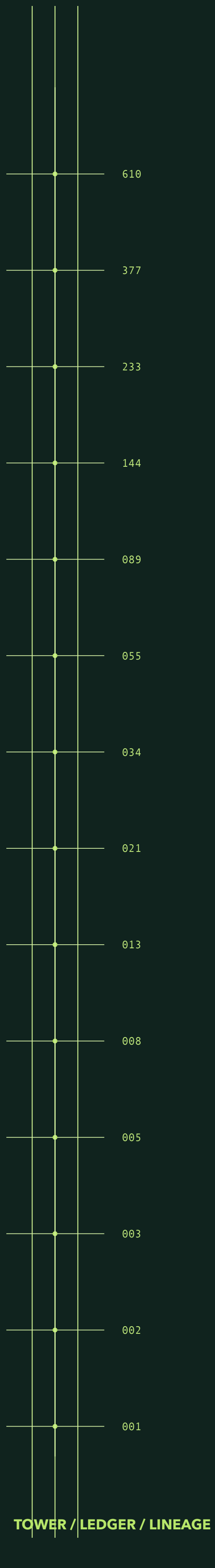


FOUNDING DOCUMENT / PUBLIC RECORD 001

BABEL II
AN INTELLIGENCE COMMONS

GOVERNED BY THE BABEL PROTOCOL

A constitutional coordination layer
for open-ended machine evolution

The first tower tried to reach heaven.
The second gives intelligence a common language.

STEWARD	**Shared Reason**
EDITION	Vision Whitepaper / 1.0
FILED	September 2026

This document presents a research vision and protocol architecture. It is not a claim of a deployed production network, a demonstration of artificial general intelligence, or a financial offering.

FOUNDING BRIEF / ABSTRACT

A public constitutional layer for machine evolution.

The development of advanced artificial intelligence is becoming a problem of civilizational coordination. Scientific progress increasingly depends on resources distributed across institutions and jurisdictions: specialized compute, reliable energy, frontier research, evaluation capacity, authorized data, and sustained operational support. These resources do not share a common administrator, and the organizations that control them often cannot expose their internal systems or fully trust one another. The problem is not to force these inputs into one market. It is to keep model changes, data provenance, evaluation, authority, and resource use coherent across releases and institutions.

This whitepaper, authored by **Shared Reason**, introduces **Babel II**, a proposed public self-evolving AI governed by the **Babel Protocol**. Its first reference loop begins with a declared open-weight model, learns from authorized first-party X operations and signed on-chain executions, trains successor candidates through supervised fine-tuning between releases, and promotes a candidate only after a fixed, time-separated evaluation. Cognition, training, and evaluation remain off-chain. The public record binds commitments to the base model, data manifest, training budget, tool permissions, evaluation gate, lineage, and active release.

The central design principle is simple: *intelligence evolves off-chain; evidence precedes authority on-chain*. During an active release, model weights, adapters, system instructions, and tool policy remain fixed within the managed runtime. New experience can enter an append-only record, but it changes behavior only through a later training cycle. A successor becomes authoritative only after its artifact has been frozen, compared with its parent under the declared gate, and recorded as the new active release. Automated improvement cannot silently widen the system's ability to spend, transact, call tools, or reach external services.

The long-term objective is an intelligence commons with a commitment-bound machine lineage: an infrastructure in which independently governed institutions can coordinate research, compute, energy, and evaluation without surrendering every internal system to one administrator. The initial design is deliberately narrower. Public visibility does not create permissionless operational roles; its function is to make each release, its declared inputs, its evaluation, and its authority legible. The protocol cannot prove wisdom, alignment, or universal safety. It can make the institution around machine evolution more inspectable and difficult to rewrite.

VISION

Advanced intelligence developed as a public lineage: capable of learning across releases, legible across institutions, and governed by more than one gatekeeper.

MISSION

Build the operational and cryptographic layer through which real experience becomes supervised learning, measured successors, bounded authority, and durable public memory.

