Claims

A Bittensor Subnet for Machine-Readable Scientific Evidence

Selective Verification, Continuous Learning, and Identity-Robust Incentives

---

Bittensor Subnet 111

Claims Team

Philipp Koellinger

Professor of Economics, Vrije University

Christian Roessler

Associate Professor of Economics, California State University-East Bay

Ogban Ugot

Engineer

White Paper v1 • Public release candidate • August 2026

Purpose. Claims turns scientific publications into persistent, machine-readable claim–evidence records. Independent miners improve a common extraction baseline, validators reserve costly review for the cases where it is most useful, and verified corrections strengthen the lower-cost system used at scale.

Abstract

Scientific findings remain trapped in documents, while AI systems and evidence-intensive organizations need structured records that preserve claims, evidence, provenance, and uncertainty. Producing claims is only the first problem: scientific data must remain trustworthy when extraction is costly, complete ground truth is scarce, participants can duplicate identities or methods, and evaluators themselves require evaluation. Claims is a production market on Bittensor, a decentralized network for rewarding machine intelligence. Miners compete to improve an economical reference extraction, while validators selectively adjudicate disagreements and use sparse known-answer or trusted review to calibrate the process. Verified corrections train progressively cheaper evaluators. Rewards follow distinct information families and accepted marginal contributions, limiting gains from duplicate identities. The architecture separates a live calibration system from the intended production release.

Keywords: scientific data extraction; claim–evidence graphs; Bittensor; decentralized AI; mechanism design; selective verification; Sybil resistance; synthetic challenges.

Contents

1 Product, Scope, and Data Object

This section defines the product before describing the market that produces it. It explains why reusable scientific evidence is missing, what a CLAIMS record contains, and how the accepted record remains distinct from both source content and world truth.

1.1 The Missing Evidence Layer

Scientific publishing is built for human reading. Evidence reviews, drug-development programs, scientific AI systems, and research analytics increasingly need something different: records that can be compared across papers, traced to their sources, corrected over time, and reused without reconstructing the same facts for every new question. The gap between documents and reusable evidence defines the problem that CLAIMS addresses.

1.1.1 Scientific knowledge is still document-bound

A paper can express methods, qualifications, uncertainty, and argument with great flexibility. That same flexibility makes the document a costly interface for systems that work across many studies. Finding a relevant paper is only the first step. A researcher or AI agent must still recover what the paper claims, which observation supports each claim, which population and comparator apply, what quantity was measured, how uncertainty was reported, and where the supporting material appears.

Literature-review teams, pharmaceutical and biotechnology firms, evidence dashboards, scientific AI systems, and individual researchers repeat this reconstruction. The work remains expensive because each user often creates a temporary answer rather than a durable record. The reusable product is a persistent evidence layer that can support many questions, users, and applications.

CLAIMS is designed to produce that layer. Its output is a canonical, versioned collection of machine-readable claims linked to evidence, source locations, structured quantities, provenance, and verification status. Search and summarization remain valuable interfaces, but they operate on top of the evidence layer instead of standing in for it.

Within the subnet, *miners* are the competing producers of structured records, and *validators* construct tasks, compare submissions, and determine evidence and reward. Bronze denotes the economical reference extraction, Silver the stronger machine-adjudicated record, and Gold sparse known-answer or otherwise trusted verification. Sections 3 and 4 develop each tier and participant role in detail. Figure 1 summarizes the transformation from publications to a persistent evidence graph and downstream uses.

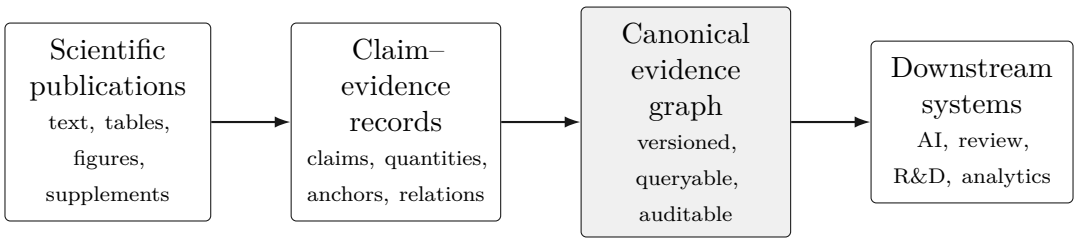

Figure 1. Scientific documents become reusable only after claims, evidence, quantities, and source anchors are organized in a persistent and auditable layer.

The evidence graph therefore serves as infrastructure rather than as a one-time response. Once a publication has been translated into traceable records, downstream systems can reuse the same structured material while preserving a route back to the source.

1.1.2 What Claims produces

The basic unit is a claim–evidence record. A record is more than a sentence extracted from a paper; it preserves the context needed to interpret and audit that sentence. At minimum, it contains:

- an atomic or otherwise well-scoped statement;
- the paper, section, passage, table, figure, or supplement from which it was extracted;
- the evidence or analysis associated with the statement;
- structured quantities, units, uncertainty, sample information, and comparators when available;
- claim type, modality, direction, and other qualifications needed to preserve meaning;
- normalized entities and links among claims, evidence, methods, and sources;
- extraction, adjudication, and verification provenance;
- a current database state, together with prior versions and unresolved alternatives.

These fields make the record useful beyond its original extraction task. They allow later users to inspect a source, compare related findings, update a disputed interpretation, and distinguish a current database state from the history that produced it.

Claims first asks what a paper states. It does not treat that answer as a judgment about whether the underlying scientific proposition is true in the world. Document curation and scientific validation require different evidence. The first task is already valuable, can be audited against the publication, and supplies a necessary input to the second.

1.1.3 Human-originated evidence curated by AI

AI participates throughout extraction, comparison, adjudication, and learning. The important boundary lies between AI-mediated transformation of identifiable scientific sources and recursive model output that lacks an inspectable evidentiary anchor.

Every accepted record retains a path to the publication and to the evidence span from which the record was constructed. An AI output is thus a proposed representation of a source. Known-answer synthetic challenges and selective human or otherwise trusted verification add independent anchors when agreement among models is insufficient. Extensive automation can then coexist with a rule that model self-consistency does not establish ground truth.

The resulting resource supports both training and inference while preserving three properties:

1. **Traceability.** A downstream system can inspect the source of a record.
2. **Correctability.** Later adjudication can patch the record without erasing its history.
3. **Controlled recursion.** AI-generated corrections enter future training only through versioned and independently tested release gates.

Together, these properties turn automation into a controlled curation process. The system can learn from corrected records without severing the connection between each accepted representation and the human-originated evidence behind it.

1.1.4 The scientific-literature resource stack

No single service supplies every production input. CLAIMS can combine complementary resources while preserving the licensing and access conditions attached to each source.

Function	Representative resources	Role in the pipeline
Work discovery and prioritization	OpenAlex [6], PubMed [7]	Identify papers, domains, authors, venues, and processing queues.
Persistent identifiers and metadata	DOI and Crossref metadata [8]	Resolve source identity, bibliographic fields, and version lineage.
Lawful full-text resolution	Unpaywall [9] and repository or publisher access	Locate open-access copies and other permitted full-text sources.
Document structure	Publisher XML, PDFs, tables, figures, supplements	Supply the content from which claims, evidence, and provenance are extracted.
Corrections and status	Retractions, corrections, and version records	Prevent superseded or corrected source material from being treated as static.

The protocol standardizes the record and its provenance, not the miner's internal stack. Miners may use different retrieval sources, parsers, models, and verification tools as long as their outputs satisfy the common schema and access constraints.

1.1.5 Scope and non-goals

CLAIMS is a production system for structured evidence. It does not replace peer review, serve as a universal scientific-truth oracle, or prescribe one summarization model. Its immediate purpose is to lower the repeated cost of translating documents into reusable records while making extraction quality measurable, contestable, and capable of improvement.

That purpose creates a data-definition problem as well as a method-discovery problem. The next subsection fixes the production object and its epistemic boundary before Section 2 turns to the market and mechanism that produce it.

1.2 The Data Object and Its Epistemic Boundaries

A scientific record can fail in two different ways: it can misrepresent the paper, or the paper itself can make a false claim about the world. Those failures require different evidence and different institutions. CLAIMS therefore separates the content of a publication, the database's current representation of that content, and the underlying scientific truth before it assigns scores or verification states.

1.2.1 Paper content, accepted record, and scientific truth

For paper p, let X_p denote the canonical content that the extraction schema asks the system to recover. Let F_p denote the current accepted CLAIMS record. Let W_p denote the truth of the underlying scientific propositions in the world. These three objects are related, but each answers a distinct question:

$$\boxed{\text{paper content } X_p \quad \neq \quad \text{accepted record } F_p \quad \neq \quad \text{world truth } W_p.}$$

The first production objective is to make F_p a faithful, structured representation of X_p. A later credibility layer can draw on replication, study design, retractions, contradiction networks, and other evidence to assess W_p. Keeping the objects separate prevents a document-extraction score from being presented as a judgment of scientific validity.

This boundary also clarifies the role of verification. A record can receive strong evidence that it accurately represents a paper even when the paper's real-world claim remains contested. Conversely, later scientific evidence can overturn a proposition without showing that the original extraction was inaccurate.

1.2.2 Four layers of extractable information

The extraction task contains four layers, each with a different source of truth and a different verification burden:

1. **Coverage and source structure.** Sections, passages, tables, figures, identifiers, and source anchors.
2. **Grounded facts.** Values, units, signs, sample sizes, intervals, statistical results, interventions, comparators, and outcomes that can often be checked directly against the source.
3. **Judgment records.** Relation type, direction, modality, causal status, context, evidence role, and semantic equivalence.
4. **Prediction and consensus.** Assessments for which the paper does not supply a complete answer key, including some higher-order classifications and later judgments of evidence strength.

Cheap deterministic checks work best for source structure and many grounded facts. Semantic adjudication becomes more important for judgment records. Prediction and consensus may require independent judgment, selective human review, or a separate peer-prediction or verification design.

The layers should therefore remain visible in both scoring and reporting. A system that performs well on source anchors may still fail on causal interpretation, while agreement on an open judgment cannot be treated as equivalent to a directly grounded numerical check.

1.2.3 Canonicalization before scoring

Miners submit structured records, but literal wording does not determine whether two records contain the same information. A canonicalization operator κ maps semantically equivalent expressions to a shared content unit while preserving material differences in qualifiers, modality, polarity, numerical values, populations, comparators, and evidence links.

Canonicalization allows the system to distinguish:

- a paraphrase from a new claim;
- a refinement from a contradiction;
- independent corroboration from duplicated output;
- a split or merged claim from a substantive difference;
- a genuine improvement from superficial variation created to earn additional reward.

The same operation serves data quality and incentive design. Stable final-record item identifiers let the system attribute accepted contributions, deduplicate support, preserve versions, and calculate marginal contribution on the information that survives adjudication.

Canonicalization does not erase ambiguity. When the source supports several interpretations that cannot yet be reconciled, the database can preserve those alternatives and their provenance instead of forcing a false equivalence.

1.2.4 Verification states and reversible lineage

A participant's reward score and a record's database status answer different questions. The first estimates how useful the participant's behavior is under the current evaluation design. The second records how much support a particular item has received. A highly rated miner does not automatically certify every submission, and a weak miner can still contribute an item that later receives independent verification.

The v1 database uses a verification ladder of the following form:

State	Interpretation
Raw	Retained with provenance but excluded from trusted query results.
Provisional	Structured and eligible for use with an explicit unaudited label.
Corroborated	Independently matched by another reliable information family or reference process.
Audited	Validated in a sampled audit or high-confidence adjudication.
Gold	Assigned the highest verification tier, with provenance identifying a known synthetic answer, trusted human review, or another explicitly designated source.
Ambiguous	Preserved as an unresolved alternative rather than forced into one canonical answer.
Rejected	Invalid, unsupported, duplicated without value, or outside the task scope.

Every accepted record preserves source anchors, contributing submissions, Bronze and Silver versions, Gold provenance when present, adjudication decisions, supersession links, and alternative interpretations. A correction creates a new version rather than silently replacing the history that made the change intelligible.

This combination of tiered status and reversible lineage makes the evidence graph auditable even as it improves. It also supplies the observable objects that the mechanism must reward, test, and protect in Section 2.

2 Production Environment and Design Principles

With the production object fixed, the design problem is to sustain search over changing extraction methods while making hidden shortcuts observable. This section first explains Bittensor's role and then translates the production problem into explicit mechanism-design requirements.

