SCAI: The Web3 Brain of Science

SCAI Foundation

Abstract. SCAI is building a next-generation data infrastructure to address the critical scarcity of high-quality, structured data for the artificial intelligence (AI) industry. We solve this by tackling the inefficiencies at the source: academic publishing. The current academic ecosystem traps valuable knowledge behind paywalls and fails to reward researchers adequately, creating a bottleneck for scientific and technological progress.
This whitepaper introduces SCAI's vision to create a new knowledge economy through **DataFi (Data Finance)**, transforming scientific data into a transparent, monetizable asset class. Our Web3-native platform provides an integrated infrastructure for researchers to collaborate, publish, and manage high-value data. Our solution integrates three core pillars: a **Decentralized Data Hub** leverages decentralized storage like Irys to ensure permanent, transparent, and censorship-resistant access to knowledge; an **AI-Driven Curation Platform** uses a combination of AI and scholarly networks to structure this knowledge for machine consumption; and a **Collaborative Research Infrastructure** streamlines the research workflow while rewarding contributors with the SCAI token.
By creating this direct incentive layer, SCAI drastically reduces publication costs, targeting 1/10th to 1/100th of the current industry average, and introduces new revenue streams for data contributors. While facing the challenge of establishing academic reputation outside traditional indexing, our progress is demonstrated by our platform (https://scai.sh) and product suite, including **SCAI-Box** and **SCAI-Search**. We are positioned to capture the growing open-access market and serve the voracious data needs of the AI economy. SCAI is not just an academic platform; it is the foundational infrastructure for the future of data science.

1 Introduction

The artificial intelligence revolution, powered by advancements in Large Language Models (LLMs), is fundamentally reshaping our world. However, this progress faces an existential threat: the depletion of high-quality data. The foundational principle of "scaling laws", which posits that bigger models fed with more data yield better performance, is beginning to fail as the reserves of usable public internet data are exhausted. The future of AI no longer lies in the sheer volume of data, but in its quality, structure, and verifiability. The industry now confronts a critical bottleneck: a severe scarcity of the specialized, ethically sourced data required to train the next generation of intelligent systems.

Simultaneously, the primary engine for producing this high-quality knowledge, the academic ecosystem, is crippled by a deeply flawed economic model.

The $32.1 billion academic publishing industry operates on an extractive framework where researchers, the producers of value, pay exorbitant fees (often exceeding $3,000 per paper) to publish their work, surrender their copyrights, and perform peer review for free. This system creates misaligned incentives, restricts knowledge access behind expensive paywalls, and delivers low-quality, fragmented digital services. It has become a barrier to innovation, failing to efficiently curate and distribute the world's most valuable data asset: scientific knowledge.

Current approaches have failed to address these challenges holistically. **Traditional publishers** like Elsevier and Springer perpetuate the problem with their high-cost, closed-access models. While they are the incumbents, their business model is the direct cause of the economic inefficiency and data friction we see today. **Preprint platforms** such as arXiv have championed open access, providing a valuable service for rapid dissemination. However, they lack formal peer-review mechanisms for quality control and offer no economic model to incentivize or reward the creation of high-quality, structured data. Lastly, **AI-driven academic tools** like Google Scholar, Semantic Scholar, and even general-purpose LLMs have improved literature discovery and analysis. Yet, they are merely sophisticated consumers of a broken data ecosystem. They operate on top of the existing flawed foundation and do not address the root problems of data quality, ownership, or the misaligned incentives in knowledge production.

SCAI (Scholarly Collaborative Academic Intelligence) is engineered to solve these two intertwined crises at their source. We are not merely building a better tool for researchers; we are building the foundational data infrastructure for the AI era by creating a new, equitable, and efficient economic model for academic research. Our platform provides a vertically integrated ecosystem that realigns the entire value chain, from knowledge creation and validation to its consumption by both humans and AI. SCAI's architecture is designed to transform the academic lifecycle into a transparent, efficient, and rewarding process for all participants.