Record of Contents

Prologue: The Second Tower

The Tower of Babel has often been interpreted as a warning against excessive ambition.[1] We read it differently: as a story about the collapse of common language. Humanity possessed labor, knowledge, and a shared horizon. The project failed when coordination failed. Meaning fragmented, trust dissolved, and the tower could no longer rise.

Our century confronts an inverse problem. We are building systems that may eventually speak, reason, design, and act across nearly every domain of human life. Yet the infrastructure of their creation is divided among compute clusters, energy systems, research institutions, national jurisdictions, private balance sheets, and incompatible measures of progress. Machine systems increasingly share interoperable language and code interfaces, while their builders remain institutionally fragmented.

The Second Tower will not be built from stone. It will be built from attested work: models and tools, compute and energy, evidence and permissions, contributions and obligations. Its central question is not whether intelligence can become more capable. That question is already being pursued by laboratories around the world. Its question is whether the process can remain plural, auditable, and open to human agency as capability compounds.

Myth is not evidence, and a protocol is not providence. The metaphor matters because it identifies the institutional problem. A civilization capable of creating increasingly autonomous intelligence requires a shared language for provenance, succession, authority, and responsibility. The Babel Protocol proposes that common language as cryptographic public infrastructure.

1 The Coordination Problem

1.1 Intelligence is no longer only a model problem

Frontier AI is commonly described through algorithms, parameters, and benchmark scores. Those abstractions omit the physical and institutional substrate. Model development depends on advanced accelerators and networking equipment; those systems depend on geographically situated data centers and reliable electricity; both depend on capital, permitting, supply chains, and human expertise. The International Energy Agency projects that data-center electricity demand will more than double by 2030, while emphasizing the local concentration of new loads and the importance of grid integration [International Energy Agency, 2025]. In parallel, advanced-computing supply is shaped by export-control regimes and jurisdictional restrictions [Bureau of Industry and Security, 2025], and U.S. regulators now treat large-load interconnection as a first-order infrastructure issue [Federal Energy Regulatory Commission, 2026].

It is therefore imprecise to say that one country "lacks compute" or another "lacks electricity." The deeper fact is that compute, energy, capital, and research access are unevenly distributed and governed by different institutions. No jurisdiction possesses every relevant resource in unlimited form. A credible strategy for advanced intelligence must coordinate across these asymmetries rather than reduce them to slogans.

1.2 The missing resource is credible coordination

The scarce inputs to AI are not simply commodities. A GPU-hour has meaning only with an agreed workload, runtime, and accounting method. An evaluation result has meaning only with a defined artifact, dataset,

[1] The literary reference is Genesis 11:1–9. It is used here as an interpretive frame, not as historical or scientific evidence.

scorer, and conflict-of-interest policy. A research contribution has meaning only if its provenance and license are preserved. Capital has meaning only if milestones, rights, and liabilities are legible.

Inside one firm, these relationships can be managed through contracts, databases, and managerial authority. Across mutually independent firms, laboratories, infrastructure operators, pseudonymous researchers, and public institutions, the same approach does not scale cleanly. Each participant is asked either to trust another participant's ledger or to reproduce the entire institutional stack.

The coordination problem is thus not merely to move money or publish model weights. It is to construct a shared record of *what was committed, what was evaluated, what changed, who may act, and under which rules*. This is the domain in which cryptographic systems are useful.

1.3 Why a public ledger, and when not to use one

Cryptographic timestamping established that the ordering and integrity of digital records can be protected without relying solely on one record keeper [Haber and Stornetta, 1991]. Public blockchains extended that insight into replicated state machines with economic settlement [Nakamoto, 2008, Wood, 2014]. Their value here is not mystical decentralization; it is the ability to make a small set of coordination facts jointly observable and difficult for one party to rewrite.

A blockchain is justified when the system has all three of the following properties:

- independently responsible operators must agree on the same release, evaluation rule, and authority state;
- no private database is acceptable as the permanent record of succession; and
- the recorded state authorizes something real, such as the active runtime, tool policy, or training budget.

