticated AI to run locally without cloud dependency [11].

The U.S. AI market is shocked. These breakthroughs signal the dawn of low-cost AI development and suggest an impending explosion in affordable AI models, driven by open-source tools and cost-effective training methods. Despite DeepSeek's promise for developers, the consumer market lags behind. A predominance of AI users still relies on subscription-based large models, spending at least $20 monthly, with an estimation of 20-30% subscribing to multiple services [12].

Meanwhile, 40% of surveyed businesses report a lack of AI solutions [13] tailored to their needs, highlighting a disconnect between supply and demand. No mature platform yet exists to bridge this gap, allowing users to "vote with their feet" and connect developers with those who need their models. While open-source platforms like Hugging Face foster collaboration and learning, they fail to provide creators with direct revenue. Our interviews with PhD students from leading universities revealed a common reluctance to open-source models they've invested significant time and computational resources into, due to the absence of financial incentives.

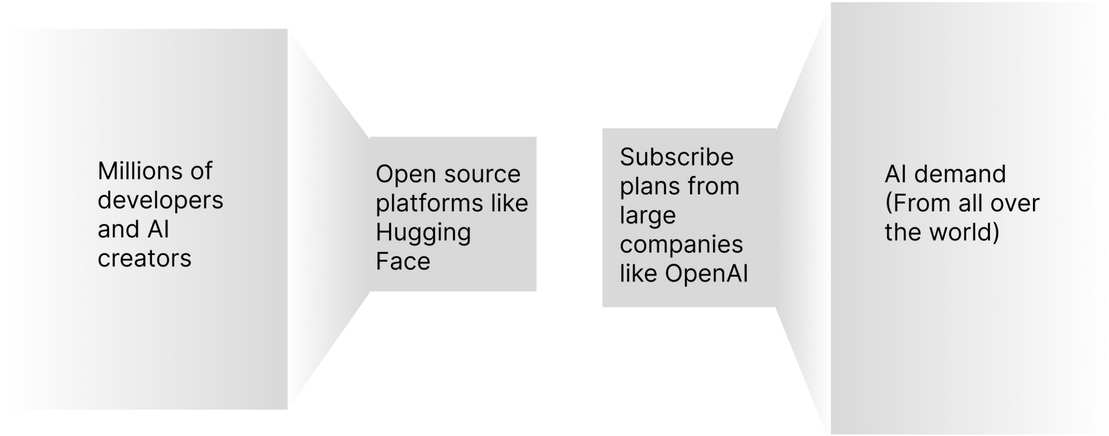
Millions of developers and AI creators
Open source platforms like Hugging Face
Subscribe plans from large companies like OpenAI
AI demand (From all over the world)

Figure 1. On the left are millions of AI model developers with limitless creativity, but the only way for them to gain recognition is by publicly releasing their work on open-source platforms. On the right, a handful of corporations control the global supply and demand of AI. Consumers cannot access models tailored to specific vertical needs directly, creating a gap between developers and users due to the dominance of these large companies.

OpenAI's 2019 pivot to a for-profit model, followed by 2023's leadership crisis exposing internal clashes between commercialization and safety, starkly deviated from its original nonprofit mission of open, ethical AI for all. This corporate turbulence underscores the risks of centralized control over AI development, contrasting with our web3 product's vision: decentralized, open-source models that democratize access and uphold transparency, aligning with OpenAI's founding ideals.

Decentralized blockchain technology offers a potent solution to these challenges. Unlike traditional AI models controlled and trained by giants, deploying models on a decentralized network enables developers and users to collaboratively train, improve, and govern them. This reduces reliance on centralized entities, lowers barriers to entry, and creates a transparent, equitable ecosystem where value is shared among contributors. By tokenizing AI models and leveraging community-driven mechanisms, I Cubed aims to unlock this potential, redefining how AI innovation is nurtured and rewarded.

2. Problem

As the AI ecosystem evolves toward modularity and decentralization, the creation of an open marketplace for AI models becomes both a promising opportunity and a complex engineering challenge. Such a marketplace must not only overcome typical platform bootstrapping issues—like the cold-start problem and liquidity imbalances—but also address domain-specific concerns such as model verification, provenance tracking, and equitable evaluation. Unlike traditional software platforms, AI model marketplaces must integrate usage-based incentives, trustless quality assurance, and dynamic benchmarking systems that reflect real-world deployment contexts. This section outlines the core infrastructural and design challenges in building a decentralized AI model marketplace and proposes directions to ensure fairness, trust, and sustainability in this emerging ecosystem.

2.1. Marketplace

A marketplace for AI models encounters the same supply-and-demand hurdles as any emerging platform, amplified by the specialized characteristics of AI. A primary challenge is the cold-start problem: for the marketplace to grow, there must be sufficient liquidity—a balanced supply of models, and demand from users—from the very beginning. To attract AI model creators and stimulate participation, the platform must offer clear usage-based rewards. This requires tracking model usage accurately and distributing proportional payments to providers, incentivizing them to contribute high-quality models.

2.2. Trustless Verification Complexities

To establish a trustless network with robust economic incentives, the marketplace must ensure that models are genuinely used and deliver high-quality results. In the absence of centralized oversight, verifying model performance and usage becomes a significant challenge. Without an effective verification system, Creators could submit low-quality or untested models, undermining the platform's credibility; Users might falsely report usage to manipulate rewards, disrupting the economic model; The lack of quality control could flood the marketplace with substandard contributions, reducing its overall value.

A reliable mechanism to measure usage (e.g., through tracked interactions like API calls or downloads) and assess quality (e.g., via performance metrics or user feedback) is critical to maintaining trust and ensuring rewards reflect real contributions.

2.3. Identity Gap in N-Creation

Current Web3 platforms lack a robust mechanism for N-creation recognition—accurately tracing and attributing multi-generation derivatives of AI models—because decentralized identifiers (DIDs) and self-sovereign identity standards remain sparsely adopted. In practice, provenance chains fall back to opaque wallet addresses [14], hindering reliable attribution across successive creations and impeding any systematic accounting of contributor lineage.