Our primary contribution is the introduction of **DataFi (Data Finance)**, a paradigm that treats high-quality, structured scientific data as a transparent, liquid, and monetizable digital asset. This is our fundamental innovation, which directly realigns incentives to reward data producers and curators. Our key contributions are:

- **Economic Model:** By leveraging the SCAI token, we create a circular economy where the consumers of high-quality data (e.g., AI companies, research institutions) can directly fund and reward the producers (researchers, reviewers, and curators). This drastically reduces publication costs and creates new revenue streams for academics.
- **Integrated AI and Web3 Infrastructure:** SCAI's uniqueness lies in the deep synergy between AI and Web3. We use AI to automate the structuring, validation, and analysis of data, transforming unstructured knowledge into machine-readable assets. We use Web3 technologies, specifically decentralized storage via Irys and smart contracts, to guarantee data permanence,

provenance, and to execute the incentive mechanisms of our DataFi economy in a transparent and trustless manner.
– **A Unified, High-Quality Data Hub:** Unlike fragmented platforms, SCAI serves as a single source of truth for high-quality, structured scientific data. By building a robust curation and peer-review layer directly into the infrastructure, we ensure the data is not only accessible but also reliable and ready for advanced AI applications.

The rest of this whitepaper outlines the architecture, tokenomics, and strategic roadmap for SCAI. We will demonstrate how our platform, through its core innovation of DataFi, solves the critical data bottleneck for the AI industry while simultaneously fixing the broken economic model of academic publishing.

Our integrated platform actualizes this vision through three pillars: a **Decentralized Data Hub** on Irys ensures permanent and open access to research; an **AI-Driven Curation Platform** structures and validates this knowledge for machine consumption; and a **Collaborative Research Infrastructure** rewards authors and reviewers with SCAI tokens for their contributions. This whitepaper details SCAI's architecture, its tokenomics, and our roadmap to build the foundational layer for the next generation of science and AI. Our progress is already demonstrated by our platform (https://scai.sh) and its core components, **SCAI-Box** and **SCAI-Search**.

2 Framework

SCAI redefines the academic ecosystem by integrating AI and Web3 technologies into a unified platform that supports the entire research lifecycle. To address the dual crises of AI data scarcity and broken academic economics, SCAI's framework is designed as a vertically integrated ecosystem for the production, curation, and distribution of high-quality scientific data. Rather than discrete technical layers, our architecture is built on three synergistic pillars that together power our **DataFi (Data Finance)** economy. These pillars are: the **Decentralized Data Hub**, the **AI-Driven Curation Platform**, and the **Collaborative Research & Publishing Infrastructure**.

2.1 Pillar 1: The Decentralized Data Hub (DDH)

The DDH is the foundational layer of trust for all data within the SCAI ecosystem. Its purpose is to serve as a permanent and verifiable public ledger for scientific knowledge, solving the pervasive problem of data impermanence. In traditional academia, references often lead to broken links (link rot), and papers can be retracted or altered without a clear public record. The DDH fundamentally solves this by transforming research outputs from ephemeral files into persistent, immutable data assets. To achieve this, we leverage decentralized storage protocols like **Irys**, which assigns a unique, permanent Content Identifier (CID) to every piece of data. Unlike a URL, which points to a location, a CID points to

the content itself, guaranteeing that the asset is forever findable and its integrity is preserved.

This guarantee of permanence is coupled with cryptographically verifiable provenance. Every asset submitted to the DDH, from a full manuscript to a single dataset, is signed by its creator's Decentralized Identifier (DID). This creates an unbroken, auditable chain of evidence for every piece of knowledge, making it profoundly resistant to fraud or ownership disputes. Researchers interact with this system primarily through **SCAI-Box**, which functions as their personal, permanent digital library. Here, they can manage, share, and track their cryptographically secured assets with confidence. The true power of the DDH, however, lies in its nature as an open and composable infrastructure. For example, a third-party fact-checking organization could build a tool that directly queries our DDH's GraphQL API to verify claims in real-time against the immutable source documents. The DDH is therefore the unfalsifiable bedrock upon which all other value in the SCAI ecosystem is built.

2.2 Pillar 2: The AI-Driven Curation Platform

If the DDH is the vault that guarantees the integrity of raw data, the Curation Platform is the refinery that transforms this raw material into high-value, structured knowledge. This pillar's primary function is to bridge the vast gap between human-readable documents—typically static, isolated PDFs—and a dynamic, interconnected, machine-readable knowledge base. It is this value-creation engine that unlocks the full economic potential of scientific data for the AI industry.

The curation process begins the moment a paper is committed to the DDH. Our pipeline deconstructs the document, converting its core components—text, tables, figures, and equations—into rich semantic embeddings. This vectorization process does more than just index keywords; it maps complex scientific concepts into a multi-dimensional "idea space," where the system can mathematically reason about their relationships, nuances, and context. Simultaneously, we construct a comprehensive knowledge graph that explicitly maps not only citations, but also the methodologies employed, the datasets referenced, the core claims asserted, and even contradictory findings across different papers. A static document is thus reborn as a dynamic, queryable data object, deeply interconnected with the entire corpus of scientific knowledge.