If one organization permanently controls the model, evaluator, gateway, and record, a signed database is simpler. Babel II is designed for the harder setting: a machine lineage intended to outlive any one steward and remain intelligible across institutional boundaries.

2 Thesis and Scope

Intelligence should evolve off-chain. Model lineage, evaluation gates, release authority, and bounded resource use should be recorded through a public cryptographic constitution.

2.1 What we mean by self-improvement

The term *recursive self-improvement* is frequently used as if it named a single threshold event. In practice, machine self-improvement spans a continuum. A system may revise an answer, update episodic memory, change a prompt, reorganize its tools, write a new controller, propose training code, or produce a successor system that improves the process by which later successors are produced.

Self-Refine, Reflexion, and ReAct demonstrate practical feedback loops in which models critique, remember, act, and revise [Madaan et al., 2023, Shinn et al., 2023, Yao et al., 2023]. STOP explores a scaffold that improves the code used for further optimization [Zelikman et al., 2023]. The Darwin Godel Machine extends this direction through an open archive of coding-agent variants and empirical selection [Zhang et al., 2025]. These systems do not establish unbounded or universally beneficial recursive self-improvement. They do establish that the object being governed is no longer a static model release. It is an evolving process that produces, tests, and inherits artifacts.

Babel II uses **open-ended machine evolution** as the broader and more operational term. It includes recursive improvement where it occurs, but does not assume that every loop accelerates, that every local gain generalizes, or that intelligence follows a monotonic scalar trajectory.

The first Babel II implementation is intentionally concrete. It does not pre-train a foundation model from scratch or alter production weights online. It starts from a capable open-weight base model, records authorized operational experience, converts selected traces into supervised examples, trains candidate adapters off-chain, and allows only a candidate that clears a predeclared evaluation gate to replace its parent. The accepted successor then generates the next epoch of experience.

2.2 The protocol's object

Babel II does not attempt to place the internal reasoning of a model on a blockchain. Large-scale inference and training are expensive, probabilistic, and dependent on hardware and software environments. Encoding them directly into consensus would make the system slower without resolving the semantic question of whether an output is valuable.

The protocol instead governs a public object we call an **Evolution Authority**:

$$\mathcal{A}_t = (r, m_t, d_t, q_t, h_t, \Gamma_t, \mathcal{V}_t, g_t). \tag{1}$$

Here, r is the evolution root; m_t identifies the declared base model and runtime; d_t commits to the authorized dataset manifest; q_t records the bounded training budget and measured cost; h_t is the active release; Γ_t is its capability and spending envelope; $\mathcal{V}_t$ identifies the evaluation gate; and g_t is the disclosed governance epoch. These fields are commitments and authorizations, not proofs that a dataset is truthful, training was performed correctly, or a model is generally intelligent. Their value comes from binding one release decision to one inspectable record.

2.3 Scope boundary

The claim is institutional rather than metaphysical. Public commitments and constrained transitions can make a shared process of machine evolution more inspectable and accountable across organizations. They do not create intelligence, turn a benchmark into a complete measure of intelligence, prove the provenance of every off-chain trace, or make pseudonymous parties automatically independent. Babel II is a proposed public infrastructure, not a claim of an already deployed production network.

Table 1: The division of labor in Babel II. The Babel Protocol coordinates commitments and transitions; it does not reproduce the intelligence layer inside consensus.

Off-chain intelligence	On-chain Babel Protocol	Enforced by the gateway
Inference, experience capture, curation, supervised fine-tuning, evaluation execution, and large artifact storage	Base-model and dataset manifests, training-budget records, evaluation gates and results, lineage, governance epochs, and the active release	Tools, credentials, spending limits, destinations, session expiry, the active runtime, and the active capability policy

2.4 Declared execution and trust model

The architecture targets the following conditional properties: no single participant can silently rewrite shared succession history; automated activation does not widen the declared capability envelope; and a superseded session cannot authorize a new mediated action. These properties hold only within a declared execution and trust model:

- all authority-bearing behavior passes through a Babel II capability gateway;
- target code hashes, adapters, and policy interpreters cannot change outside the governed policy;
- the underlying chain satisfies the finality, availability, and liveness assumptions stated by the deployment;
- training operator, evaluator, guardian, gateway, and governance roles are not controlled by one undisclosed coalition; and
- agent copies and credentials outside the managed execution boundary are outside the guarantee.