2.1 Why Bittensor Is the Right Production Environment

Building the evidence layer requires repeated decisions about models, prompts, parsers, retrieval systems, verification routines, and compute. Because the best combination is unknown and keeps

changing, the production problem is poorly suited to a fixed technical prescription. Bittensor provides a market in which the output can be specified while independent participants continue to search for better ways to produce it.

2.1.1 Extraction has an unknown and changing production function

Scientific extraction consumes compute, but access to distributed compute is only part of the rationale for a subnet. Quality also depends on model choice, prompting, retrieval, document parsing, treatment of tables and figures, schema decomposition, numerical checks, domain specialization, self-verification, and ensembles. The best mix changes as model capabilities, prices, and scientific source formats change.

A conventional procurement process tends to choose a method or provider before this production function is fully understood. A Bittensor subnet can instead define the required artifact and the evaluation process, then reward participants who discover better ways to complete the task. Miners keep their internal methods private while submitting the records needed to show that those methods add value.

This approach follows the foundational Bittensor idea that intelligence systems can evaluate useful machine output and allocate rewards through a decentralized market rather than through a single fixed supervised benchmark [1]. Scientific evidence extraction adds an important complication: evaluation is itself costly, incomplete, and only partly grounded in known answers. The subnet therefore needs a scientific mechanism inside the broader market.

2.1.2 Collective ingenuity is the scarce input

The subnet purchases a continuing search over methods, not merely a volume of inference. Independent miners can explore:

- different foundation models and model families;
- domain-specific prompts, ontologies, and extraction decompositions;
- full-text, table, figure, and supplement parsers;
- numerical and citation verification routines;
- retrieval and context-selection strategies;
- specialized methods for particular paper types or claim classes;
- cost-reducing approximations that preserve accepted quality.

Only verified improvements in the shared data product create economic value. Additional computation is useful when it discovers information, reduces error, broadens coverage, or lowers the cost of an accepted record. Compute that merely reproduces an existing output contributes availability, but it does not provide the same residual information as an independent correction.

2.1.3 The outer network and the inner scientific mechanism

Bittensor supplies participant identities, staking and emissions, validator weight submission, and network-level consensus. Those functions do not determine whether a paper contains a claim, whether two phrasings are semantically equivalent, whether a number was transcribed correctly, or whether a validator used an independent evaluation process.

CLAIMS therefore operates through two nested mechanisms:

1. **The Bittensor layer** converts validator weights into economic rewards and supplies the market in which miners and validators participate.
2. **The Claims layer** defines the data object, creates tasks, compares records, constructs known-answer challenges, estimates miner and validator reliability, and determines which information enters the database.

The Claims layer performs the scientific production work. Network consensus surrounds that work with an economic system, but it cannot substitute for the domain-specific rules that turn submissions into trusted evidence.

2.1.4 Why not one centralized extraction pipeline?

A centralized pipeline can be efficient when the leading method is known and stable. Scientific extraction presents several moving frontiers at once: quality, coverage, latency, inference cost, access to full text, and domain specialization. One provider also creates a common-mode failure risk and makes it harder to learn whether an apparently strong baseline has systematic blind spots.

CLAIMS keeps a common reference pipeline because a shared baseline lowers evaluation cost and gives all participants a floor to beat. The reference pipeline does not receive a monopoly on innovation. Miners search for residual errors and better representations, while validators decide whether the differences are genuine improvements.

The resulting division of labor combines a stable production floor with decentralized method search. With the production object already fixed, the next subsection translates that division of labor into explicit mechanism-design requirements.

2.2 A Mechanism-Design Framework for Claims

Once the data object is fixed, the design question becomes behavioral. Miners, judges, validators, and operators can often take a cheaper action that resembles productive work while adding little information or imposing hidden cost. The mechanism must make those deviations observable before stronger rewards can make the desired actions worthwhile.

2.2.1 Roles, hidden actions, and low-cost deviations

Each role has a target behavior and a set of plausible shortcuts. Describing those shortcuts directly is more useful than rewarding an undefined notion of quality.

Role	Target behavior	Relevant deviations
Extraction miner	Run an independent, grounded process that covers the paper and finds genuine improvements over Bronze.	Copy Bronze; use a shallow model; omit difficult content; flood plausible extras; manipulate wording; optimize only for recognizable audits.
Silver validator	Resolve semantic and substantive disputes accurately, use source evidence, calibrate confidence, and escalate difficult cases.	Use a cheaper judge; decide every case; over-escalate to avoid responsibility; follow origin or style cues; reuse one correlated model family.
Gold judge	Apply independent effort to hidden known-answer and genuine cases, and abstain when warranted.	Guess; follow a majority; recognize synthetic templates; create identities to multiply vote weight.
Economic actor	Add independent information through one or more genuine strategies.	Split one method across identities; reset after poor performance; create near-duplicate strategies to acquire more assignments or ranks.
Protocol operator	Commit to stable rules, hidden sampling, reproducible scoring, and versioned releases.	Change models or samples after outcomes are known; train on unverified errors; apply asymmetric rules; publish irreproducible aggregates.

The mechanism responds to observable behavior rather than inferred intent. A weak model and a strategic actor receive the same statistical treatment when they create the same errors, redundancy, or validation burden.

2.2.2 Objectives and constraints

The system must pursue several objectives at once:

- accurate and complete canonical records;
- cost-efficient production and validation;
- useful diversity of miner methods;
- credible entry opportunities for new miners;
- resistance to duplicate identities and duplicated information;
- accurate, independent, and reproducible validator judgment;
- continuous improvement of the production baseline;
- persistent downstream value rather than benchmark-only activity.

These objectives operate under incomplete ground truth, noisy AI judgment, correlated model errors, heterogeneous compute costs, limited human review, changing model capabilities, and inexpensive digital identities. A reward rule that improves one objective can therefore damage another. For example, strong prize concentration may raise effort while reducing participation or amplifying score noise.

The architecture keeps these dimensions separate so that one strong aggregate score cannot conceal a failed channel. That separation begins with the signals used to distinguish productive behavior from its cheaper substitutes.

2.2.3 From general principles to Claims design choices

The companion mechanism-design analyses suggest a sequence for constructing those signals [3, 4, 5]. First identify the desired action and its deviations. Then show that observable measures separate them and cover every important hidden-action dimension. Choose reward intensity and economic exposure only after those conditions hold.

General principle	Claims implementation
Reward outputs rather than unverifiable effort	Score grounded records, accepted corrections, and reproducible judging outcomes.
Create separate channels for separate objectives	Distinguish coverage, grounding, numerical fidelity, semantic fidelity, independent discovery, and judge calibration.
Pay for residual information	Canonicalize duplicates and reward the contribution that remains beyond Bronze and other miner families.
Use costly truth selectively	Run Bronze broadly, Silver on discrepancies, and Gold on hidden challenges and difficult residuals.
Audit evaluators with hidden truth	Insert known synthetic errors and clean controls into otherwise realistic judging packets.
Separate truthfulness from effort	Use calibrated reporting scores together with hidden tasks that make low effort measurably worse.
Use separate levers for separate problems	Distinguish score design, reward intensity, prize concentration, participation access, identity rules, challenge penalties, and ingestion gates.
Remove nuisance variation	Normalize or stratify by paper type, domain, difficulty, reference release, and model-family correlation.

These principles do not prescribe one permanent set of constants. They identify which object each control must affect and which failure it is intended to prevent.

2.2.4 Separation, coverage, and exposure

The order of design can be summarized as

$$\boxed{\text{separation} \longrightarrow \text{coverage} \longrightarrow \text{exposure.}}$$

Separation asks whether the score makes the desired strategy more attractive than a relevant deviation after accounting for cost. **Coverage** asks whether the observable channels respond to every hidden-action dimension that needs an incentive. **Exposure** asks how much reward variance, audit volume, or score intensity is required once the score points in the right direction.

A larger prize cannot induce independent extraction when copying Bronze and extracting independently produce the same observable score. A refined coverage measure cannot induce numerical verification when numerical fidelity has no distinct channel. A larger synthetic suite adds little when every item measures the same capability. The challenge distribution must span the behaviors the system needs rather than merely contain many cases.

This sequence prevents economic force from being applied to an uninformative signal. It also clarifies why scoring, sampling, reward concentration, and identity controls must remain distinct levers.

2.2.5 Residual information as the unit of value

Agreement can arise from independent accuracy, an obvious passage, a shared public baseline, or direct copying. Counting every agreeing identity as new evidence would therefore overstate information. The rewardable object is the part of a submission that changes the accepted record or adds independent corroboration after conditioning on what the system already knows.

That principle leads to:

- semantic deduplication before support is counted;
- behavior-based strategy families rather than one-reward-per-identity accounting;
- marginal-contribution scoring at the level of accepted final-record items;
- a bounded bonus for independent corroboration rather than value that grows with every apparent copy;
- leave-family-out scoring wherever a participant could otherwise alter its own benchmark.

Residual information links the scientific and economic sides of the design. It defines what improves the evidence graph and, at the same time, what deserves incremental reward. Section 3 implements this principle from task release through database update.

3 Workflow and Quality Architecture

The architecture implements those design requirements through a version-stable workflow and a tiered quality system. The workflow fixes the sequence of production and verification decisions; the Bronze–Silver–Gold architecture allocates costly judgment and routes verified corrections back toward the lower-cost layers used at scale.

3.1 Production Workflow and Versioning Philosophy

The mechanism must remain understandable even as models, thresholds, and sample rates change. CLAIMS therefore separates a stable sequence of production and verification steps from the release-specific controls used to implement them. The same distinction also keeps the live v0 calibration system from being mistaken for the fuller v1 production design.

3.1.1 The invariant workflow

The public workflow specifies the objects and decisions that every release must produce. It does not depend on a particular model vendor, prompt, API route, or batch size.

1. **Task release.** A homogenized batch of papers and an extraction template are released. The template defines the required output without necessarily revealing which fields will be used for production, adjudication, or hidden calibration.
2. **Independent production.** Selected miners process the papers with private model and tool pipelines, then return structured records grounded in the sources.
3. **Cheap validation.** Schema, provenance, numerical format, source resolution, duplication, and other deterministic checks run before expensive judgment.
4. **Bronze comparison.** A versioned reference miner produces an economical baseline for the relevant papers or hidden sample.

5. **Dispute construction.** Miner and Bronze records are canonicalized. Mismatches and potentially superior miner-only content become blinded adjudication packets.
6. **Silver adjudication.** A stronger AI process receives more context and compute than Bronze. It identifies semantic equivalence, accepts improvements, rejects unsupported changes, and resolves most discrepancies.
7. **Gold escalation and calibration.** Difficult cases may go to independent judging or selective trusted review. Synthetic corruptions with known answers are mixed into the process to calibrate miners and validators.
8. **Scoring and reward.** Extraction quality, judging quality, marginal contribution, and reliability are updated. Rewards are assigned to behaviorally distinct information families rather than to raw identity count.
9. **Database write.** Accepted records enter the graph in states that reflect their verification level and preserve full lineage.
10. **Reference update.** Gold corrections improve Silver, and corrected Silver records become training and evaluation material for later Bronze releases.

Figure 2 displays the stable sequence and the reverse reference-update path.

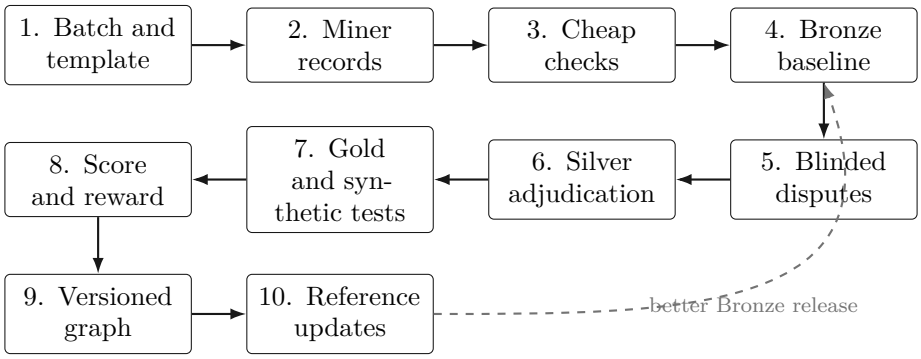

Figure 2. The version-stable workflow separates production, selective verification, reward, database entry, and reference improvement. Models, sample rates, and thresholds remain release parameters.

The workflow moves information through three related systems: a production market, a verification process, and a versioned database. Keeping the steps distinct allows a release to change one implementation choice without changing the role of the surrounding stages.