2.4. Neutral Evaluation

Exceptional AI models deserve sufficient exposure—especially to their intended audience—and fair assessment of their value. In many platforms, visibility is dictated by algorithms or editorial decisions, which can introduce bias and favor well-known contributors. To address this, a neutral voting system is necessary, enabling impartial evaluation of models based on merit. Without such a system, high-quality models from emerging creators may remain unnoticed, reinforcing the dominance of established players; Unfair or opaque evaluations could distort perceptions of model quality, discouraging participation; Lack of transparency erodes trust, a vital component of any marketplace.

A community-driven, neutral evaluation process ensures that models are judged fairly, promoting diversity and innovation across the platform.

2.5. Limitations of Conventional AI Benchmarking

Mainstream evaluation relies on static benchmarks—e.g., GLUE [15] for language and ImageNet [16] for vision—that measure a model's stand-alone performance on isolated tasks. These vertical metrics were sufficient when models were deployed singly, but they falter in today's multi-model era:

First, static benchmarks such as GLUE measure isolated task accuracy but ignore a model's systemic contribution to end-to-end workflows. Models that score near the GLUE ceiling frequently underperform in production settings demanding richer context and multi-step reasoning [17], which led to the stricter SuperGLUE suite. Yet even SuperGLUE [18] remains far from human parity; meaningful progress will require evaluations centred on multi-task transfer and self-supervised adaptation rather than single-task scores alone. Second, existing tests cannot quantify peer-workflow compatibility: how efficiently a model interoperates [19] with others once standards such as the Model Context Protocol (MCP) [20] enable large-scale composition. Third, because most benchmarks are detached from deployment data, their results create no feedback loop [21] for fine-tuning or post-training, especially for proprietary models. Consequently, the classical paradigm both mis-prices a model's collaborative utility and impedes its iterative improvement.

2.6. Related Work and Gaps

Several blockchain-based platforms have begun tokenizing AI models but only solve fragments of the create-to-earn equation. AIOZ Network [22] tokenises models and meters on-chain usage, yet rewards stop at first-order calls and ignore downstream derivatives. Ritual [23] "enshrines" model provenance on-chain but emphasises integrity proofs over price discovery, offering neither workflow benchmarks nor incentive-weighted ranking. ReelMind [24] tokenises output licences, yet its ledger does not clarify how royalties propagate when a model is fine-tuned or merged into successors. In sum, lineage-aware revenue sharing, objective cross-workflow benchmarking, and permissionless composability remain missing primitives—gaps that I Cubed Modelverse directly addresses.

3. Solution

3.1. Overview

I Cubed is building a **decentralized, community-driven marketplace** for AI model development and usage, powered by **blockchain technology** for transparency and democracy. It offers AI model creators a platform to earn income and reputation through their creation and partial ownership transfer, while the users can get rid of prevalent expensive subscriptions from large companies, search for the best models in niche areas, only pay for their usage, and stake the models they look to promising future. Governance is managed by the DAO to ensure **trust and fairness** for all participants.

For Model Creators:

Unlike Web2 platforms where creators work under centralized ownership without recognition or sustainable income, I Cubed provides a Web3-native infrastructure where creators retain ownership and tokenize their AI models as digital assets. Developers can: 1)Deploy fully on-chain autonomous AI models validated via Proof of Intelligence. 2)Monetize through create-to-earn rather than "share-for-free". 3)Retain partial model ownership while transferring shares to fund development. 4)Benefit from exposure through a community-curated recommendation engine, promoting model discovery. 5)Compete fairly in an open, DAO-driven AI benchmark that highlights performance through community voting.

For Community Users:

Community users can: 1)Discover and use high-quality, niche models without committing to large, flat-fee subscriptions. 2)Stake tokens in models they believe in—earning returns if the models

doi:10.20944/preprints202506.1717.v1

perform well. 3)Remix, fork, and experiment with open-source models, contributing to an evolving ecosystem. 4)Participate in democratic governance by voting on benchmarks, model evaluations, and development proposals. 5)Become active co-creators rather than passive consumers.

3.1.1. Layers of I Cubed

The I-Cubed platform is built on a multi-layered architecture to create a decentralized AI Modelverse. The Modelverse System consists of 4 layers. The **Control Layer** provides the on-chain foundation for governance and ownership, representing models with unique decentralized identifiers (DID) and managing royalties and derivatives. The **Service Layer** enables innovation through an off-chain orchestrator where users can compose complex AI pipelines and mint them as on-chain secondary creation, i.e., workflows. These models and workflows rely on the **Storage Layer**, which uses decentralized networks like IPFS and Filecoin to persistently store model checkpoints and intermediate artifacts off-chain. Finally, the **Execution Layer**, a decentralized physical infrastructure network (DePIN) of GPU nodes, securely executes these multi-step pipelines, with the orchestrator scheduling tasks and passing data between steps via DIDs from the Storage Layer.

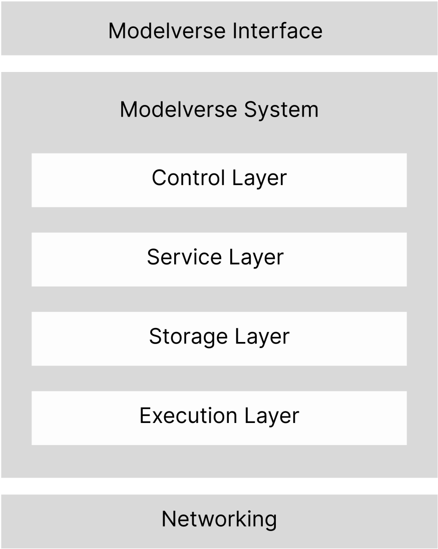

Figure 2. Layers of I Cubed

3.1.2. Key Workflows

Upon uploading model weights to IPFS and recording the accompanying metadata on-chain, the developer triggers a smart contract that mints a fixed supply of ownership tokens—49% immediately allocated to the creator and 51% reserved for public acquisition. Community participants may deposit funds into this pool; once aggregate staking reaches the 51% threshold, the contract releases the model's IPFS hash under an open-source licence and escrows the staked tokens for a predefined vesting period (or until a subsequent financing event). In addition, any user may invoke the model by paying a usage fee, which the contract distributes proportionally among all token-holders.