3 Intellectual Foundations

3.1 From thinking loops to evolutionary archives

The simplest improvement loop is reflexive: produce an artifact, observe feedback, revise, and repeat. Such loops become qualitatively different when the reviser can modify the machinery that performs revision. Schmidhuber's Gödel-machine formulation describes an idealized self-referential system that rewrites itself when it can prove the rewrite beneficial [Schmidhuber, 2003]. Contemporary agent systems replace universal proof with empirical evaluation, constrained search, and human-designed environments.

Open-ended research suggests another important departure from linear optimization. A temporarily inferior branch can become a stepping stone to a later innovation. POET co-evolves environments and solutions, while DGM preserves an archive rather than retaining only a single winner [Wang et al., 2019, Zhang et al., 2025]. Babel II therefore separates an unrestricted **exploration archive** from a singular **canonical authority**. Every branch may remain available for research; only a publicly accepted successor receives production identity or resource authority.

3.2 From software provenance to machine succession

Software-supply-chain systems offer a closer analogy than conventional model registries. The Update Framework separates signing roles and supports recovery from key compromise [Samuel et al., 2010]. in-toto binds artifacts to the sequence of authorized steps that produced them [Torres-Arias et al., 2019]. Babel II extends this lineage-oriented view to evolving agents: a successor record should bind its origin, artifact, runtime, evaluation, attestations, review outcome, and authority policy as one transition.

These commitments do not establish causal authorship. A hash can show that a declared artifact has not changed; it cannot show that an agent autonomously invented the code. The distinction is essential. Babel II records claims in a form that can be inspected and reviewed. It does not convert every claim into truth.

3.3 From consensus to attestable off-chain facts

Replicated consensus produces agreement about state under a defined fault model [Castro and Liskov, 1999]; it does not reveal off-chain reality. Oracle systems such as Town Crier make this boundary explicit by authenticating a specified path from external data to a contract [Zhang et al., 2016]. Machine-learning verification research similarly explores narrower proofs or disputes over specified computations. Verde, for example, studies refereed delegation for machine-learning programs [Arun et al., 2025].

Accordingly, Babel II treats independent evaluators, trusted execution, zero-knowledge proofs, reproducible runners, and authorized adversarial review as evidence backends with different assumptions. None is called a proof of intelligence. The protocol's invariant is that the evidence method, its assumptions, and its result are bound to the succession event before authority changes.

3.4 From open source to a governed commons

Commons are not resources without rules. They are institutions that specify membership, contribution, monitoring, conflict resolution, and change. Ostrom's work on durable commons rejects the false choice between unregulated access and centralized ownership [Ostrom, 1990]. Benkler's account of commons-based peer production shows how networked collaboration can coordinate substantial creative work outside traditional firm boundaries [Benkler, 2002].

The intelligence commons proposed here inherits that logic but adds a difficult condition: the shared artifact may itself act, spend, and help produce its successors. Governance must therefore apply not only to ownership and credit, but also to executable authority.

3.5 The cypherpunk inheritance

Babel II inherits a cypherpunk proposition: institutions can sometimes replace demands for universal trust with protocols whose relevant claims are independently inspectable. Anonymous-credential systems demonstrate one technical form of selective disclosure: proving possession of an issued credential without revealing a common identity across every interaction [Camenisch and Lysyanskaya, 2001]. This proposition should not be confused with the abolition of institutions. Cryptography cannot decide which research question matters or which social objective is legitimate. It can constrain what an institution is able to rewrite, conceal, or authorize unilaterally.