The tangible output of this curation engine is experienced through applications like **SCAI-Search**, the primary interface for interacting with our living knowledge base. Instead of wrestling with keywords, a user can pose complex questions in natural language, such as, "What are the main competing theories for protein folding published since 2023?" The platform can then deliver a synthesized summary, complete with direct links to the source claims, which are immutably stored and verifiable in the DDH. This capability, alongside features for identifying knowledge gaps and comparing methodologies, is demonstrated on our platform prototype at https://scai.sh. It transforms research from a manual search-and-read process into a dynamic dialogue with the entire body of scientific literature.

2.3 Pillar 3: The Collaborative Research & Publishing Infrastructure

If the DDH provides the foundation of trust and the Curation Platform creates value, this third pillar serves as the dynamic engine for growth. It is here that new, high-quality knowledge is produced, validated, and exchanged, governed by the transparent economic principles of DataFi. This infrastructure replaces the extractive, gatekept model of traditional publishing with a community-driven, economically aligned system that powers a self-sustaining flywheel of scientific production.

The SCAI token is the medium of exchange that fuels this new data economy. The value flow is direct and transparent, best illustrated by an example: an AI firm licenses a curated dataset on quantum computing via our API, paying in SCAI tokens. A significant portion of this payment is then automatically routed via smart contract to the digital wallets of the researchers and peer reviewers whose work constitutes that dataset. This mechanism directly solves the chronic problem of uncompensated academic labor, creating a powerful incentive for researchers to contribute their best work to the ecosystem.

Quality and efficiency are ensured through a hybrid peer-review model that augments human expertise, rather than replacing it. An incoming manuscript is first analyzed by our AI, which not only screens for plagiarism but also verifies that data citations link to accessible datasets in the DDH, analyzes the statistical methods against established best practices, and even suggests potential human experts for validation based on content similarity. This frees up human reviewers to focus on the high-level scientific merit of the work. Each successful review, publication, or curation task is a verifiable transaction that contributes to a researcher's on-chain reputation, linked to their Decentralized Identifier (DID). Unlike a traditional academic title, this reputation is portable and universally recognized, empowering researchers with credentials that can be used to access grants, join collaborations, or participate in the governance of the SCAI ecosystem.

In summary, the three pillars of the SCAI framework operate in concert to create a self-sustaining value chain designed for the new economy of scientific data. Each pillar performs a distinct but interconnected role:

- **The Decentralized Data Hub** serves as the foundational layer of *trust*. It guarantees that all scientific knowledge is stored as permanent, immutable, and verifiable assets, forming the bedrock of the entire ecosystem.
- **The AI-Driven Curation Platform** acts as the *value-creation* engine. It transforms raw, unstructured data from the Hub into enriched, interconnected, and AI-ready knowledge, unlocking its true economic and scientific potential.
- **The Collaborative Infrastructure** provides the economic and social *incentives for growth*. By implementing the DataFi model, it fuels the production and validation of new, high-quality data, creating a positive-feedback loop that continuously enriches the entire ecosystem.

3 Tokenomics

The SCAI token is the backbone of the SCAI ecosystem, enabling a decentralized, incentivized, and sustainable platform for academic research. By leveraging Web3 principles, the SCAI token facilitates low-cost transactions, rewards contributors, and ensures equitable access to services such as literature retrieval, AI-driven analysis, and publication. This chapter outlines the token's economic model, including its allocation, utility, fee structure, deflationary mechanisms, vesting schedule, and governance, ensuring transparency and alignment with the platform's mission to revolutionize academic research.

3.1 Token Overview

The SCAI token is designed to power the platform's operations, incentivize participation, and foster long-term value creation through a deflationary model. Key details include:

- **Symbol**: SCAI
- **Total Supply**: 1,000,000,000 tokens
- **Initial Circulating Supply**: Determined by the vesting and unlock schedule (detailed below)

The token operates on a blockchain compatible with SCAI's decentralized infrastructure (e.g., integrated with Irys for storage), ensuring secure and transparent transactions.