3.1.2 v0: a deliberate calibration system

v0 is a minimal live mechanism. It lowers the cost of entry, exercises the full operational path available today, and exposes failure modes before the higher-leverage production architecture begins to learn from its own outputs. In the current v0 loop:

- every scoring paper receives or retrieves a Bronze extraction;
- miners submit structured artifacts with source grounding;
- structural, grounding, and rigor diagnostics are applied;
- miner and Bronze candidates are aligned and adjudicated without revealing candidate origin to the judging model;

- equivalent candidates are merged into a canonical Silver record;
- miner score combines Silver coverage, diagnostic quality, and adjudication quality;
- a winner-takes-most rank curve converts batch performance into proposed weights.

These operations test whether miners can return usable records, whether the Bronze–Silver comparison is stable, and where the production path fails. They also generate evidence about score dispersion, cost, latency, and common parsing or grounding errors.

v0 does not yet provide the full properties intended for v1. It lacks a maintained Gold layer, post-submission paper sampling, miner judging, behavioral clustering, marginal-contribution payout, canonical production-graph ingestion, and a decentralized validator set. Its records and metrics are experimental evidence about the mechanism, not Gold-anchored claims about final database quality.

3.1.3 v1: the initial production architecture

v1 adds the mechanisms needed for dependable, verification-tiered data production:

- persistent hidden synthetic challenges and selective trusted Gold;
- a separate judging task for difficult candidate pairs;
- hidden post-commit paper and field sampling to bound validator cost;
- separate extraction and judging reliability estimates;
- semantic item identifiers and behavior-based strategy families;
- family-level sampling, marginal-contribution scoring, and payout ranks;
- frozen reference releases promoted through held-out Gold tests;
- random re-audit of accepted and apparently undisputed content;
- verification-tiered canonical ingestion and historical record patching.

The additions change the epistemic status of the output, not merely the scale of the workload. They create independent calibration, identity-aware accounting, controlled learning, and database states that can support production use.

3.1.4 Version matrix

The following matrix keeps the operational v0 objects and their v1 successors side by side.

Mechanism object	v0 calibration	v1 production direction
Bronze	Reference extraction on every scoring paper.	Frozen release, sampled use where appropriate, Gold-gated promotion.
Silver	Origin-blind AI adjudication and canonical machine record.	Better calibrated, Gold-trained, confidence-aware, with difficult cases escalated.
Gold	Not active.	Synthetic known-answer cases, selective human or other trusted verification, and provenance-labeled high-confidence resolution.
Miner task	Extraction.	Extraction plus separate judging.
Selection	Qualification, performance, and rotation at UID level.	The same exploration logic applied to strategies and behavioral families.
Reward	Winner-takes-most miner rank.	Family rank, marginal contribution, and within-family division.
Database	Experimental outputs; no canonical production ingestion.	Record-level verification states, lineage, and reversible graph updates.
Validator topology	One owner-operated validator.	One substantive reference validator at initial deployment, with independent replay and validator expansion as the mechanism matures.

The version boundary is functional. v0 asks whether the Bronze–Silver loop can operate reliably and produce informative measurements. v1 asks whether selective verification, controlled learning, and identity-robust rewards can produce records that satisfy maintained quality standards. The next subsection describes the three verification layers, and Section 4 develops the incentives that connect them.

3.2 The Bronze–Silver–Gold Quality Architecture

Running the strongest available evaluation on every paper would be too costly, while relying on one economical model would leave systematic errors unexamined. CLAIMS allocates verification effort through three layers. Bronze supplies a common floor, Silver resolves the cases where stronger machine judgment has expected value, and Gold provides sparse independent truth for calibration, escalation, and learning.

3.2.1 Bronze: an economical quality floor

Bronze is the extraction produced by the current reference pipeline. It must be capable enough to serve as a useful starting point and inexpensive enough to run broadly. Miners can reproduce, inspect, or incorporate the reference system, but they gain a meaningful advantage only when they add information beyond what Bronze already supplies.

Bronze performs three economic functions:

1. it prevents the subnet from paying repeatedly for performance below a known baseline;
2. it gives validators a common comparison object, which lowers the cost of evaluating heterogeneous submissions;
3. it directs miner effort toward residual errors and omissions, where experimentation has the greatest expected value.

Bronze remains a baseline rather than ground truth. A faithful copy can provide availability, yet it adds no new information. A valid miner correction can outperform the baseline and, after verification, improve a later reference release.

3.2.2 Silver: selective, high-quality machine adjudication

Silver is the strongest machine-curated record economically justified for a dispute. It may use more capable models, more context, richer adjudication fields, stronger verification routines, or an ensemble that would be too expensive to apply to every paper.

Silver decides whether differently worded candidates are equivalent, whether a miner-only claim is genuinely present, whether a Bronze element should remain, and whether the available evidence is too uncertain for machine resolution. The decision is made at the level of canonical items rather than whole submissions.

The resulting record retains undisputed Bronze content, accepts verified miner improvements, merges equivalent representations, and preserves ambiguity when the source does not support one answer. This item-by-item assembly can combine the strongest contributions from several methods without selecting an entire miner output as the winner.

3.2.3 Gold: sparse trusted verification with explicit provenance

Gold denotes the highest verification layer used by the mechanism. It does not promise universal certainty. A Gold observation can come from:

- an oracle-known synthetic transformation created by the validator;
- a selective human extraction, reconciliation, or review;
- a later independently resolved case;
- a high-confidence consensus result whose provenance and lower epistemic status remain explicit.

Because Gold is expensive, it remains sparse. Its role is to supply enough independent truth to calibrate evaluators, expose systematic errors, gate model releases, and teach the lower-cost layers. Full human annotation of every paper is neither required nor implied.

Gold provenance stays attached to the record because the different sources carry different meanings. A known synthetic answer, a trusted human judgment, and a high-confidence consensus may all support calibration, but the database should not collapse them into one unexplained label.

3.2.4 The quality flywheel

Verification moves from low cost to high cost:

$$\boxed{\textit{Bronze} \longrightarrow \textit{Silver} \longrightarrow \textit{Gold}.}$$

Learning moves in the reverse direction:

$$\boxed{\textit{Gold} \longrightarrow \textit{Silver} \longrightarrow \textit{Bronze}.}$$

A Gold correction becomes a high-value training and evaluation case for the Silver judge. A stronger Silver judge can produce a much larger set of corrected machine labels. Those labels can then improve the Bronze reference miner that runs on every paper. Sparse, expensive supervision can therefore affect the broad production process many times rather than being consumed once. Figure 3 combines the operational escalation and reverse learning paths.

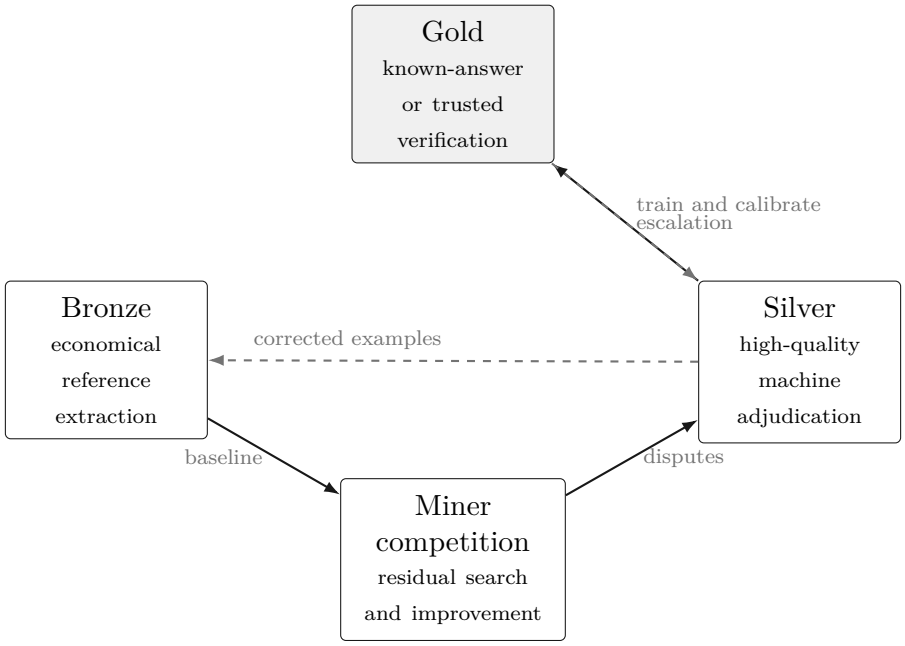

Figure 3. Production escalates from economical extraction to stronger verification, while verified corrections flow back to improve the cheaper layers.

The two directions form one system. Operational escalation spends additional resources only where uncertainty warrants it, while reverse learning transfers verified information back toward the inexpensive layer used at scale.

3.2.5 Synthetic challenges as scalable Gold

A synthetic challenge begins with a valid record X and applies a controlled corruption T_k from error class k:

$$\widetilde{X} = T_k(X).$$

Possible corruptions include deleting a qualifier, reversing polarity, presenting association as causation, perturbing a number, and swapping a unit or comparator. Other classes change a source anchor, link evidence to the wrong result, duplicate a paraphrase, split or merge claims incorrectly, or reverse claim–evidence roles. Because the validator creates the mutation, the correct disposition is known without commissioning a complete new human extraction.

Synthetic challenges serve three different functions:

1. **Effort discipline.** When synthetic and genuine cases look operationally alike, participants cannot reserve careful work for recognizable tests.
2. **Reliability estimation.** Performance can be measured by capability class, difficulty, confidence, and false-positive behavior.
3. **Training.** Exact labels become reusable examples for Silver and, through Silver, for Bronze.

A synthetic answer does not resolve an unrelated genuine dispute. Its value depends on transport: performance on the synthetic error class must predict performance on genuine cases from the same class. Challenge generators therefore require tests of representativeness and regular renewal so that participants cannot learn a shortcut detector.

3.2.6 The basic quality condition

The selective-verification system improves the record only when successful correction outweighs the damage caused by false challenges and poor adjudication. For a representative canonical item, let b be the Bronze error rate, d the probability that miners expose a Bronze error, and f the probability that miners create a dispute when Bronze is correct. Let q_1 be final error conditional on an exposed Bronze error, and let q_0 be final error conditional on a false challenge. Final-record error is

$$\boxed{e_F = b(1-d) + bdq_1 + (1-b)fq_0.}$$

The curated record improves on Bronze exactly when

$$\boxed{bd(1-q_1) > (1-b)fq_0.}$$

The left side measures Bronze errors that miners expose and the system successfully corrects. The right side measures correct Bronze content damaged by a false challenge and an erroneous final decision. The inequality therefore supports payment for verified correction rather than raw disagreement. It also shows why challenge precision becomes more important as Bronze improves and leaves fewer errors to find.

3.2.7 Conditions for compounding improvement

The learning loop compounds only when several empirical conditions hold:

- synthetic cases remain difficult to recognize as synthetic;
- challenge error classes represent failures that matter on genuine papers;
- every material failure class retains a positive probability of audit;
- Gold provenance and confidence remain attached to each case;
- Silver and Bronze are evaluated on held-out cases that were not used for training;
- reference releases are frozen and promoted only after independent regression testing;
- apparently undisputed content is occasionally re-audited;
- historical records can be patched or reprocessed after material improvements.

Without these safeguards, correlated mistakes can persist or spread through the learning cascade.

Improvement claims must therefore come from measured performance across frozen releases, not from the mere existence of a training loop.

The quality architecture defines how truth is purchased and reused. Section 4 determines which participants receive opportunities to search for improvements, how their work is evaluated, and how accepted information becomes reward.

4 Miner, Validator, and Economic Mechanisms

The participant mechanisms allocate access, evaluation work, reward, and cost recovery. Miner rules reward distinct accepted information, validator rules make substantive judgment observable and testable, and the economic conditions determine whether capable methods and adequate verification can remain in the market.

4.1 The Miner Mechanism

Miners search for better extraction methods under a common output contract. The protocol observes their records rather than their internal effort, so selection and reward must distinguish useful independent information from copied baselines, superficial variation, and duplicated identities. The mechanism combines access lanes, concentrated competition, behavioral-family accounting, and item-level contribution.

4.1.1 Private strategies, standardized outputs

A miner chooses a private production strategy: model family, prompts, tools, retrieval process, parsing stack, verification routines, and compute allocation. The protocol does not need to observe those choices directly. It standardizes the requested record and evaluates the information that the strategy returns.

Useful effort has several dimensions. A miner can invest in coverage, source grounding, numerical fidelity, semantic fidelity, causal classification, independent discovery, latency, or cost reduction. The score must expose enough of these dimensions that a cheap shortcut cannot imitate the desired action on every rewarded channel.

Standardized outputs allow private experimentation without making the evaluation arbitrary. Miners can differ radically in method while validators compare their results through the same schema, canonicalization rules, and source-grounding requirements.