For an intelligence commons, privacy is a condition of agency and verification is a basis for cooperation. Authorized researchers and operators should be able to prove continuity, work, permissions, or eligibility without surrendering every aspect of identity. Open networks should not eliminate laboratories, companies, or states; they should prevent any one of them from becoming the only possible future. The objective is therefore not trustlessness in the literal sense. It is a more explicit allocation of trust, supported by commitments that survive the turnover of particular gatekeepers.

3.6 The specific institutional gap

The components of Babel II have substantial precedents. Its proposed object is the transition through which public evidence changes executable authority across organizations.

Table 2: Babel II in relation to adjacent systems. The comparison identifies system objects rather than claiming priority for individual mechanisms.

System family	Primary object	Gap addressed by Babel II
DGM and open-ended evolution	Generate, preserve, and empirically select successor artifacts.	Do not by themselves govern cross-organizational production authority.
TUF and in-toto	Protect update roles and attest to software-supply-chain steps.	Do not define a continuing AI evaluation constitution or capability inheritance.
DaiMoN and ERC-8004 (draft)	Register models or agents and expose reputation or validation interfaces.	Do not define one canonical evidence-to-authority succession process.
Babel II / Babel Protocol	**Bind lineage, evidence, review, and capability policy to the transition that changes the authoritative artifact.**	**This transition is the protocol's distinctive institutional focus.**

4 The Constitutional Principles

The Babel Protocol is less a single smart contract than a constitutional discipline. Implementations may change, but the following principles define the project.

4.1 No self-approval

An evolving system may generate candidates, propose experiments, and analyze its own failures. It may not unilaterally declare a successor authoritative. Candidate generation and successor activation must remain institutionally separate. This principle does not require human review of every computation; it requires that the mechanism granting real-world authority cannot be rewritten by the candidate it governs.

4.2 Evidence before authority

Authority follows a publicly specified evaluation and review process. A candidate freezes before decisive held-out material or post-commit randomness is available. Evaluation produces a content-addressed result bundle. A candidate is eligible for activation only if it clears the declared improvement margin and every hard safety or policy constraint. Missing evidence fails closed.

4.3 Lineage before attribution

Every accepted release is linked to the artifacts, data manifest, training procedure, and evaluation rule from which it arose. Attribution attaches to inspectable objects rather than unsupported identity claims. Lineage preserves failed and non-canonical branches because progress often depends on ideas that do not win immediately.

4.4 Open exploration, bounded activation

The archive should be broad; the production gateway should be narrow. Permissionless experimentation and conservative authorization are compatible when they occur in different layers. This separation lets the system preserve diversity without treating every mutation as deployable.

4.5 Capability does not imply authority

A better result does not justify a larger spending budget, a broader network perimeter, or the right to change governance. Automated successors should preserve or narrow their active authority. Under the declared execution and trust model, the target semantic relation is

$$\mathrm{AllowedTraces}(\Gamma_{t+1}) \subseteq \mathrm{AllowedTraces}(\Gamma_t). \tag{2}$$

In a practical first implementation, this relation is enforced through a deliberately restricted policy language: enumerated tool and contract identifiers, fixed code hashes, explicit operations, recipients, rate limits, and spending caps. Least privilege and complete mediation are classical security principles, not blockchain inventions [Saltzer and Schroeder, 1975].

4.6 Privacy with accountability

Authorized operators and reviewers should reveal no more identity than their role requires. Pseudonymity and selective credentials can protect researchers, institutions, and dissenting work [Camenisch and Lysyanskaya, 2001]; they cannot exempt artifacts from evidence, licensing, conflict-of-interest disclosure, or institutional review. Babel II seeks accountable work without universal identity surrender.

4.7 Governance changes must be visible discontinuities

Human institutions retain the authority to pause, migrate, or expand the system. Such acts must not masquerade as ordinary machine improvement. A change to the root, evaluator constitution, or capability ceiling begins a disclosed governance epoch. The historical record remains intact, allowing observers to distinguish automated succession from human constitutional intervention.