3.2 Token Allocation

The SCAI token's allocation is structured to balance community incentives, ecosystem development, and operational needs. The following table summarizes the distribution:

Table 1. SCAI Token Allocation

Category	Percentage	Amount	Remarks
Initial Airdrop	17%	170,000,000	Community Airdrop
VC	20%	200,000,000	Strategic investments
Raydium Pool	10%	100,000,000	Fair Launch Loss
Staking Rewards	9%	90,000,000	Eventually Release
Community Foundation	36%	360,000,000	Ecosystem incubation and incentives
- Initial Release	6%	60,000,000	Immediate allocation
- Locked	30%	300,000,000	1% monthly unlock
Marketing	8%	80,000,000	Airdrops, Events, Market making

3.3 Token Utility

The SCAI token is deeply integrated into every function of the platform, ensuring its utility is directly tied to the ecosystem's activity and growth across our three pillars:

- **Platform Access & Payments:** The token is the primary method for payments. This includes fees for premium services on our Curation Platform (e.g., advanced AI analysis), publication and peer-review services on our Publishing Infrastructure, and high-volume storage access on the Data Hub.
- **Incentivizing Production:** In line with our DataFi model, tokens are used to directly reward the producers of value. This includes compensating authors for high-impact publications, reviewers for timely and rigorous feedback, and data curators for their contributions.
- **Staking and Governance:** Token holders will be able to stake their SCAI to receive a share of platform revenue and participate in the governance of the ecosystem. This allows the community to vote on key decisions, such as feature development, fee structures, and the allocation of foundation funds.
- **Community Engagement:** Tokens are used to incentivize community-building activities, such as participating in scholarly discussions, referring new users, and contributing to open-source development.

By integrating these use cases, the SCAI token creates a vibrant ecosystem where researchers, reviewers, and other stakeholders are rewarded for their contributions, aligning incentives with platform growth.

3.4 Vesting Schedule

To align incentives and prevent market flooding, SCAI implements a vesting schedule for key allocations:

- **Community Foundation (36%, 360,000,000 tokens)**:
  - **Initial Release**: 6% (60,000,000 tokens) available immediately for ecosystem development.
  - **Locked Portion**: 30% (300,000,000 tokens) locked, with 1% (3,000,000 tokens) unlocked monthly to fund ongoing initiatives.
- **Staking Rewards (9%, 100,000,000 tokens)**: Release 0.3% pre-month, designed to incentivize long-term participation and platform stability.

The vesting schedule ensures controlled token release, balancing immediate ecosystem needs with long-term sustainability.

3.5 Multi-Sig Governance

To ensure transparency and security, all major token operations, including large transfers and burns, are managed through multi-signature (multi-sig) wallets:

- **Multi-Sig Address**: To be announced, ensuring secure and decentralized control.
- **Operational Integrity**: Multi-sig governance requires multiple stakeholder approvals, preventing unilateral actions and enhancing trust.

This governance model aligns with Web3 principles, ensuring that the SCAI ecosystem remains transparent and community-driven.

By combining low-cost services, tokenized incentives, and a deflationary model, SCAI's tokenomics fosters a robust and equitable academic ecosystem, poised to disrupt traditional publishing and empower researchers globally.

4 Comparison

SCAI redefines academic research by integrating artificial intelligence (AI) and Web3 technologies into a unified platform that addresses the inefficiencies of traditional academic publishing and fragmented tools. By offering an end-to-end solution for literature retrieval, analysis, and publication, SCAI outperforms existing systems in cost, efficiency, accessibility, and equity. This chapter compares SCAI with traditional publishers, preprint platforms, and AI-driven academic tools, using data-driven insights to highlight its competitive advantages. We also address challenges, such as establishing academic reputation, to provide a balanced perspective on SCAI's potential to transform the academic ecosystem.

4.1 Preprint Platforms

Preprint platforms like arXiv, bioRxiv, and Crypto ePrint have grown in popularity, hosting over 2 million papers annually and offering free, open-access repositories for early-stage research. SCAI builds on their strengths while addressing their limitations, providing a more comprehensive and incentivized ecosystem.

- **Integrated Functionality**: Preprint platforms focus on hosting manuscripts, with limited tools for search, analysis, or peer review. SCAI integrates these functions into a single platform, enabling seamless transitions from literature retrieval to AI-driven analysis and publication. For example, a researcher can use SCAI's prototype (https://scai.sh, user ID: b94c251b-9ed1-46c9-92e6-350e07260ae9) to search a vectorized database of 20 million articles, store papers in `scibox.store`, and engage in multi-turn AI dialogues for insights, all within one ecosystem.
- **Decentralized Infrastructure**: Unlike arXiv's centralized servers, which rely on institutional funding (e.g., Cornell University's $1 million annual budget for arXiv), SCAI leverages Irys's decentralized storage for permanence, censorship resistance, and zero-cost access. This ensures long-term availability without dependence on single-entity funding.
- **Incentivized Ecosystem**: Preprint platforms lack formal peer review or contributor incentives. SCAI's tokenized system rewards reviewers with SCAI

tokens, ensuring high-quality feedback within 1–2 weeks. Authors benefit from low-cost publication and NFT-DOI issuance for unique identification, enhancing the value of preprints.
– **Reputation Potential**: While arXiv enjoys academic credibility without SCI/EI indexing, it took years to build trust. SCAI aims to replicate this by prioritizing quality, affordability, and adoption. Its 20 million article database and AI-driven tools provide immediate value, attracting researchers and building reputation over time, as seen with arXiv's 30% annual growth in submissions.