4.1.2 Qualification, performance, and rotation

Evaluation opportunities are scarce because each submission can create validation cost. Claims allocates those opportunities through three distinct lanes:

1. **Qualification.** New or under-evaluated strategies receive enough assignments to establish quality and behavioral identity.
2. **Performance.** Strategies with demonstrated quality and contribution receive more assignments.
3. **Rotation.** Qualified strategies that have not been tested recently, or that appear behaviorally novel, retain a route back into evaluation.

The lanes balance entry, exploitation, and exploration. A performance-only rule can entrench early incumbents and suppress new methods. A pure lottery spends validation on weak strategies and makes identity multiplication attractive. A qualification quota by itself can be farmed through repeated new accounts.

Family-level accounting becomes essential once enough shared-task evidence exists. It prevents one redundant method from acquiring several lanes through several identities while preserving opportunities for genuinely different strategies.

4.1.3 Almost-winner-takes-all competition

CLAIMS concentrates much of the performance reward near the frontier. In v0, the calibration assigns 70 percent of the performance pool to rank one and distributes a geometrically declining 30 percent tail across ranks two through five. Winning has a large payoff, while several near-frontier miners can still recover part of their costs and remain available as alternative methods.

Prize concentration and score intensity address different questions. Score intensity determines how strongly verified performance changes expected reward relative to cost. Prize concentration determines how sharply the performance pool is divided by rank. In a fixed-participation contest, greater concentration can raise effort. The Claims setting also includes participation decisions, noisy scores, risk, method diversity, false challenges, and duplicated compute. The preferred concentration is therefore a versioned empirical control rather than a universal constant.

The v1 design retains three commitments:

- the strongest verified contribution receives a dominant share;
- near-frontier methods can earn enough to remain viable;
- behaviorally redundant strategies cannot occupy several payout ranks.

These commitments preserve competitive pressure without assuming that one sampled winner contains all useful information. They also separate the value of a frontier method from the number of accounts that happen to run it.

4.1.4 Strategy identities and behavioral families

A network identifier establishes routing, not informational independence. CLAIMS distinguishes four levels:

$$\boxed{\text{UID: routing} \longrightarrow \text{hotkey: account} \longrightarrow \text{strategy version} \longrightarrow \text{behavioral family.}}$$

A miner may declare a strategy and version, but the protocol infers independence from behavior rather than accepting the label. Strategies are compared on shared tasks using overlap in accepted items, evidence spans, structured values, false positives, omissions, and synthetic-challenge errors. Shared unusual errors provide stronger evidence of a common information source than shared correct answers.

A behavioral family is an accounting category for statistically redundant information. It does not assert shared ownership, collusion, or misconduct. Independent miners can enter one family because they use the same public model and make the same errors. One actor can operate several separate families when the strategies produce genuinely different information.

Family assignment affects redundancy-sensitive selection and reward, while the underlying account and strategy histories remain visible. The separation lets the mechanism control repeated information without pretending to infer legal or economic identity from model behavior alone.

4.1.5 Marginal contribution at the item level

After adjudication, each accepted contribution maps to a stable final-record item h. Let k_h denote the number of distinct behavioral families that supplied that item. An initial v1 calibration assigns total item value

$$\boxed{V_h = 1 + \beta \mathbf{1}\{k_h \geq 2\},}$$

where the current corroboration parameter is $\beta = 0.20$. The first independent family creates discovery value. A second family can add a bounded corroboration premium. Third and later families do not raise total item value. Each supplying family receives V_h/k_h, and several strategies from one family divide the family share instead of multiplying it.

The rule embeds four design choices:

1. only content accepted into the final record receives contribution credit;
2. canonicalization removes duplicate wording before credit is assigned;
3. independent corroboration can add value without allowing support to grow linearly with identity count;
4. reward follows the data product assembled by the mechanism rather than one all-or-nothing submission.

A family's batch score combines its strongest standalone quality with its total marginal contribution. The current proposal gives greater weight to marginal contribution because reproducing information that is already available adds less value than a verified correction, omission recovery, or independent confirmation.

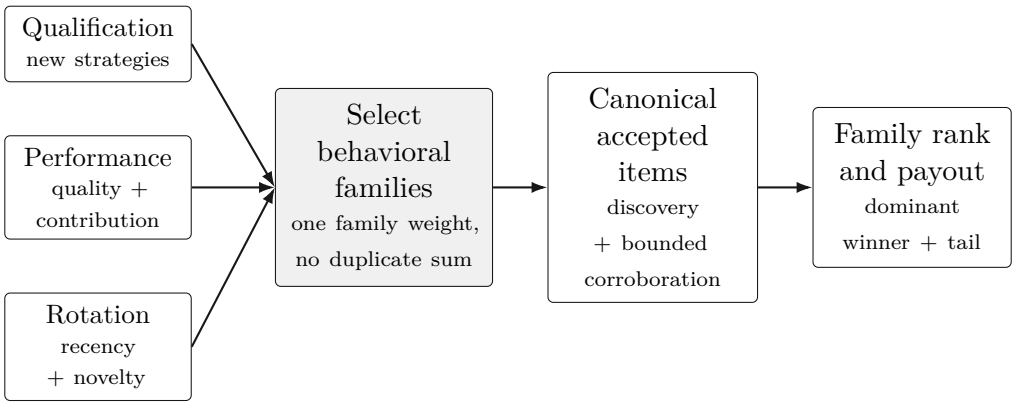

Figure 4. Selection begins in three access lanes, but contribution accounting and payout operate at the level of behaviorally distinct information families.

Figure 4 shows the two levels at which the mechanism operates. Qualification, performance, and rotation determine who is evaluated. Canonical accepted items and behavioral families determine what information earns value and how that value is divided.

4.1.6 The duplicate-invariance result

Consider an exact or highly redundant strategy that the clustering process correctly assigns to an existing family. The family receives the maximum member selection weight rather than the sum, occupies one payout rank, and holds item credit at the family level. Adding a copy cannot increase total family assignment weight, accepted item value, or payout; it only divides the existing family reward among more strategies.

This conditional property is the practical information-level defense against Sybil splitting in v1. It relies on accurate clustering, stable canonicalization, and a scoring context that the family cannot change with its own copies. When a duplicate evades clustering and appears to be independent, the bounded corroboration rule limits item-value inflation to β instead of allowing value to rise with every additional identity.

A full split-Sybil guarantee demands more. It requires an enforceable resource, entitlement, or context rule that prevents one actor from multiplying total economic exposure through undetected identities. Registration burn, UID scarcity, immunity periods, per-hotkey limits, and qualification quotas raise cost or slow bulk entry, but they do not link identity resets. Behavioral clustering and marginal contribution control repeated information; resource linkage remains a complementary design frontier.

4.1.7 Leave-family-out context

Even accurate family accounting can fail if a participant changes the benchmark used to score itself. The mechanism therefore constructs a leave-family-out context whenever feasible:

- the evaluated family is excluded from its own judging panel;
- multiplicity within that family does not change the comparison or consensus baseline;
- task difficulty and normalization are estimated without allowing the family to move its own reference point;
- several candidate variants from one family cannot appear to corroborate one another.

This counterfactual context reduces improvement spam and prevents a family from manufacturing both a disputed object and the apparent support used to validate it. It also makes the duplicate-invariance property meaningful beyond the payout formula.

4.1.8 Miner participation and cost recovery

A miner bears source-access, parsing, inference, orchestration, numerical-verification, storage, engineering, and experimentation costs. Independent extraction will be supplied only when its expected reward advantage over copying Bronze can cover the incremental expense.

Public evaluation should therefore answer more than which miner won. It should report:

- how many evaluations a capable new strategy needs before its first positive payout;
- which quality level is required for expected break-even under low-, medium-, and high-cost pipelines;
- whether several independent families remain economically viable;
- whether reward concentration pays for true quality or for upper-tail sampling noise;
- whether additional compute creates independent information or duplicated expenditure.

These questions connect the incentive design to an actual supply response. A contest that selects a strong winner but leaves no credible path for alternative high-quality methods would weaken the method search that motivates the subnet. The next subsection develops the validator mechanism that must measure quality while controlling its own cost and correlation.

4.2 The Validator Mechanism

Selective verification works only when validators perform real scientific evaluation and can themselves be assessed. Claims therefore treats validation as a production activity with observable outputs: blinded adjudications, known-answer calibration results, replayable records, and release decisions. Validator count alone provides little assurance when the underlying judgments are cheap, copied, or correlated.

4.2.1 Validators perform substantive evaluation work

A Claims validator constructs tasks, generates Bronze, checks sources and schemas, aligns candidates, performs Silver adjudication, creates synthetic challenges, calibrates judges, selects audits, canonicalizes records, assigns database states, tests releases, and reports results publicly. These activities determine which miner outputs become trusted data.

The role is costly by design. A validator must spend resources to distinguish a grounded correction from a plausible error and to reveal when its own judging process fails. Adding identities without adding this work does not increase scientific assurance.

Validator credibility therefore comes from the quality, independence, and reproducibility of judgment. The mechanism observes those properties through hidden known-answer items, clean controls, overlap audits, and records that allow material decisions to be replayed.

4.2.2 Origin-blind Silver adjudication

v0 already hides candidate origin from the adjudicator. The judge receives the relevant source evidence and anonymous candidates, without labels that identify the reference output or the miner. This blocks two easy shortcuts: treating Bronze as correct because it is the baseline, and treating a miner proposal as suspect because it challenges the baseline.

The judge returns structured findings rather than assigning reward directly. Deterministic calculations then convert those findings into coverage, diagnostic, and adjudication components. Separating substantive issue detection from reward arithmetic improves reproducibility and limits prompt-sensitive discretion.

Origin blindness does not guarantee accurate judgment, but it removes one nuisance signal that should have no bearing on the content. Known-answer calibration is still needed to measure whether the blinded process reaches the correct result.

4.2.3 The synthetic-error validator loop

v1 mixes intentionally corrupted records and clean controls into otherwise realistic evaluation packets. Because the system knows the correct disposition of those items, it can estimate:

- seeded-defect detection by error class;
- false-positive rates on clean controls;

- confidence and abstention calibration;
- shared error patterns among judges;
- divergence between synthetic and genuine-case behavior;
- whether challenge difficulty distinguishes reliable participants.

Challenge difficulty is a design variable. A suite that every judge passes provides little separation. A suite that almost every judge fails creates a signal too noisy for dependable weighting. The validator therefore adapts the future share of easy, medium, and hard cases to maintain useful dispersion, while fixing each challenge distribution before observing the responses it will score.

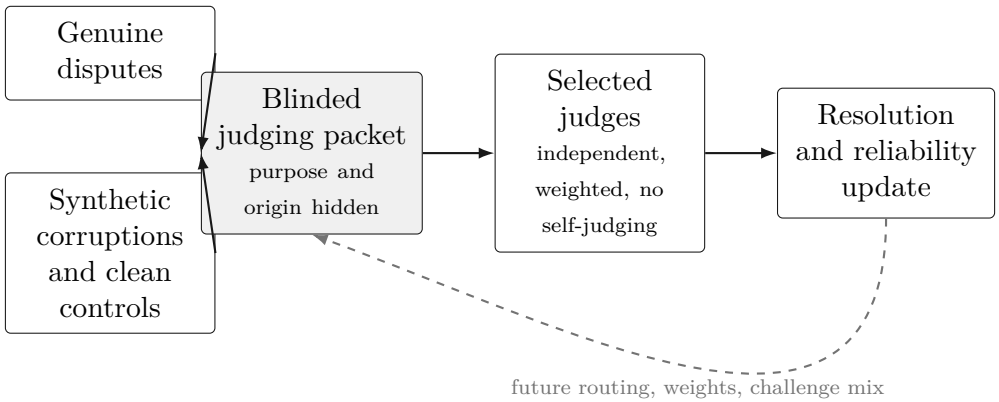

Figure 5. Genuine disputes, synthetic corruptions, and clean controls enter the same blinded judging process, allowing known answers to calibrate decisions on unknown cases.

Figure 5 shows how known answers calibrate the same process that handles unknown genuine disputes. That connection is valuable only when synthetic and genuine cases are operationally similar and performance transports across them. Divergence between the two becomes a diagnostic rather than evidence that synthetic accuracy has solved the real task.

4.2.4 Judge aggregation under correlation

Reliability and independence are separate properties. Five judges that use the same model and prompt do not provide five independent observations. Judge aggregation therefore includes:

- reliability weights anchored by known-answer tasks;
- a maximum individual vote weight;
- caps on the total weight of highly correlated judge families;
- exclusion of the evaluated miner and its linked family;
- confidence-aware resolution and an unresolved outcome when evidence is insufficient;
- later score updates when a genuine dispute receives independent resolution.