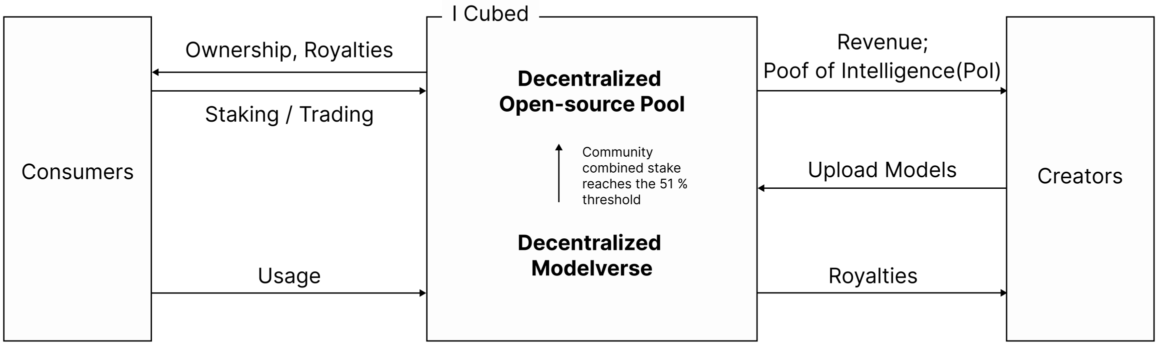
I Cubed
Consumers
Ownership, Royalties
Staking / Trading
Usage
Decentralized Open-source Pool
Community combined stake reaches the 51 % threshold
Decentralized Modelverse
Revenue; Poof of Intelligence(PoI)
Upload Models
Royalties
Creators

Figure 3. Modelverse Common Workflow: Creators upload models to the I Cubed Modelverse; consumers access them via usage fees or by staking to gain ownership. Once 51% of a model is publicly owned, it becomes open-source.

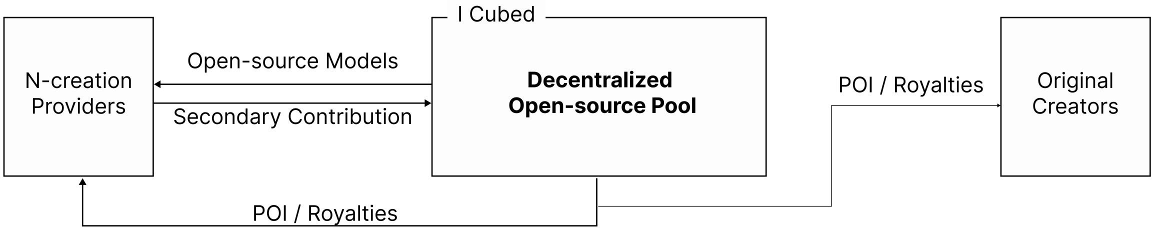

Figure 4. Secondary/N-Creation Workflow: Creators can build their own creations based on existing open-source models. Royalties from these derivative works are distributed proportionally to all contributing original creators.

3.2. Decentralized AI modelverse

3.2.1. Architecture

Control Layer (Asset Provenance and Governance) — Serves as the authoritative on-chain ledger for model identity, ownership, and intellectual-property entitlements. Core elements include: (i) decentralised identifiers (DIDs) that immutably bind models to their creators; (ii) smart-contract suites governing IMO thresholds, transfers, and royalty flows; (iii) a JobManager contract that escrows and releases compute payments; and (iv) a token-based governance module enabling community stewardship over protocol parameters.

Service Layer (Composition and Orchestration) — Facilitates the creation of complex, composite AI systems from discrete models. It abstracts the complexity of workflow design through an off-chain orchestration engine coupled with an on-chain registry, effectively transforming computational pipelines into transferable digital assets. Most importantly, MCP Orchestrator provides an abstracted interface for workflow construction, allowing users to define multi-step computational pipelines by logically connecting individual AI models.

Storage Layer (Decentralized Data Persistence) - Leverages off-chain, decentralized storage networks for data persistence. This layer is responsible for storing the artifacts required for the execution and verification of computational workflows. Data is content-addressed using DIDs, enabling the Execution Layer to deterministically retrieve the precise inputs and outputs for each step.

The Execution Layer - functions as the system's computational substrate, comprising a Decentralized Physical Infrastructure Network (DePIN) of globally distributed GPU nodes. This layer provides the resources necessary for the secure and efficient processing of complex AI pipelines. On-chain ZK-ML proofs or attestation hash will be used to verify the referenced weights.