SCAI extends the open-access ethos of preprint platforms while offering integrated tools, decentralized infrastructure, and incentives, making it a more versatile and scalable solution.

4.2 Centralized Data Providers

While SCAI is set to redefine academic research, its most significant impact lies in creating a new paradigm for the AI data industry. To understand this, we must contrast our model with today's industry titans, such as **Scale AI**. These companies have become essential by providing "Data-as-a-Service" for the current generation of AI, successfully meeting the demand for human-annotated data. However, SCAI addresses a more advanced and pressing need: the shift from labeled data to structured, verifiable knowledge. We are not a competitor to these services; we are the foundational infrastructure for the next generation of AI.

The fundamental distinctions begin with the data itself. Centralized providers typically operate on proprietary or generic raw data from their clients—such as sensor logs or internet images—and their primary value-add is applying a human workforce to *annotate* or *label* it. In stark contrast, SCAI's source material is the global corpus of scientific research, a repository of knowledge that has already undergone rigorous expert validation. Our primary value-add is not manual labeling, but using our AI Curation Platform (Pillar 2) to *deconstruct, structure, and interconnect* this highly specialized information. In essence, while they provide labeled data, SCAI provides structured knowledge. We are moving from "Data-as-a-Service" to "Knowledge-as-a-Service."

This leads to a profound difference in the economic model. Centralized providers operate on a traditional B2B service model where value is captured by the central entity and distributed as wages to a contract-based workforce. **SCAI**, conversely, is a decentralized, community-driven ecosystem governed by the principles of **DataFi**. Our model is circular and regenerative. The "workforce" is the global community of researchers who are the original producers of the data. Through the SCAI token, the value generated from the consumption of curated data is distributed directly back to these contributors. We are not a service provider managing a workforce; we are the infrastructure for a self-sustaining economy of knowledge creators.

Finally, this distinction redefines the nature of trust. Trust in a dataset from a centralized provider is based on their brand reputation and contractual agreements (SLAs). Trust in a dataset from SCAI is established through cryptographic truth. Every piece of data is linked back to its source publication on our Decentralized Data Hub (Pillar 1), with its provenance and authorship immutably recorded and auditable by anyone. This provides a level of verifiability critical for building robust and explainable AI. In summary, centralized providers sell a trusted service; SCAI provides the infrastructure for verifiable truth.

5 Conclusion

SCAI represents a groundbreaking approach to addressing the longstanding challenges of academic and high-quality data. SCAI built on three synergistic layers—Application, Storage, and Incentive—leverages AI-driven tools, decentralized storage via Irys, and the SCAI token to create a cost-effective, efficient, and equitable ecosystem.

Our solution is a vertically integrated ecosystem, built upon three synergistic pillars that power our core innovation, **DataFi (Data Finance)**. The **Decentralized Data Hub** establishes a permanent and unfalsifiable bedrock of scientific truth. Our **AI-Driven Curation Platform** transforms this raw knowledge into structured, interconnected, and machine-readable assets of immense value. Finally, our **Collaborative Infrastructure** implements the DataFi model, creating a self-sustaining circular economy where the consumers of data directly fund and reward its creators. This architecture systematically dismantles the barriers of the old paradigm, replacing it with a system that is transparent, equitable, and designed for growth.

The impact of this paradigm shift is profound. By fixing the broken incentives within academic publishing, we unlock a sustainable, high-volume pipeline of the world's most valuable data. This creates a powerful flywheel: researchers are empowered and fairly compensated, fostering the creation of more high-quality knowledge. In turn, this provides the AI industry with the trusted, structured fuel it desperately needs to build the next generation of robust, explainable, and truly intelligent systems. Our working platform at https://scai.sh serves as the first proof-point of this vision.

SCAI represents more than a platform; it is a foundational knowledge layer for an intelligent future. Our mission is to accelerate the pace of both scientific discovery and technological progress by creating a world where knowledge is not just accessible, but is also a regenerative and empowering asset for its creators. We are building the infrastructure to power the next great leap forward, and we invite you to join us in redefining how humanity creates, shares, and builds upon its collective intelligence.