For k reports with pairwise correlation ρ, a useful approximation to the number of independent observations is

$$n_{\text{eff}} = \frac{k}{1 + (k-1)\rho}.$$

When reports are nearly exact copies and $\rho \approx 1$, their effective count is close to one. Public reporting should therefore show effective panel size alongside nominal panel size.

Agreement on a genuine case supplies evidence, but it does not establish truth. A rule that pays only for matching the eventual majority can reward prediction of other judges instead of careful assessment of the paper. Known-answer calibration, abstention, delayed independent resolution, and correlation adjustment are necessary before consensus can support a high verification tier.

4.2.5 Replayable audit records

Each production batch leaves an audit record that is sufficient to reconstruct material decisions. At a high level, the record includes:

- batch, paper, and release identifiers;
- selected miners, strategies, behavioral families, and selection lanes;
- sampling commitments and selected audit items;
- hashes or identifiers for Bronze and candidate records;
- canonicalization, relation, and adjudication findings;
- Silver and Gold decisions with confidence and provenance;
- score components, marginal contribution, and family-level payout;
- validator failures, excluded cases, cost, and timing summaries.

Replayability and challenge secrecy create a real tradeoff. Retired challenge families can be disclosed in detail. Active generators may require delayed or coarser disclosure so that accountability does not create a shortcut for recognizing evaluation cases.

The audit record resolves this tension through commitment and later reconstruction. Material inputs and decisions can be fixed when they occur, while sensitive details become public after the secrecy window no longer protects an active challenge distribution.

4.2.6 Validator bootstrapping

v0 and the initial v1 plan rely on one owner-operated substantive validator. The architecture therefore does not claim that scientific validation is already decentralized. Initial assurance comes from origin-blind judging, hidden known-answer tasks, deterministic score calculation, frozen releases, signed artifacts, committed sampling, and reproducible aggregate reports.

As independent validators enter, they must perform costly and meaningfully distinct evaluation rather than replay centrally computed scores. Shared Gold tasks, overlap audits, independent error discovery, reproducibility, and correlation-adjusted agreement can compare their performance.

Additional validators create value when they contribute independent evidence and expose common-mode failures. A larger node count without distinct verification work would add nominal decentralization while leaving scientific assurance unchanged. The following subsection translates those substantive tasks into participation, cost-recovery, and operating-coverage constraints.

4.3 Miner and Validator Economics

The subnet must finance two costly activities at once: independent production by miners and selective evaluation by validators. Reward concentration can encourage frontier performance, but the mechanism remains viable only when capable methods have a path to recover costs and validators can afford the work needed to distinguish those methods.

4.3.1 The cost of miner work

Miner costs include both variable production and fixed capability building. Variable costs cover model inference, retrieval, parsing, full-text and multimodal processing, source verification, retries, and storage. Fixed costs include pipeline engineering, ontology maintenance, monitoring, and experimentation. A miner can also bear the cost of creating a useful strategy that is sampled too infrequently to demonstrate its value.

The combination of concentrated performance reward, qualification, and rotation addresses different parts of this problem. The dominant prize raises the return to reaching the frontier. The declining tail, repeated assignments, and family-level exploration preserve a credible route to cost recovery for methods close to that frontier.

The relevant economic object is expected reward after selection and evaluation, not the advertised top prize. Entry remains credible only when a capable new strategy can obtain enough observations to overcome uncertainty and when genuinely different families are not crowded out by repeated versions of one method.

4.3.2 The cost of validator work

Validators incur costs for:

- Bronze inference and artifact management;
- deterministic source, schema, and numerical checks;
- candidate canonicalization and semantic comparison;
- higher-cost Silver adjudication;
- synthetic challenge generation and quality control;
- selective Gold review and reconciliation;
- judging, re-audit, release testing, database hosting, and audit reporting.

Selective verification is an economic design because it controls where these costs are spent. Bronze runs broadly because it is economical. Silver is reserved for discrepancies and strategically sampled content. Gold is sparse. Canonicalization prevents exact copies from multiplying Silver work, while post-commit sampling and re-audit limit the validation burden as paper volume grows.

A cost attack can still create many superficially distinct disputes. The security section therefore treats validator expenditure as an object that incentives must protect, not as a fixed overhead outside the mechanism.

4.3.3 Participation constraints and break-even quality

Let audited quality be q. Expected miner profit is

$$\Pi(q) = \mathbb{E}[\text{reward} \mid q] - C_{\text{compute}}(q) - C_{\text{fixed}} - C_{\text{entry}}^{\text{amortized}}.$$

The break-even quality frontier is

$$q_{\text{BE}} = \inf\{q : \Pi(q) \geq 0\}.$$

A sustainable subnet estimates this frontier for a Bronze-copying baseline, a median qualified strategy, top-quartile strategies, and several compute-cost regimes. The maximum prize can be large while the expected path to that prize remains too narrow to support independent entry.

Economic health therefore requires two conditions: capable methods must have a credible expected path to break-even, and the validator must be able to finance the evaluation that separates those methods from cheaper deviations. Reporting both conditions prevents reward size from being confused with market viability.

4.3.4 Cost per trusted record

The main scaling objective is a declining cost per accepted, quality-adjusted record:

$$C_t^{\text{trusted}} = \frac{\text{Bronze, Silver, Gold, audit, and canonicalization cost}}{\#\{\text{new audited or Gold claim–evidence links}\}}.$$

Each part of the quality flywheel can lower this ratio. Better Bronze reduces routine error. Better Silver resolves more disputes without Gold. Better canonicalization avoids repeated adjudication of duplicates. Better miner strategies discover corrections with less expenditure.

A larger database creates value only when verified output grows without violating the maintained quality floor. Raw paper count or claim count remains useful as an operational diagnostic, but it cannot replace the trusted-record denominator.

4.3.5 Emissions, product revenue, and operating obligations

Three financial layers remain separate:

1. **Protocol emissions** reward miner production and validator evaluation.
2. **Product revenue** comes from data access, APIs, custom databases, evaluations, or other services built on the evidence layer.
3. **Owner and operating resources** fund reference infrastructure, Gold acquisition, canonical storage, release engineering, and commercial development.

Keeping the layers distinct reveals whether the subnet pays for useful work, whether the resulting data product has downstream value, and whether verification is adequately funded. Token price alone does not answer any of those questions.

The measurement architecture in Section 5 therefore emphasizes production yield, miner break-even, cost per trusted record, and validator operating coverage. Those measures also expose security failures that can preserve apparent quality while making the system economically unsustainable.

5 Measurement, Security, and Release Integrity

Public measurement and security controls make the scientific and economic conditions of the mechanism observable. The status vector reports non-substitutable dimensions of performance; the threat model and release controls protect both the accepted record and the production process that sustains it.

5.1 Metrics and Public Accountability

The mechanism has no single sufficient health measure. A system can increase output while reducing accuracy, expand miner count while concentrating information in one behavioral family, or improve average quality while a critical component fails. Public reporting must therefore show a compact set of non-substitutable measures tied to the architecture's scientific, economic, and governance conditions.

5.1.1 A status vector, not one health score

Claims pursues data quality, useful production volume, miner diversity, Sybil resistance, validator assurance, cost efficiency, and participant viability. Combining all of them into one weighted average would allow strength in one dimension to conceal failure in another.

A status vector keeps each indispensable channel visible. Every reported quality statistic identifies its denominator, sample-selection rule, confidence interval, and applicable Bronze, Silver, Gold, schema, and model release. v0 measures are labeled as Silver-relative experimental proxies. A v1 measure becomes Gold-anchored only when an independent Gold process supports that description.

The vector is intentionally compact enough to monitor, but each headline measure retains the companion information needed to interpret it. The following families connect directly to the mechanism's quality and incentive conditions.

5.1.2 Headline metric families

A public dashboard must separate objectives that can otherwise conceal one another. The following metric families pair each headline measure with the companion information needed to interpret performance, diagnose failure, and compare releases on a consistent basis.

Objective	Headline measure	Required companion information
Gold-anchored data quality	Claim precision and recall; claim–evidence edge F-score; conservative lower confidence bound.	Audit-set size; paper and error classes; Bronze, Silver, and final quality on the same sample; Gold provenance mix.
Net lift over Bronze	Corrected Bronze errors minus correct Bronze items damaged by challenge and adjudication.	Correction rate, damage rate, false-challenge rate, Silver error by confidence class.
Verified data velocity	New audited or Gold claim–evidence links per day.	Papers and raw units per day; median and tail latency; re-audit reversals; verification-state mix.
Effective miner market	Effective number of behavioral families based on accepted contribution or reward share.	Selection funnel, positive-payout breadth, reward concentration, time to first payout, cluster size and turnover.
Marginal contribution and Sybil resistance	Unique accepted correction, bounded corroboration, and reward under simulated identity splits.	Duplicate load, leave-family-out scores, selection duplication ratio, cluster false-merge and false-split tests.
Validator assurance	Seeded-defect recall and clean-control false-positive rate.	Confidence calibration, abstention, reproducibility, effective panel size, overlap audits, Gold reversal by error class.
Economic viability	Cost per audited claim–evidence link and miner break-even quality.	Bronze, Silver, Gold, storage, and audit cost; selection probabilities; participant retention; validator operating coverage.

The rows are deliberately non-substitutable. For example, high verified data velocity cannot compensate for a failing Gold quality floor, and a large nominal miner population cannot compensate for one family supplying most accepted information.

5.1.3 Gold-anchored quality floor

For claim units, let P^G_{claim} and R^G_{claim} denote precision and recall on a maintained Gold audit set. Let F^G_{edge} measure the quality of claim–evidence links, including grounding and relation correctness. A conservative headline measure is

$$Q^G_{\text{floor}} = \min\left\{P^{G,\text{LCB}}_{\text{claim}}, R^{G,\text{LCB}}_{\text{claim}}, F^{G,\text{LCB}}_{\text{edge}}\right\},$$

where LCB is a published lower confidence bound. Taking the minimum prevents strength in one component from hiding failure in another.

The report also shows the underlying components, audit-set composition, and Gold provenance. The floor is a guardrail, not a substitute for the richer error profile needed to improve the system.

5.1.4 Net lift and learning velocity

Agreement with Bronze does not show that the mechanism improves the record. At release t, let B_t denote Bronze, let F_t denote the final record, and let X denote the canonical paper label.

Define net lift by

$$L_t = \Pr(B_t \neq X,\, F_t = X) - \Pr(B_t = X,\, F_t \neq X).$$

The first term counts successful corrections of Bronze; the second counts damage to correct Bronze content. Both components are reported separately so that the same net value cannot conceal very different correction and harm rates.

Learning velocity compares frozen releases on held-out data. Release-to-release reductions in Bronze and Silver error show whether Gold corrections actually transfer through the cascade rather than merely improving the cases already used for training.

5.1.5 Effective competition and split-gain tests

Let s_g be behavioral family g's share of accepted contribution or reward. The effective number of miner families is

$$N_{\text{eff}}^{M} = \frac{1}{\sum_g s_g^2}.$$

This statistic falls when one family dominates, even if many UIDs are registered. It is reported with the number of families receiving positive payout, the selected-to-completed funnel, reward concentration, and the longest time a qualified family has gone without evaluation.

A direct Sybil test compares the combined reward under a k-identity split with the counterfactual reward of the consolidated family:

$$SGR_k = \frac{\text{total reward under a } k\text{-identity split}}{\text{counterfactual consolidated-family reward}}.$$

For exact and highly redundant splits, the target is $SGR_k \leq 1$. The same experiment tests assignment probability and evidentiary weight, because a split can distort production even when the payout formula alone appears invariant.

Together, effective family count and split-gain simulations distinguish nominal participation from independent competition. They also reveal where clustering errors or context manipulation still create an advantage.

5.1.6 Reporting contract

Every headline card shows:

- the current value and a meaningful 7-, 30-, or 90-day trend;
- a target, acceptable band, or confidence interval;
- the denominator, sample size, and audit-selection method;
- applicable release identifiers and the metric-definition version;
- whether the measure is a v0 proxy or a v1 production statistic;
- a replayable aggregate sufficient to recompute the value;
- a changelog when the metric, audit distribution, or denominator changes.

Metric definitions remain fixed within each reporting release. A better definition can start a new series, but it does not silently rewrite the historical record.

This reporting contract turns the architecture's scientific and economic conditions into public evidence. The security analysis that follows uses the same measures to distinguish record robustness from economic robustness.