doi:10.20944/preprints202506.1717.v1

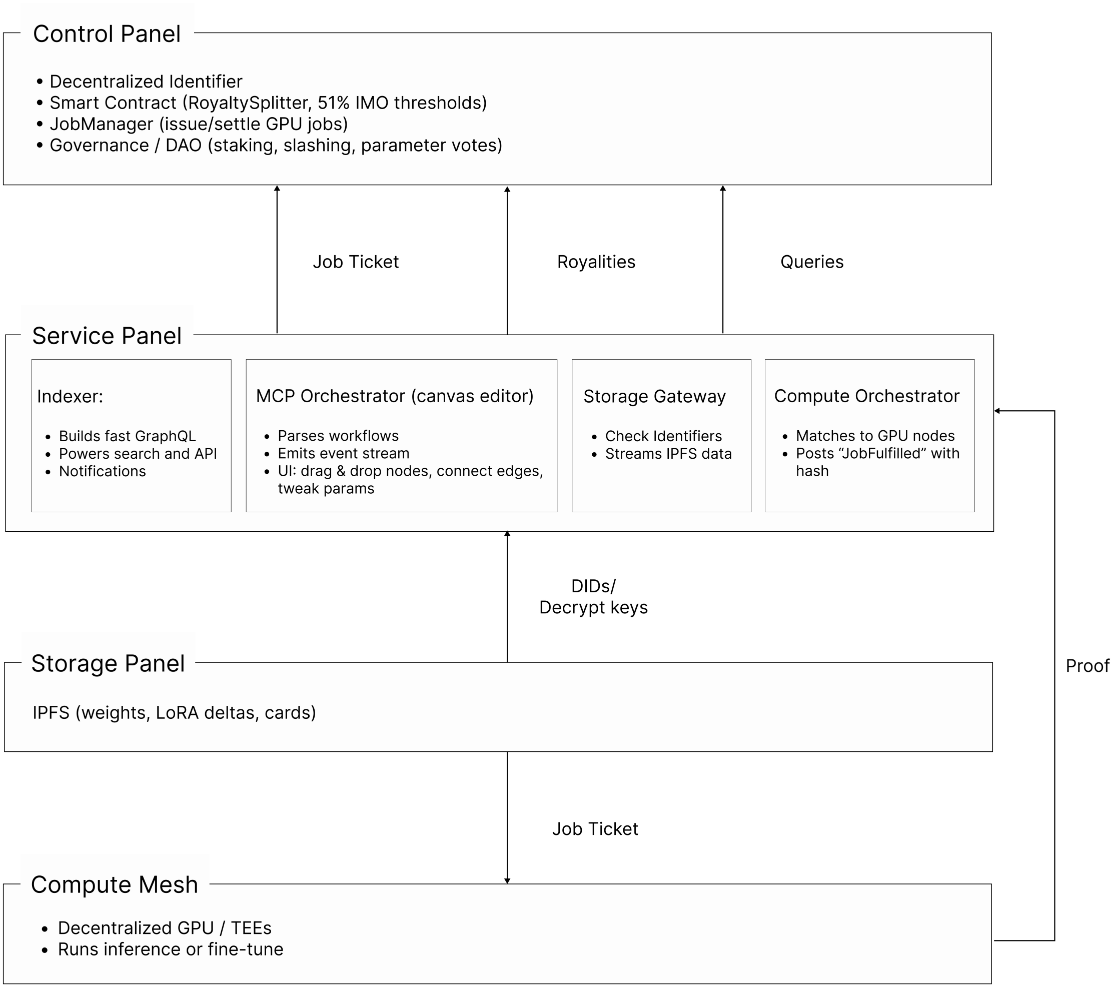
Control Panel
• Decentralized Identifier
• Smart Contract (RoyaltySplitter, 51% IMO thresholds)
• JobManager (issue/settle GPU jobs)
• Governance / DAO (staking, slashing, parameter votes)
Job Ticket
Royalities
Queries
Service Panel
Indexer:
• Builds fast GraphQL
• Powers search and API
• Notifications
MCP Orchestrator (canvas editor)
• Parses workflows
• Emits event stream
• UI: drag & drop nodes, connect edges, tweak params
Storage Gateway
• Check Identifiers
• Streams IPFS data
Compute Orchestrator
• Matches to GPU nodes
• Posts "JobFulfilled" with hash
DIDs/
Decrypt keys
Proof
Storage Panel
IPFS (weights, LoRA deltas, cards)
Job Ticket
Compute Mesh
• Decentralized GPU / TEEs
• Runs inference or fine-tune

Figure 5. I Cubed Architecture

3.2.2. From Open-Source Visibility to Incentive-Aligned Value Creation

I Cubed adopts a Hugging Face–like model space framework while fundamentally rethinking its architecture through decentralization and tokenized ownership. Traditional Web2 platforms often rely on creators voluntarily open-sourcing their work to gain visibility or handing over ownership for compensation. In contrast, I Cubed ensures that creators retain intellectual property rights and earn income through usage-based rewards—creating long-term incentives for sustainable innovation and community self-governance.

This approach transforms the "open-source for exposure" model into a "create-to-earn" ecosystem. As models perform well, they naturally attract more usage and community-driven staking to unlock open-source access. Developers not only gain visibility and reputation but also earn from each usage, unlike on platforms like Hugging Face where models are shared without direct monetization.

At the same time, consumers benefit from pay-per-use access to niche, high-quality models—without committing to expensive multimodal subscriptions from tech giants or overpaying for one-time model access.

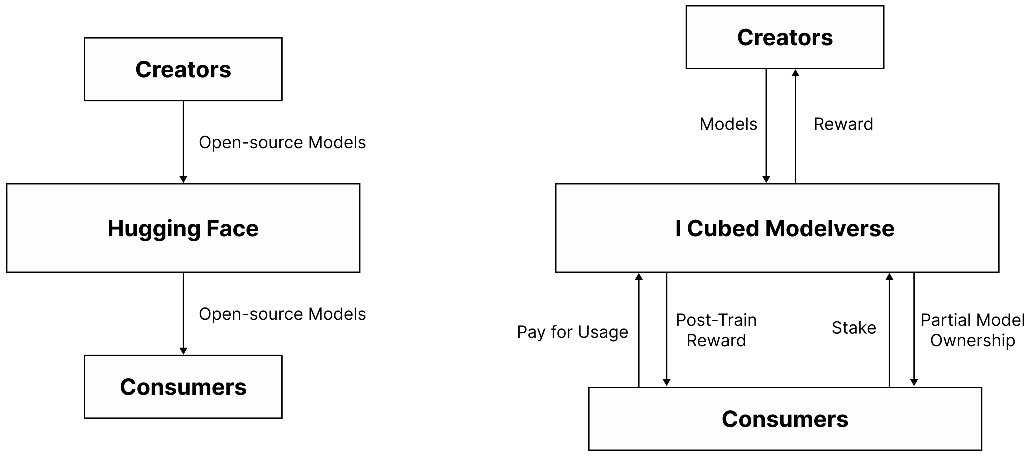

Figure 6. Key Differences from Hugging Face

Feature	Hugging Face	I Cubed Modelverse
Incentivization	Developers often open-source without direct rewards	Developers and creators earn from usage, staking, and remixing activities
Ownership	No native ownership mechanism	Blockchain-based model ownership and transferable token stakes via IMO
Developer Value	Exposure and community feedback	Royalties, usage fees, and community recognition via on-chain tracking
Consumer Value	Free access but limited incentive feedback loop	Pay-per-use pricing and potential post-train rewards for high-impact data contributions

3.3. Initial Model offering

I Cubed introduces a novel **Initial Model Offering (IMO)** mechanism that enables AI models to be offered, traded, and collectively owned—much like early-stage equity in a startup. This process not only helps creators monetize their work early but also aligns community incentives around promising AI systems.

3.3.1. Process

A model creator begins by uploading a model to the I Cubed Model-Verse and launching an Initial Model Offering (IMO). After setting an indicative token price—augmented by I Cubed's recommendation engine, which benchmarks traffic patterns and comparable assets—the IMO pool opens for staking. Users reserve ownership shares at this fixed price; under an anti-rollback rule, staked tokens may be traded among participants but cannot be withdrawn. When cumulative stakes reach the 51 percent threshold, the model enters a one-day soft-lock that ends at 00:00 PT on the following business day, after which the weights are open-sourced to IPFS and pricing becomes market-driven. From that point onward, model tokens trade freely in a secondary market analogous to post-IPO equity trading, allowing participants to enter, exit, or consolidate positions as demand evolves.