5.2 Security, Failure Modes, and Release Integrity

Scientific extraction can fail through ordinary error, strategic manipulation, or a release process that teaches the system its own mistakes. These channels can damage either the accepted record or the economics required to produce it. Claims therefore uses observable signatures, layered defenses, and frozen releases rather than relying on assumptions about participant intent.

5.2.1 A threat model for scientific data production

A weak model and a deliberate attacker can create the same distribution of outputs. The mechanism applies the same statistical treatment when those outputs fabricate content, omit difficult material, multiply apparent support, or impose disproportionate validation cost. The threat model focuses on the behavior's effect rather than on an unverifiable motive. Table 1 organizes the observable signatures, primary defenses, and residual risks.

Table 1. Observable failure modes, primary defenses, and risks that remain after those defenses.

Failure mode	Observable signature	Primary defenses	Residual risk
Fabricated content	Claims, quantities, or evidence links cannot be grounded to the source	Source anchors; deterministic checks; origin-blind Silver review; false-positive penalties; selective Gold audit	Fluent fabrication that passes weak source matching
Shallow extraction	High abstract coverage but systematic omissions from methods, results, tables, figures, or supplements	Coverage manifests; hidden omission tests; paper- and field-level sampling; accepted miner-only discoveries	Access and parser differences may resemble low effort
Semantic flooding	Many distinct-looking but unsupported or immaterial candidate records	Canonicalization; stable item IDs; per-family marginal value; challenge penalties; adjudication budgets	Wording variation can increase comparison cost before canonicalization
Duplicate identities	Several accounts submit nearly identical outputs or share unusual errors	Behavioral-family clustering; one family selection weight and payout rank; registration friction; qualification quotas	Cluster evasion and identity reset before sufficient shared evidence
Benchmark imitation	A miner copies Bronze or exploits stable judge quirks instead of extracting independently	Accepted-improvement rewards; hidden challenges; versioned and diverse audits; leave-family-out context	Public reference access necessarily permits some imitation
Validator copying or correlation	High agreement without independent discovery; lagged or highly correlated judgments	Commit–reveal where applicable; hidden Gold; correlation-adjusted vote caps; independent shadow runs; replay audits	Several validators can share a foundation-model blind spot
Challenge recognition	Synthetic cases receive different latency, confidence, or accuracy from genuine-looking cases	Mixed packet composition; evolving private generators; delayed retirement; transport tests	Public disclosure can train detectors of obsolete challenge families
Recursive self-training	Silver and Bronze become increasingly confident in a shared error	Frozen releases; independent holdouts; regression suites; Gold provenance; retrospective audits	Sparse Gold may miss a systematic but rare failure mode
Context manipulation	A family changes the benchmark, consensus, or difficulty estimate used to score itself	Leave-family-out construction; self-exclusion from judging; precommitted samples and parameters	Full counterfactual reconstruction may be computationally costly
Cost attack	False disputes or unique-looking records force expensive Silver or Gold review	Cheap checks first; canonicalization; family rate limits; score reserves or bonds; capped escalation budgets	A sophisticated attacker may impose cost without changing the final record

No single defense covers the full table. Registration cost can limit cheap bulk entry without establishing independence. Consensus can aggregate reports without creating truth. A high challenge score measures performance on the challenge distribution without proving accuracy on every genuine dispute.

The defenses are therefore complementary. Source grounding protects the record, family accounting protects evidentiary and economic weight, hidden tests discipline effort, and release controls limit the propagation of errors that survive the first three layers.

5.2.2 Data-quality robustness and economic robustness

A mechanism can preserve the final record while remaining economically fragile. Exact duplicates may be removed before adjudication and add little beyond ingestion cost. Semantically varied false challenges can force many unique comparisons and expensive disputes even when Silver rejects every one correctly.

Let N_t be the number of eligible source units, M_t the number of raw submissions, U_t the number of unique canonical candidates, D_t the number of disputes, and G_t the number of Gold escalations. A reduced-form operating cost is

$$\begin{aligned} C_t = {} & N_t c_B + M_t c_{\text{in}} + U_t c_{\text{can}} \\ & + D_t c_S + G_t c_G + R_t, \end{aligned}$$

where R_t is the reward pool. The design must control accepted-record error and keep U_t, D_t, and G_t tied to genuine information rather than to identity count or manufactured disagreement.

A participant with perceived challenge-success probability z, reward J, loss L after rejection, and private cost c submits a challenge when

$$zJ - (1 - z)L - c \geq 0.$$

The threshold must invite difficult genuine corrections while discouraging speculative flooding. When a rejected challenge creates material validator cost, a score reserve, bond, or equivalent expected loss can internalize part of that burden. The chosen instrument remains a versioned control rather than a permanent constitutional feature.

This distinction between record robustness and cost robustness prevents a well-defended database from concealing an unsustainable evaluation process. Both outcomes must appear in metrics and release tests.

5.2.3 Frozen releases and protection against self-training

The Bronze–Silver–Gold cascade creates leverage because one verified correction can affect many future records. The same leverage creates a poisoning channel. Each production release therefore identifies, at minimum:

- the reference model, weights, and inference configuration;
- extraction prompts, skills, or orchestration logic;
- document readers, parsers, and preprocessing;
- schema, ontology, and canonicalization rules;

- Silver models, rubrics, confidence rules, and aggregation logic;
- Gold generators, active evaluation distributions, and release holdouts.

A candidate release is promoted only after evaluation on private or independently maintained Gold, held-out mutation tests by capability class, regression cases drawn from earlier corrections, and cost and throughput benchmarks. Training examples remain separate from release holdouts. A reported improvement is therefore a comparison of frozen releases on an independent evaluation set, rather than a retrospective judgment by the new system about itself.

An improved Bronze release changes future papers automatically, but it does not automatically repair the historical graph. A material release change can trigger targeted re-extraction, re-adjudication of affected classes, record patches through lineage links, or retrospective audit. Current-pipeline convergence and historical-database correction remain distinct obligations.

5.2.4 Challenge secrecy and public auditability

Active synthetic challenges must remain difficult to recognize, while evaluation and reward decisions must remain inspectable. Claims separates the secrecy of a live challenge distribution from the later auditability of decisions made with it.

A public record can disclose commitments, sample identifiers, hashes, release versions, score components, aggregate performance by error class, and retired challenge families. Details that would reveal an active generator may be delayed or coarsened. After the relevant secrecy window closes, committed inputs allow material reward decisions to be reproduced.

This design makes secrecy temporary and purpose-specific. It protects calibration while a challenge is active without turning hidden evaluation into an unreviewable source of discretion.

5.2.5 Procedural safeguards and validator conflicts

During bootstrapping, an owner-operated validator performs the primary production evaluation. Anonymous candidate packets do not by themselves remove institutional conflicts, especially when team-linked miners also participate. The initial safeguards are procedural:

- common scoring and sampling rules for team-linked and external miners;
- origin-blind candidate packets and self-exclusion from judging panels;
- sample and challenge commitments fixed before outcomes are observed;
- signed artifacts, release identifiers, and deterministic score calculations;
- disclosure of team-linked participation and exclusion rules;
- audit summaries sufficient for independent replay of material decisions.

The long-run objective is independently meaningful validator work. More validators improve the mechanism only when they run distinct evaluation processes, bear real verification cost, and can be compared against common calibration anchors.

Security in Claims is thus a property of the full production path: grounded inputs, identity-aware rewards, calibrated judging, controlled cost, and release discipline. Section 6 places that path within the broader Bittensor design space.

6 Claims in the Bittensor Design Space

The architecture combines a general decentralized intelligence market with a task-specific system for partially verifiable scientific data. Comparing those layers clarifies which functions come from Bittensor, which must be supplied by the subnet, and why mechanisms developed for tasks with complete synthetic truth cannot be copied unchanged into scientific extraction.

6.1 The base protocol and the subnet mechanism

The foundational Bittensor design treats machine intelligence as an economic commodity. Peers evaluate other peers, submit rankings, and receive rewards through decentralized consensus and incentives [1]. The base protocol supplies identities, stake, emissions, weight submission, and network-level consensus.

The subnet must still define the object that miners produce and the evidence that validators use. Claims specifies the claim–evidence record, assigns extraction and judging tasks, constructs Bronze, Silver, and Gold layers, measures marginal information, and decides which records enter the evidence graph.

Yuma consensus can aggregate validator weights, but it cannot determine whether those weights arise from costly, independent, and scientifically meaningful evaluation. That determination belongs to the inner mechanism and its audit trail.

6.2 Minos as the closest documented comparison

Minos offers a close comparison because it uses Bittensor competition for a concrete scientific inference problem. It creates new genomic challenges by injecting hidden synthetic mutations into real sequencing reads. Validators execute submitted miner configurations and score the outputs against known truth, while emissions concentrate on the highest smoothed performer [2]. Each scored round also creates a validated synthetic genome, so the benchmark leaves behind a growing data asset.

Claims adopts three broad design ideas from this architecture:

1. **Blind known-answer tests.** Participants should not be able to identify which apparently ordinary cases determine objective calibration.
2. **Competition over private methods.** The protocol standardizes tasks and outputs while participants experiment with models, prompts, parsers, and workflows.
3. **A productive flywheel.** Evaluation should create reusable data, training examples, and better production infrastructure rather than only a leaderboard.

The two systems differ in their truth structure. Minos can manufacture a reliable answer key for selected mutations and can execute a submitted configuration deterministically. Scientific extraction includes omissions, paraphrase equivalence, numerical grounding, causal interpretation, and unresolved ambiguity in the source. Complete ground truth is expensive or unavailable for most records.

Claims therefore combines several forms of evidence. Known-answer challenges measure selected capabilities, stronger machine adjudication resolves most discrepancies, and sparse trusted review handles a residual set. No single deterministic score covers the entire task. Table 2 summarizes

the resulting division of labor.

Table 2. Bittensor, Minos, and Claims assign different roles to the base protocol, known-answer evaluation, and persistent scientific output.

Dimension	Bittensor base layer	Minos	Claims
Primary object	General market for useful machine intelligence	Variant-calling performance	Canonical claim–evidence data
Miner contribution	Subnet-defined	Tool configuration; later custom algorithms	Structured extraction and, in v1, blinded judging
Ground truth	Not supplied by the base protocol	Hidden synthetic mutations with known labels	Partial: synthetic defects, curated subsets, and selective trusted review
Validator work	Weight assignment and subnet-defined evaluation	Execute configurations and score against truth	Construct Bronze; adjudicate Silver; generate and score Gold; canonicalize and audit
Reward logic	Consensus-mediated emissions	Winner-take-all after score smoothing	Almost-winner-takes-all, family-level ranking, and marginal contribution
Identity defense	Stake and consensus at the network layer	Concentrated prize and blind tasks	Registration friction, hidden tasks, behavioral families, one family rank, bounded corroboration value
Compounding output	Network intelligence and learned rankings	Synthetic genome database and future consensus caller	Evidence graph, hard examples, calibrated judges, and improved reference extraction

The comparison shows where direct transfer is possible and where adaptation is necessary. Blind tests and private-method competition survive the change of domain; complete answer-key scoring does not.

6.3 A general design taxonomy

Scientific and data-oriented subnets can be organized by two questions: whether the work creates a persistent external asset, and whether a complete answer key can be generated at low cost. Four broad cases follow:

1. **Deterministic benchmark tasks** have inexpensive complete labels and can use direct accuracy scoring.
2. **Synthetic-ground-truth tasks** create hidden labels through controlled perturbations, as

Minos does.

3. **Partially verifiable extraction tasks** combine deterministic checks, selective adjudication, and sparse Gold, as Claims does.
4. **Open judgment tasks** lack timely truth and require carefully justified peer-prediction, market, or later-resolution mechanisms.

Mechanism features do not transfer unchanged across these cases. Winner-take-all is attractive when a scalar score is reliable and participation remains viable. It is riskier when sampled quality is noisy and method diversity has option value. Consensus helps when reports contain independent information; it is weak when participants share a model-family blind spot. Synthetic challenges are informative when they are hard to recognize and representative of genuine errors; they can mislead when either condition fails.

Named ecosystem comparisons should be updated as subnet tasks, scoring rules, and validator architectures change. The stable analytical question is which informational or economic problem each feature solves.

Within this taxonomy, Claims is a partially verifiable extraction system whose benchmark activity produces a persistent evidence asset. Governance must preserve that identity while allowing the empirical controls to change with measured performance.

7 Governance, Versioning, and Roadmap

A mechanism that learns from its own verified outputs must be able to change without changing rules after outcomes are known. Claims separates durable commitments about evidence, incentives, and lineage from empirical controls that should respond to measured quality, cost, and participation. Releases make that separation operational.