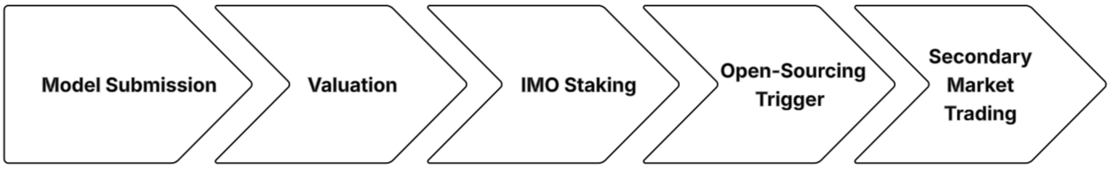

Figure 7. IMO Process

3.3.2. Model Valuation Recommendation Engine

Each model entering the IMO pool requires an **initial valuation** set by the creator. To guide this decision, I Cubed offers a recommended price range using a **dynamic valuation algorithm** inspired by secondhand marketplaces and recommender systems.

$$P_{model} = \alpha(\hat{D}_{pre}) + \beta(P_{similar\ market}) + \gamma(R_{creator}) \tag{1}$$

where α, β, and γ represent positive correlation functions that map three key inputs to the price space:

- $\hat{D}_{pre}$: predicted pre-IMO demand (inferred from page views, API trials, and wishlist activity),
- $P_{similar\ market}$: the average token price of functionally similar models, and
- $R_{creator}$: the creator's historical reputation or track record of successful model launches.

3.3.3. Anti-Rollback & Lock-in Rules

To maintain pricing integrity and prevent speculative gaming, the IMO process enforces a non-refundable participation policy: During the IMO window, users can freely trade their pre-staked shares (at a fixed price) with others. However, redemptions or refunds to the IMO contract are strictly prohibited—ensuring capital commitment is real.

3.3.4. Post-IMO Market Dynamics: Equity-like Ownership & Liquidity

Once the 51% ownership threshold is reached, the stake pool is locked until 00:00 PT on the same business day. At 00:00 PT on the next business day, the model is open-sourced via IPFS and verified by the platform and unlocked for full trading, enabling secondary ownership transfers.

Stakeholders can now: 1)Trade their shares freely with others on the marketplace. 2)Offer partial sales (like equity vesting schedules or unlisted option grants). 3)Benefit from price appreciation as usage, exposure, and community trust grow. This mimics a pre-IPO equity structure where stake tokens function as non-public equity pre-open-source. Once open-sourced, the tokens become liquid and act like public stock.

3.4. Proof of Intelligence

Proof of Intelligence (PoI) is the cornerstone of I Cubed's decentralised trust framework. Whenever a model or multi-step workflow is uploaded to the I Cubed Model-Verse, it is assigned a unique decentralised identifier (DID) that immutably records its provenance, ownership and subsequent derivatives. The platform actively promotes secondary creation: through the I Cubed Canvas interface, researchers can frictionlessly fine-tune, remix, or combine existing models to generate new artefacts while retaining clear attribution. PoI fulfils two complementary functions. It operates as a canonical ledger that anchors every intelligence asset, together with its entire lineage, on-chain; and it enables a tokenised market in which usage events, performance metrics and royalties are automatically routed to all upstream contributors. Each time a model is embedded in a composite workflow, the downstream creator assigns a peer-workflow compatibility score that reflects how smoothly the component interoperates with other models.

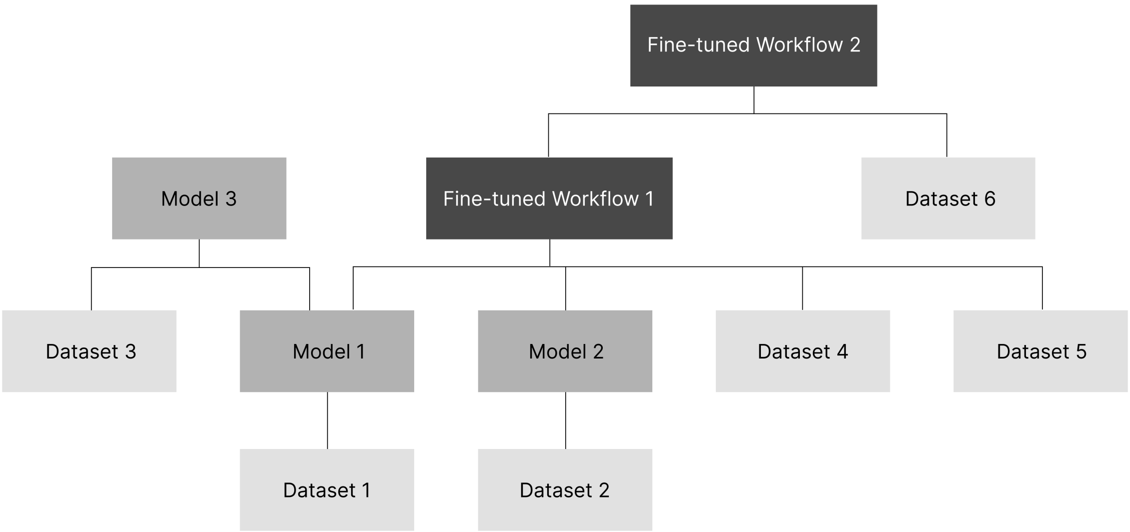

Figure 8. This diagram illustrates a basic secondary-creation process in which a creator combines existing models, fine-tunes them with custom data, and thereby produces a new workflow.

For example, Creator A uploads a weight file that becomes DID-Model-1. Creator B then combines DID-Model-1 with a LoRA adapter to construct a novel pipeline, registering it as DID-Workflow-1. Buyer C invokes this workflow via a one-off API call; decentralised GPU nodes execute each step, return outputs, and publish ZK-ML proofs. The royalty is released to Creator B and Creator A by a certain percentage.

3.5. Neutral Evaluation & Democratic Pricing

Evaluating AI models across diverse domains is inherently difficult, especially when comparing niche single-purpose models with large, multimodal systems. Multimodal models often gain outsized popularity due to their broad utility, overshadowing specialized models that perform exceptionally well in focused tasks but lack mass visibility. To counter this imbalance, the I Cubed modelverse introduces a diversified recognition and evaluation framework.