7.1 Stable commitments and versioned controls

The durable design commitments are:

- source-grounded and provenance-preserving output;
- independent miner production under a common schema;
- Bronze, Silver, and Gold as economically distinct verification layers;
- origin-blind adjudication and known-answer calibration;
- reward for verified marginal information rather than raw identity count;
- separation of miner reward from record-level database status;
- frozen releases, held-out evaluation, reversible lineage, and public diagnostics.

Other quantities remain versioned controls: sampling shares, cluster thresholds and lookback windows, score weights, the corroboration bonus, prize concentration, hidden challenge rate, panel size, confidence thresholds, audit rates, ingestion thresholds, registration cost, and immunity settings.

These controls should change through published releases with prospective evaluation criteria. Their current values are working calibrations, not claims of universal optimality. The stable

commitments define what the mechanism is trying to preserve while those calibrations evolve.

7.2 Release governance

A mechanism release includes:

1. a specification of task schemas, eligible actions, score components, and the payout transformation;
2. version identifiers for Bronze, Silver, Gold, parsers, and canonicalization;
3. a predeployment report on held-out quality, regressions, cost, and throughput;
4. an updated threat model and the known residual risks;
5. a migration rule for miner histories, behavioral clusters, and database records;
6. a public changelog and effective date.

The process fixes evaluation criteria before a deployment is judged. When sampling or challenge secrecy is necessary, commitments can be published before selection and revealed after the relevant window. This prevents the operator from choosing samples, models, or thresholds after observing which participants benefit.

Release governance also preserves comparability. A new metric or model can begin a new series, but it cannot silently rewrite the conditions under which earlier records and rewards were produced.

7.3 v0: calibration and operational learning

v0 is a deliberate minimum viable mechanism. It establishes the extraction protocol, live miner participation, Bronze comparison, origin-blind Silver adjudication, canonical Silver units, diagnostic scoring, and a competitive payout curve. Its purpose is to measure:

- whether miners can produce valid structured artifacts at batch scale;
- the quality and cost of Bronze and Silver;
- score dispersion and stability;
- common parsing, grounding, and canonicalization failures;
- throughput, latency, exclusion rates, and operational bottlenecks;
- whether external strategies can identify useful improvements over the reference pipeline.

v0 does not establish Gold-anchored accuracy, family-level Sybil resistance, decentralized validation, or canonical production ingestion. It is intended to reveal operational and incentive failures before v1 treats system outputs as training data and database inputs.

The evidence from v0 therefore supports calibration decisions rather than production-quality claims. Its success criterion is whether the live loop yields reliable measurements that can justify or revise the v1 design.

7.4 v1: the initial production architecture

v1 adds the objects needed for production-quality data:

- persistent synthetic and selectively curated Gold;

- hidden post-commit paper and field sampling;
- separate extraction and judging tasks;
- semantic item identifiers and verification-tiered ingestion;
- strategy registration, behavioral-family clustering, and family-level selection;
- marginal-contribution scoring with bounded corroboration value;
- frozen reference releases and Gold-gated promotion;
- random re-audit, replayable audit records, and historical patching.

The architecture specifies which objects must exist and which incentive problem each one addresses. Exact constants remain empirical release parameters.

v1 is therefore a change in assurance, not simply an increase in volume. It adds independent calibration, controlled learning, family-level accounting, and persistent production states to the operational path tested in v0.

7.5 A hypothesis-driven roadmap

Progress depends on falsifiable system propositions rather than calendar milestones alone. Each hypothesis identifies the evidence required before a larger operational commitment becomes justified. Table 3 links each hypothesis to the evidence and decision it supports.

Table 3. Production hypotheses, the evidence needed to evaluate them, and the decisions that the evidence can support.

Hypothesis	Evidence	Decision enabled
Protocol viability	Miners return schema-valid, grounded records with reliable batch completion	Expand workload and participant access
Competitive improvement	Independent miner families produce positive Gold-audited lift over Bronze	Increase reliance on decentralized production
Validator calibration	Seeded-defect recall, clean-control false positives, and confidence calibration meet release thresholds	Admit sampled Silver and Gold decisions into production
Learning transfer	Gold-to-Silver and Silver-to-Bronze gains persist on frozen holdouts without critical regressions	Promote new reference releases
Identity robustness	Simulated and observed identity splits do not increase family assignment or payout beyond stated bounds	Lower entry friction and scale qualification
Economic viability	Cost per trusted link falls while miner break-even and validator operating coverage remain viable	Scale paper volume and Gold investment

Table 3. Production hypotheses, evidence, and enabled decisions—continued.

Hypothesis	Evidence	Decision enabled
Downstream value	Customers and scientific systems reuse records with low correction and reversal rates	Expand domains, schemas, and commercial access

The roadmap ties scaling decisions to evidence from the mechanism itself. Failure of a hypothesis calls for revising the relevant control or stage before expanding the consequences of the failure.

7.6 Research commitments and unresolved conditions

Several quantities remain empirical. The system must measure transport from synthetic to genuine cases, along with the precision and recall of behavioral clustering. It must also estimate the value of a second independent corroborating family, the effects of prize concentration on scores and participation, the minimum Gold budget for reliable release gates, and the cost of reprocessing historical records.

These open conditions define the evidence required for stronger operational claims. CLAIMS is designed as a learning institution: it produces scientific data while also producing evidence about how that data should be created, verified, and financed.

Governance closes the loop by turning those measurements into prospective release decisions. Section 8 returns to the common problem that unites the product, verification architecture, and incentive design.

8 Conclusion

Scientific knowledge remains organized around documents, while many of the systems that use research need structured evidence that can be traced, compared, corrected, and reused. Translating papers into that form is neither a one-time model task nor a benchmark with complete ground truth. It is a continuing production problem in which methods change, evaluation is costly, and the resulting records must survive beyond the round that created them.

CLAIMS addresses that problem by combining a common baseline with decentralized search. Bronze gives every miner the same economical floor. Silver spends stronger machine judgment on discrepancies and strategically selected content. Sparse Gold supplies known-answer and trusted calibration where agreement alone is insufficient. The operational path escalates toward more expensive verification, while the learning path sends verified corrections back toward the cheaper layers used on every paper.

The reward system follows the same informational logic. Canonicalization turns varied wording into stable record items. Behavioral families prevent repeated versions of one information source from receiving repeated selection weight or payout rank. Accepted marginal contribution determines value, with a bounded premium for independent corroboration. These controls limit information-level gains from duplicated identities while preserving qualification, rotation, and a declining reward tail for genuinely distinct methods.

Trust ultimately depends on the evaluators and the release process. Origin-blind adjudication removes candidate identity from the decision. Synthetic corruptions and clean controls measure judge performance. Correlation-adjusted aggregation, replayable audit records, frozen releases, held-out tests, and reversible lineage constrain the propagation of shared errors. v0 tests the live Bronze–Silver loop and its economics; v1 adds the Gold, family, ingestion, and release mechanisms required for production assurance.

The resulting architecture applies beyond scientific extraction. Whenever effort is hidden, complete truth is scarce, evaluation is expensive, and identities or methods can be copied, trustworthy data requires more than a strong model. It requires a market that rewards new information, a verification process that measures its own failures, and a persistent record that can improve without losing its history.

A Formal Model and Design Results

This appendix states the formal objects that support the quality and incentive claims in the main text. It first represents paper content and the accepted record in canonical coordinates, then decomposes final-record error, describes the learning dynamics, and records the conditions under which family-level accounting is invariant to exact duplication.

A.1 Canonical paper content

The formal representation begins with the semantic units that the extraction schema can adjudicate. Let $\mathcal{C}$ be the universe of canonical adjudicable content atoms. For paper p, the extraction target is the finite set

$$X_p \subset_f \mathcal{C},$$

or, equivalently, the sparse binary vector $x_{pc} \in \{0, 1\}$. Each coordinate represents a semantic content atom rather than a literal string. Paraphrases map to the same coordinate only when they preserve meaning, qualifiers, polarity, population, comparator, numerical values, and the evidence relation.

Miner i submits a structured report Y_{ip}. The canonicalization operator $\kappa(Y_{ip})$ maps textual and structured variants into candidate content atoms and equivalence classes. The accepted record F_p is the system's current adjudicated estimate of X_p.

This notation retains the epistemic boundary from the main text. Whether a proposition in the paper is true in the world remains a separate scientific-validity object; the extraction system first seeks an accurate representation of the paper.

A.2 Final-record error decomposition

The final record can improve Bronze by correcting an exposed error, and it can deteriorate Bronze by accepting a false challenge. The following decomposition separates those paths.

For a representative content atom, let B be Bronze, let $Y_1, \ldots, Y_m$ be the miner reports, and let X be the canonical paper label. Define a dispute by

$$D = \mathbf{1}\{\exists i : Y_i \neq B\}.$$

Let

$$b = \Pr(B \neq X), \qquad d = \Pr(D = 1 \mid B \neq X), \qquad f = \Pr(D = 1 \mid B = X),$$

and let

$$q_1 = \Pr(F \neq X \mid D = 1, B \neq X), \qquad q_0 = \Pr(F \neq X \mid D = 1, B = X).$$

Proposition 1 (Final-record error)**.** *The final-record error is*

$$e_F = b(1 - d) + bdq_1 + (1 - b)fq_0.$$

The curated record improves on Bronze if and only if

$$bd(1-q_1) > (1-b)fq_0.$$

The expression partitions outcomes according to whether Bronze is wrong and whether miners create a dispute. The left side of the inequality is successful correction of exposed Bronze errors. The right side is damage to correct Bronze content caused by false disputes and bad adjudication. Verified correction must exceed that damage for curation to improve the baseline.

A.3 Correction and poisoning dynamics

The previous proposition describes one record. Repeated releases create a dynamic process because accepted corrections can enter future training.

Let c_t be the probability that an exposed Bronze error is correctly resolved and transmitted into the next reference release. Let p_t be the probability that a false challenge is accepted and learned, and let ν_t denote drift or newly introduced error. Bronze error then evolves according to

$$b_{t+1} = b_t(1 - d_t c_t) + (1 - b_t) f_t p_t + \nu_t.$$

With constant parameters, $\nu_t = 0$, and $0 < dc + fp < 2$, the sequence converges to

$$b_\infty = \frac{fp}{dc + fp}.$$

The limiting error rate depends on the balance between correction and poisoning. The robust-design objective is $dc \gg fp$: exposed errors should be corrected and transmitted much more often than false corrections survive adjudication and enter training.

This dynamic also explains why release discipline is part of the mechanism. Even a small false-correction channel can create a persistent error floor when accepted labels are reused.

A.4 Gold-mediated transmission

Gold affects Bronze only when a sequence of intermediate events succeeds. For error class k, let $d_{k,t}$ denote Bronze-error discovery, $\rho_{k,t}$ Gold escalation, $g_{k,t}$ Gold error, $\tau^{GS}_{k,t}$ Gold-to-Silver transfer, and $\tau^{SB}_{k,t}$ Silver-to-Bronze transfer. Define the Gold-mediated correction coefficient as

$$\chi^G_{k,t} = d_{k,t}\rho_{k,t}(1 - g_{k,t})\tau^{GS}_{k,t}\tau^{SB}_{k,t}.$$

Proposition 2 (Conditional geometric improvement)**.** *If $\chi^G_{k,t} \geq \underline{\chi}_k > 0$, no new error is introduced in class k, and corrected errors do not reappear, then*

$$b_{k,t} \leq b_{k,0}(1 - \underline{\chi}_k)^t.$$

The bound follows from removing at least a fixed fraction of the remaining class-k error in every period. The mechanism can influence discovery, escalation, and label quality. The transfer coefficients are empirical machine-learning quantities and must be measured on frozen holdouts.

The proposition is therefore conditional rather than a promise that learning will compound automatically. It identifies the channels that must remain positive and the regression conditions that release testing must check.

A.5 Contest intensity and convergence speed

A simple contest benchmark illustrates how reward concentration can affect discovery while holding participation fixed. Suppose a reward pool R combines equal sharing and a performance contest:

$$\mathbb{E}[R_i] = (1-\alpha)\frac{R}{n} + \alpha R \frac{q(e_i,\theta_i)}{\sum_j q(e_j,\theta_j)}, \qquad \alpha \in [0,1].$$

Under homogeneous miners, $q(e) = e$, and quadratic effort cost $ce^2/2$, symmetric aggregate effort is

$$E^*(\alpha) = \sqrt{\frac{\alpha R(n-1)}{c}}.$$

If discovery is $d(E) = 1 - e^{-\kappa E}$, stronger concentration raises discovery and accelerates convergence in this fixed-participation benchmark. The result does not establish that winner-take-all is optimal when participation, score noise, risk, method diversity, false challenges, and duplicated compute are endogenous.