Instead of simply asking, “How many people use Model X?”, we propose a more meaningful question: “How many critical outcomes uniquely depend on Model X?” This reframing highlights functional importance over superficial popularity. In a well-functioning system, there should be a mechanism where deriving a result is complex, but verifying it is simple. Futarchy[25], a voting mechanism that was originally introduced by Robin Hanson as "vote values, but bet beliefs", fits naturally in this decentralized system.

(Briefly explained: this mechanism selects a set of goals—which can be any measurable metrics—and combines them into a target metric M. When a decision needs to be made (say, a YES/NO outcome), three prediction markets are opened:

1. whether YES or NO will be chosen;
2. the value of M if YES is chosen (otherwise 0);
3. the value of M if NO is chosen (otherwise 0).

From these, the system can infer which decision the market believes will be more beneficial for M.)

Thus, we bring all community participants into this system to vote values and bet beliefs. I Cubed’s crowdfunding mechanism aims to let the network transparently and fairly evaluate your impact. Once a model enters the Initial Model Offering (IMO) pool, 51% of ownership is opened for public staking. Anyone from the community can stake on models they personally believe to be promising. When 51% has been fully staked, ownership is temporarily locked, the model is open-sourced, gains broader exposure, and stakeholders can further trade their stakes and earn dividends.

doi:10.20944/preprints202506.1717.v1

The more people rely on and recognize your model, the more likely it is to be open-sourced through community crowdfunding—and its valuation will rise with market expectations. This approach ties model evaluation to both market activity and democratic consensus, turning pricing power into a collective decision.

How does this mechanism reward and encourage "correct" behavior? Those who stake early on models that later prove impactful will gain exposure, dividends, and token appreciation. Those who stake on poor-performing or low-utility models will see minimal returns or incur opportunity costs. This mechanism creates built-in incentives for truthful signaling and discourages speculation or manipulation.

Our democratic benchmark system ensures that well-performing vertical models receive proper exposure. Every niche model can be surfaced and discovered—not just high-profile general-purpose models. This prevents vertical excellence from being neglected in favor of popularity alone.

3.5.1. Model-Peer Benchmark: A Three-Tier Ecosystem Valuation Framework

To translate this democratic evaluation into a transparent pricing and discovery mechanism, we adapt the peer-benchmarking methodology long used in business analytics and introduce the Model-Peer Benchmark (MPB) framework, which evaluates an AI model not in isolation but as a participant in an ecosystem of peer models. Economic value is inferred from both lateral comparison and collaborative performance. MPB comprises three tiers: 1)Vertical Indices – Each model domain (e.g., NLP, Computer Vision) produces its own AI-Vertical Index, serving as a performance benchmark and tradable market reference. 2)In-Category Leaderboards – Models within the same domain are ranked based on a weighted score combining user feedback, usage volume, and community reputation. New models are granted enhanced visibility during an initial "Starter Pool" period, avoiding "winner-takes-all" dynamics and fostering ecosystem diversity. 3) Peer-Workflow Compatibility (PWC) — quantifies a model's empirical performance when integrated into multi-model pipelines, thereby measuring its suitability for downstream composition and secondary use.

Vertical Classification and Industry Indices

Each major category will generate an AI-Vertical Index (e.g., AI-NLP Index, AI-CV Index), allowing developers to benchmark model performance within a field, investors to track sector trends or hedge across verticals, and the platform to issue ETF-style model bundles or perpetual contracts.

In-Category Model Ranking

Each model receives a composite score used in search, discovery, airdrop targeting, and reputation signals. For example:

$$Socre_i = 60 * R_i + 30 * U_i + 10 * S_i \tag{2}$$

where:

- R_i : user rating
- U_i : usage volume
- S_i : social signal or citation frequency

Weights can be fine-tuned via A/B testing or Bayesian optimization. Metrics are gathered daily, composite scores and rankings recomputed weekly; models with greater than ±20% shifts receive manual audit, and each weekly release publishes public leaderboards (top-100 global and per-vertical). Newly listed models enter a Starter Pool with boosted exposure and curated discovery support. This helps promising but unknown models surface quickly without being drowned out by incumbent volume.

Peer-Workflow Compatibility

By mining workflow logs recorded under the Model Context Protocol (MCP)—optionally verified with privacy-preserving zero-knowledge ML—we quantify each model's marginal contribution to multi-model collaboration. The core metric is:

$$PWC_i = \alpha * StandaloneAccuracy_i + \beta * CollaborationSuccessRate_i + \gamma * DownstreamGain_i \tag{3}$$

where:

- PWC_i: Final Peer-Workflow Compatibility score for model i.
- $StandaloneAccuracy_i$: Model i's conventional benchmark score (Tier 1).
- $CollaborationSuccessRate_i$: Fraction of workflows involving model i that complete end-to-end tasks, derived from MCP logs.
- $DownstreamGain_i$: Observed lift in business KPIs (e.g., conversion or throughput) when model i is included in a pipeline.

α, β, and γ are tunable. A "sharpshooter" model may receive a higher α, whereas a "glue" model—mediocre in isolation yet critical for orchestration—earns its edge through larger β and γ. Bayesian optimisation on market feedback can update these weights, ensuring the metric remains dynamically fair.

This framework is the first to integrate a model's solo competence with its team utility, providing a quantifiable basis for economic incentives.

4. Decentralization

Web3 supplies the missing institutional primitives—immutable property rights, frictionless liquidity, permissionless governance, transparent benchmarking, automatic multi-generation attribution, and open composability—that a create-to-earn model for AI unequivocally requires yet Web2 cannot deliver. Tokenized identifiers convert every algorithm into a tradeable, provenance-secured asset; smart contracts clear royalties and secondary sales in real time; DAOs replace platform gatekeepers with stakeholder rule-making; public ledgers record both standalone accuracy and workflow contribution, preventing opaque ranking bias; and lineage-aware protocols distribute value to all upstream contributors while inviting downstream remix.

4.1. Why Web3 Is the Only Viable Path for Create-to-Earn

If we truly want creators and developers—from any domain—to create and earn through AI model sharing and innovation, **Web3 is not just an option—it's a necessity**. The Web2 framework is fundamentally flawed for this use case. Earnings are often locked within walled platforms, where withdrawal is restricted, slow, or taxed by excessive fees. For most creators, liquidity is low, value is non-transferable, and monetization is at the mercy of gatekeepers. We see this failure across creative sectors. In both gaming and AI, developers dedicate time and expertise to build impactful products, only to receive non-cashable points or exposure that never translates into real income. Without reliable monetization, even great projects can't survive—and neither can their communities. Web3 introduces a fundamentally different system. Through tokenized ownership, permissionless earning, and portable value, creators are empowered to earn with clarity, autonomy, and aligned incentives. It's not just about making money—it's about restructuring economic logic at the protocol layer, enabling a creator-first innovation economy to emerge.

4.2. Decentralization as an Economic and Moral Imperative

Today's AI ecosystem increasingly resembles early industrial monopolies: power is concentrated, access is gated, and platforms extract value while offering little in return. A handful of tech giants determine who builds, who profits, and who gets seen. This centralization not only limits innovation but replicates the inequities of the Web2 era—where users and builders alike have minimal agency. Web3 reimagines this structure. Just as 20th-century companies evolved from robber-baron monopolies into employee-owned firms with equity incentives, Web3 ushers in a new era of distributed ownership and governance. In this model, participants are no longer passive users—they are stakeholders. I

Cubed embodies this vision through model tokenization, pay-per-use mechanics, and DAO-based governance, where value and voice are shared, accountable, and transparently coordinated.

4.3. Rebuilding the Order of the AI Industry

The AI industry's existing power dynamic is broken and unsustainable. It slows innovation, restricts access, and rewards only the few. Conversations around AI safety and alignment are important—but so are the economic rights of small creators, the decentralization of training data, and infrastructure access for independent contributors. Today's model mirrors a digital feudal system, where big tech and VC-backed startups act as lords, and everyone else builds under their terms. As Vitalik Buterin noted, AI must evolve into a space where every actor—developers, researchers, users—can be self-sufficient, democratic, and organically networked into a regenerative, open community. This shift isn't just moral—it's structural. For AI to scale as a sustainable industry, we must rebuild its economic order—redefining how value is created, distributed, and governed. The Revival of Distributed Capital and Intelligence Decentralization is not merely a tech shift—it's the return of capital democracy. Just as public equity markets allowed millions to co-own and benefit from corporations, I Cubed enables communities to co-own intelligence itself. A model is no longer a static asset locked in a corporate vault—it. It becomes a living, evolving protocol, backed by belief, utility, and market validation. Ownership fragmentation isn't a weakness—it's a strength. When thousands of people stake in a model's success, the system gains momentum, accountability, and resilience. Open-source models become public goods powered by private incentives. I Cubed revives the best parts of Web2—collaboration, iteration, scale—while rejecting its worst: gatekeeping, exploitation, and central control. True AI democratization doesn't begin with better models. It begins with fairer systems to own, govern, and grow them—together.

5. Discussion: Fostering AI Co-Creation in the I Cubed Modelverse

The current paradigm of AI development is shifting from a centralized, developer-driven model to a more democratized ecosystem where users are not merely consumers but active creators. I Cubed Modelverse allows contributors not only to public-list models for visibility and royalties, but also to generate N-creations—successive derivatives that extend, adapt, or fuse existing capabilities. Our platform provides a flexible "Canvas" for workflow construction but is fundamentally agnostic to the creation methodology. There are three principal methods, each grounded in current ML literature.

Prompt-Level Conditioning: At the simplest tier, a user may "tune" a foundation model by injecting specialised instructions, exemplars, or parameter flags—much as one creates a bespoke GPT in the OpenAI Playground [26] by adjusting temperature, system prompts, or few-shot exemplars. Prompt tuning has emerged as a fast, data-free technique to steer large models toward domain-specific behaviours without modifying core weights [27].

Vertical Fine-Tuning with Domain Data: For higher fidelity, creators upload niche datasets and invoke parameter-efficient fine-tuning methods such as LoRA or adapter layers, which introduce only a few million trainable parameters while keeping the backbone frozen [28]. Our platform extends this principle to entire AI workflows. Models within a workflow are interconnected via a Model Communication Protocol (MCP) [20], allowing them to function as a cohesive unit. Users can introduce new, high-value data as post-training material for the entire workflow, refining not just a single model but the collaborative intelligence of the Model Merging and Skill Fusioninterconnected system. This creates a powerful feedback loop where the workflow continuously improves its specialized capabilities.

Model Merging and Skill Fusion: Advanced users can merge multiple fine-tuned checkpoints—adding, subtracting, or averaging weight deltas—to create composite models that inherit disparate skills without retraining from scratch [29]. Recent studies demonstrate that weight-space merging, task-vector arithmetic, and low-rank averaging can efficiently combine competencies while mitigating catastrophic interference [30]. The technique offers a data-independent route to fuse "skills" learned in separate domains, enabling rapid assembly of versatile agents.

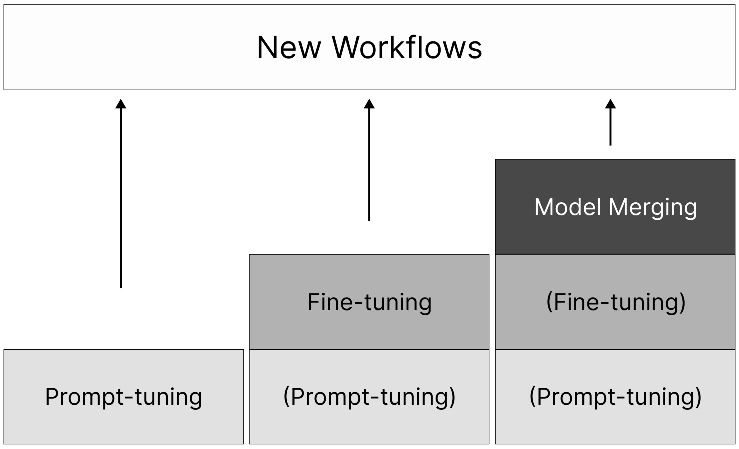

Figure 9. Three Approaches to N-Creation

Each path—from prompt conditioning to vertical adaptation to weight-space fusion—writes its lineage to the decentralized identifier (DID) graph. This guarantees that upstream contributors are automatically cited and rewarded as the derivative chain lengthens, while giving builders flexibility to choose the level of effort and customisation appropriate to their use-case.