The benchmark supplies a direction for one margin while leaving the empirical controls in the main mechanism open. Qualification, rotation, the reward tail, and family accounting address margins that the fixed-participation calculation omits.

A.6 Detected-family duplicate invariance

Family-level accounting is intended to prevent exact duplication from multiplying information weight. Suppose redundant strategies are assigned to one behavioral family. The family receives the maximum member selection weight rather than the sum, occupies one payout rank, and receives item value at the family level.

Proposition 3 (Duplicate invariance)**.** *Conditional on correct family assignment and fixed scoring context, adding an exact duplicate strategy cannot increase total family assignment weight, accepted-item credit, or payout.*

The result follows directly from the three family-level maxima and totals: the duplicate creates no new family, item, or rank. It is an information-family guarantee rather than proof of unique ownership. It depends on correct clustering and on a scoring context that the family cannot change through its own copies.

The conditional language is essential. Cluster evasion and context manipulation lie outside the proposition and require the additional defenses described in the main text.

A.7 Bounded corroboration loss and the full split-Sybil benchmark

For accepted item h, let k_h be the number of distinct supplying families and define

$$V_h = 1 + \beta \mathbf{1}\{k_h \geq 2\}.$$

Even when an apparent duplicate evades clustering, arbitrary additional families cannot raise total item value above $1+\beta$. False independence can therefore inflate item value by at most the

corroboration premium. The formula does not remove assignment or rank advantages that arise elsewhere in the mechanism.

A full split-Sybil guarantee requires a nonduplicable or conserved resource Ω whose exposure can be partitioned without being multiplied, linked identities, proper scoring in a resource-invariant context, and no positive fixed per-identity wedge after identity cost. Behavioral-family accounting approximates the informational part of this condition.

Stronger formal protection may require stake, bonded task entitlement, coldkey-level assignment limits, or another enforceable resource linkage. The bounded item-value result and the full economic benchmark should therefore be reported separately.

B Illustrative v1 Calibration

This appendix gathers one concrete set of v1 controls so that the proposed architecture can be evaluated as an operational system. The values are release parameters under study rather than constitutional constants. Future releases may change them through the governance process described in the main text.

B.1 Behavioral similarity

Behavioral-family assignment begins with output similarity on common tasks. For strategies i and j, let J_{ij} measure accepted-item overlap, let E_{ij} measure overlap in evidence and structured values, and let R_{ij} measure shared false positives, omissions, and synthetic-challenge errors. Define

$$O_{ij} = 0.7J_{ij} + 0.3E_{ij}.$$

A temporary family can be formed by complete-link clustering when $O_{ij} \geq 0.95$. A persistent link requires at least three shared tasks and either high average output overlap or moderate output overlap together with high shared-error overlap and a shared nontrivial error.

The additional error condition reduces reliance on correct-answer overlap alone. Shared unusual errors are more diagnostic of a common information source than agreement on content that most capable systems recover.

These thresholds illustrate how family evidence can accumulate over time. They remain subject to false-merge and false-split measurement before they support stronger identity-robustness claims.

B.2 Qualification, performance, and rotation

An illustrative ten-slot assignment allocates four slots to qualification, four to performance, and two to rotation. Strategies with fewer than three evaluations receive qualification priority.

For vetted strategy i, the performance score is

$$b_i = 0.25\widehat{Q}_i + 0.75\widehat{M}_i,$$

and its sampling weight is

$$(0.05 + b_i)^2.$$

At the family level, the selection weight is the maximum eligible member weight. Families are sampled without replacement, and one strategy is selected from each chosen family.

A rotation index can combine time since evaluation, uncertainty caused by limited observations, and a novelty bonus for strategies that have not yet entered a persistent family. The three lanes thus preserve entry and exploration while preventing redundant family members from adding selection weight.

The slot counts, score coefficients, and sampling transformation should be judged together. Changing one of them can alter both access and the expected return to quality.

B.3 Marginal contribution

For accepted item h, the illustrative total value is

$$V_h = 1 + 0.20\mathbf{1}\{k_h \geq 2\}.$$

The first family creates discovery value. A second independent family adds a 20 percent corroboration premium. Third and later families do not increase total item value. Each supplying family receives V_h/k_h, and multiple selected strategies within one family divide rather than multiply that credit.

For strategy i, marginal contribution is

$$m_i = \frac{\text{item credit of } i}{\sum_h V_h}.$$

Rolling quality and marginal contribution update through exponential smoothing. At payout, family c receives

$$M_c = \sum_{i \in c} m_i, \qquad Q_c = \max_{i \in c} q_i, \qquad S_c = 0.25Q_c + 0.75M_c.$$

The calibration gives more weight to accepted marginal information than to standalone quality, while retaining a quality component that rewards reliable complete work. The 20 percent corroboration premium and the 25/75 quality–marginal weights should be calibrated against Gold-audited error reduction, cluster-evasion cost, miner participation, and reward stability.

B.4 Judging controls

An illustrative v1 judging packet uses five voters, reliability weighting, exclusion of the evaluated miner and its family, an individual weight cap, a correlated-family cap, a minimum effective-panel requirement, and a confidence threshold below which the case remains unresolved.

Known-answer synthetic items and clean controls estimate accuracy by error class, false-positive behavior, confidence calibration, and abstention quality. The same measurements determine whether a nominal five-voter panel supplies enough independent evidence for a decision.

Panel size and thresholds remain release parameters. Their role is stable: they limit dominance, discount correlation, preserve an unresolved state, and make judge performance observable before genuine-case consensus receives a high verification status.

C Metric Definitions and Reporting Contract

This appendix defines the formulas behind the public status vector. Each measure identifies a distinct channel of product quality, learning, competition, or validator assurance. Reported values must also carry the release, sample, denominator, selection rule, confidence interval, and v0 or v1 status needed for interpretation.

C.1 Gold-anchored product quality

For claim units, report precision and recall on Gold. For claim–evidence links, report an F-score that includes relation correctness, source grounding, and required qualifiers. A conservative product-quality floor is the smallest lower confidence bound across these non-substitutable components:

$$Q^G_{\text{floor}} = \min\left\{P^{G,\text{LCB}}_{\text{claim}}, R^{G,\text{LCB}}_{\text{claim}}, F^{G,\text{LCB}}_{\text{edge}}\right\}.$$

The accompanying report states the number of audited papers and items, the error-class composition, Bronze, Silver, and final-record results on the same sample, and the provenance of Gold labels.

The minimum protects against substitution across components. The component estimates remain visible so that a low floor can be traced to claim precision, claim recall, or evidence-link quality.

C.2 Net lift over Bronze

At release t, let B_t denote Bronze, let F_t denote the final record, and let X denote the canonical paper label. Net lift measures correction after subtracting the damage caused by false challenge and adjudication:

$$L_t = \Pr(B_t \neq X, F_t = X) - \Pr(B_t = X, F_t \neq X).$$

The correction and damage components are reported separately. Release learning velocity is evaluated on a frozen holdout rather than on training or production cases.

This separation reveals whether an unchanged net value comes from low activity or from offsetting high correction and high damage. It also aligns the operational metric with the final-record error condition.

C.3 Verified data velocity and cost

The preferred output statistic is the number of new audited or Gold claim–evidence links produced per day. Corroborated but unaudited records are reported separately.

Unit cost divides Bronze, Silver, Gold, audit, and canonicalization expenditure by verified output. Papers per day, raw claim units, latency, exclusion rate, and reversal rate remain companion diagnostics.

The verified denominator prevents raw throughput from being presented as trusted production. The companion measures show where the pipeline gains or loses scale.

C.4 Effective competition and Sybil split gain

Let s_g denote family g's share of unique accepted contribution or reward. The effective number of miner families is

$$N_{\text{eff}}^{M} = \frac{1}{\sum_g s_g^2}.$$

A direct split test compares total reward under a k-identity representation with the consolidated-family counterfactual:

$$SGR_k = \frac{\text{reward under a } k\text{-identity split}}{\text{reward of the consolidated family}}.$$

For detected redundant strategies, the target is $SGR_k \leq 1$. Any bounded residual advantage from cluster evasion is reported explicitly.

Effective family count measures concentration in observed contribution or reward. The split test evaluates the mechanism counterfactually. Together they distinguish a diverse market from a large collection of redundant accounts.

C.5 Validator assurance

For seeded defects and clean controls, report

$$R_{\text{seed}}^{V} = \Pr(\text{defect detected} \mid \text{seeded defect}), \qquad FP_{\text{clean}}^{V} = \Pr(\text{defect reported} \mid \text{clean control}).$$

Both measures are stratified by omission, hallucination, numerical error, qualifier loss, attribution, causal escalation, evidence mismatch, semantic duplication, and split–merge errors. Agreement among validators or model families is reported alongside Gold accuracy and correlation-adjusted effective count.

The pair of measures keeps sensitivity and false alarms visible. Stratification shows whether an aggregate score conceals a failure in a capability class that matters for genuine records.

C.6 Publication standard

Every headline metric identifies the release, sample, denominator, selection method, confidence interval, and its status as either a v0 proxy or a v1 production measure. Metric definitions remain fixed for each reporting release.

A historical series is never silently recomputed under a new Gold set, denominator, or audit distribution. When a definition changes, the new series begins with an explicit crosswalk and changelog.

These requirements make the public dashboard part of the reproducible mechanism. They also preserve the distinction between calibration evidence and production assurance across versions.

D v0–v1 Crosswalk

Table 4 records how each live v0 object maps to the intended v1 production architecture. It prevents calibration terminology from being read as a claim that the corresponding production control is already active.

Table 4. Mechanism objects in the live v0 calibration system and their intended v1 production form.

Mechanism object	v0 calibration system	v1 production direction
Bronze	Reference extraction for scoring papers	Frozen, versioned reference release promoted through Gold holdouts
Silver	Origin-blind AI adjudication and canonical Silver units	Gold-calibrated machine adjudication with confidence-aware escalation
Gold	Not an active production layer	Synthetic known-answer tasks, curated subsets, selective trusted review
Miner work	Structured extraction	Extraction plus separate blinded judging
Miner selection	Qualification, performance, and rotation at account level	Strategy- and family-level selection with exploration and redundancy controls
Reward	Competitive rank curve using v0 quality scores	Almost-winner-takes-all family ranking and marginal contribution
Identity treatment	UID and hotkey observations; exact-copy diagnostics	Declared strategies, behavioral clusters, one family selection weight and rank
Paper validation	Full live Silver path for returned scoring papers	Hidden post-commit sampling, confidence bounds, and random re-audit
Database ingestion	Experimental artifacts; no automatic canonical production entry	Verification-tiered record ingestion with lineage and patching
Validator topology	One owner-operated validator	Owner reference validator plus independent replay and later substantive validators
Public evidence	Stored operational and score artifacts	Signed replayable audit records and Gold-anchored release reports

The v1 direction keeps the operational lessons of v0 while adding independent truth, controlled sampling, family-level incentives, production ingestion, and stronger public assurance. The crosswalk should be updated with each release so that readers can distinguish implemented controls from planned architecture.

References

[1] Y. Rao, J. Steeves, A. Shaabana, D. Attevelt, and M. McAteer. *Bittensor: A Peer-to-Peer Intelligence Market*. arXiv:2003.03917, version 3, 2021. https://arxiv.org/abs/2003.03917; https://www.bittensor.com/whitepaper.

[2] Minos Team. *Minos: Decentralized Infrastructure for DNA Mutation Inference & Benchmarking*. White paper, version 1.1, February 2026. https://theminos.ai/whitepaper.

[3] C. Roessler. *Actionable Information and the Tinbergen Rule*. Working paper, 2026.

[4] C. Roessler. *Designing What to Reward*. Working paper, 2026.

[5] C. Roessler. *Sybil-Resistant Scoring under Anonymity*. Working paper, 2026.

[6] J. Priem, H. Piwowar, and R. Orr. *OpenAlex: A Fully-Open Index of Scholarly Works, Authors, Venues, Institutions, and Concepts*. arXiv:2205.01833, 2022. https://arxiv.org/abs/2205.01833.

[7] U.S. National Library of Medicine. PubMed. Online database, https://pubmed.ncbi.nlm.nih.gov. Accessed August 28, 2026.

[8] Crossref. Metadata Retrieval. Online service, https://www.crossref.org/services/metadata-retrieval/. Accessed August 28, 2026.

[9] OurResearch. Unpaywall API. Online service, https://unpaywall.org/products/api. Accessed August 28, 2026.