6. Community Building

A thriving decentralized ecosystem cannot exist without an active, empowered community. I Cubed is committed to cultivating a developer- and creator-first culture by launching a dedicated DAO pool to fund community-driven initiatives. This pool will support activities ranging from model benchmarking and educational content to open-source tooling and community-led curation, ensuring that contributors at all levels are recognized and rewarded. Beyond token incentives, the DAO will govern long-term infrastructure and economic policies, allowing stakeholders to co-shape the platform's evolution. Community members will have voting rights not only over resource allocation but also over how visibility, validation, and value flow within the modelverse. Looking ahead, I Cubed will also invest in auxiliary tools to support model training, such as decentralized dataset libraries, model evaluation sandboxes, and reproducibility frameworks. These efforts aim to lower the barrier to entry for AI creators and foster an environment where innovation emerges from the bottom up, driven by shared purpose rather than corporate control. In I Cubed, community is not an afterthought—it is the protocol.

7. Future Ecosystem

To enable a truly decentralized and sustainable AI model economy, I Cubed is building an end-to-end infrastructure that supports not only model ownership and usage, but also scalable deployment, cost-efficient compute, and privacy-preserving execution. Our future ecosystem consists of three key components:

7.1. Hardware Membership

In the near future, I Cubed will introduce proprietary hardware equipped with Trusted Execution Environments (TEE). These edge devices serve as gateways to a membership-based system, allowing users to run AI models they own or license directly on secure hardware. By executing models locally, these devices reduce dependency on centralized hosting, lower inference costs, and enhance privacy by keeping user data within the device. If misuse or tampering is detected, devices can be remotely suspended, enforcing access control without compromising decentralization. This model strengthens data sovereignty while offering creators a controlled and protected runtime environment—paving the way for distributed AI compute infrastructure at the edge.

7.2. *Decentralized Compute Power Integration*

Upstream, I Cubed will collaborate with decentralized GPU platforms to create a shared computing layer that dynamically allocates resources based on real-time model demand. Through partnerships with networks such as Akash or io.net, the platform will enable cost-efficient compute for model training, fine-tuning, and on-demand demos without relying on traditional cloud monopolies. This flexible compute structure reduces overhead for creators and ensures the modelverse remains operationally scalable and permissionless.

7.3. *Inference Endpoints and Public Demo Spaces*

At the application layer, I Cubed will provide seamless tools for deploying AI models into production. Developers can convert any model in the modelverse into live inference endpoints, host public demos, or integrate models into real-world workflows. These endpoints will connect with cloud providers like Azure and Google Cloud, as well as I Cubed's own hardware infrastructure. Usage data will be transparently tracked, and revenues fairly distributed to model stakeholders, ensuring continuous incentives for both developers and infrastructure providers. In doing so, I Cubed closes the loop from open innovation to real-world impact—ensuring models not only live in code but also operate, earn, and evolve in decentralized environments.

8. Contributors

The contributors are listed in alphabetical order by last name. The authors can be contacted via the email address.

- Fernando Jia
- Rebekah Jia
- Florence Li
- Tianqin Li
- Jade Zheng

Corresponding author is Fernando Jia and Tianqin Li.
Email: fernando.jia@intelligencecubed.com;tianqinl@cs.cmu.edu.

References

1. First Page Sage. Top generative ai chatbots by market share – may 2025, May 2025. Accessed: 2025-06-01.
2. Shubham Singh. Chatgpt statistics 2025 – dau & mau data (worldwide), May 2025. Accessed: 2025-06-01.
3. Will Knight. Openai's ceo says the age of giant ai models is already over. *WIRED*, April 2023. Accessed: 2025-06-01.
4. Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021.
5. AI Hive. Build llm from scratch, 2023.
6. DeepSeek-AI. Deepseek-v3, 2025. Accessed: 2025-06-01.
7. Sungmin Woo. The deepseek shock: A 'cost-effective' language model challenging gpt, 2 2025.
8. Andrii Rusanov. Bypassing sanctions: Deepseek flashmla improves the performance of nvidia h800 ai chips by 8 times, 2 2025.
9. Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 1 2025.
10. Seek AI. Understanding deepseek: What enterprises need to know, 2 2025.
11. Allied Insight. Allied insight, 2025.
12. Planable. 77 ai statistics & trends to quote in 2025 + own survey results, 3 2025.
13. Tim Tully, Joff Redfern, and Derek Xiao. 2024: The state of generative ai in the enterprise, 11 2024.
14. Kar Balan, Andrew Gilbert, and John Collomosse. Content arcs: Decentralized content rights in the age of generative ai. *arXiv preprint arXiv:2503.14519*, 2025.

15. Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.
16. Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.
17. Code Labs Academy. What is the glue benchmark?, 2024.
18. Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32, 2019.
19. Minsu Kim, Evan L Ray, and Nicholas G Reich. Beyond forecast leaderboards: Measuring individual model importance based on contribution to ensemble accuracy. *arXiv preprint arXiv:2412.08916*, 2024.
20. Anthropic. Model context protocol, 2024.
21. Shreya Shankar and Aditya Parameswaran. Towards observability for production machine learning pipelines. *arXiv preprint arXiv:2108.13557*, 2021.
22. AIOZ Network. Aioz network launches aioz ai: A marketplace for web3 ai models and compute, 2025.
23. Ritual Foundation. Model marketplace, 2025.
24. Reelmind.ai. Prime video license ai: Understanding the challenges of digital rights management, 2025.
25. Robin Hanson. Futarchy: Vote values, but bet beliefs, 2000.
26. OpenAI Platform. Text generation and prompting, 2024.
27. Kaiyan Chang, Songcheng Xu, Chenglong Wang, Yingfeng Luo, Xiaoqian Liu, Tong Xiao, and Jingbo Zhu. Efficient prompting methods for large language models: A survey. *arXiv preprint arXiv:2404.01077*, 2024.
28. Siwei Li, Yifan Yang, Yifei Shen, Fangyun Wei, Zongqing Lu, Lili Qiu, and Yuqing Yang. Expressive and generalizable low-rank adaptation for large models via slow cascaded learning. *arXiv preprint arXiv:2407.01491*, 2024.
29. Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022.
30. Jiho Choi, Donggyun Kim, Chanhyuk Lee, and Seunghoon Hong. Revisiting weight averaging for model merging. *arXiv preprint arXiv:2412.12153*, 2024.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.