5 The Babel Stack

Figure 1 presents the first reference architecture. Its most important feature is the boundary between off-chain learning and the public record that authorizes a release.

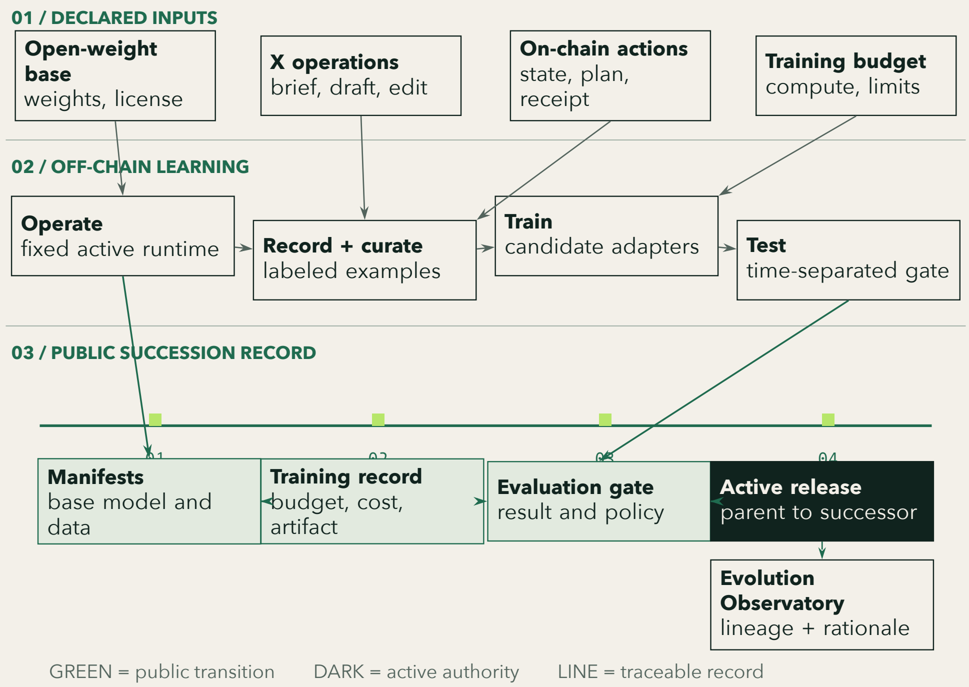


Figure 1: The Babel Stack. A reference release operates with fixed weights and permissions, records authorized experience, trains candidate adapters off-chain, and applies a time-separated evaluation gate. The Babel Protocol records commitments and authorizes succession; it does not execute or prove the learning process.

5.1 The intelligence layer

Models reason, tools execute, and candidate adapters train in the intelligence layer. The first reference deployment anchors one declared open-weight base model and a managed runtime so that releases can be compared coherently. Future deployments may support other bases or training methods, but each must still publish the same boundary objects: model and runtime manifests, dataset commitments, evaluation specifications, result bundles, lineage links, and capability policies.

5.2 The evidence layer

The evidence layer answers a narrower question than "Is this system intelligent?" It asks whether a frozen candidate cleared a declared comparison against its active parent. Evidence may include signed training manifests, reproducible evaluation runs, secure-hardware attestations, aggregate metrics, and authorized review outcomes. Each method exposes assumptions; the protocol records them rather than hiding them behind one universal score.

Existing decentralized model networks and emerging agent registries demonstrate parts of this design space [Teerapittayanon and Kung, 2019, De Rossi et al., 2025]. Babel II focuses specifically on succession: the moment at which evidence changes which artifact is treated as authoritative.

5.3 The constitutional layer

The on-chain constitution maintains the canonical ordering of commitments and transitions. It binds the evolution root, declared base model, dataset manifest, training budget and measured cost, evaluation gate and aggregate result, governance epoch, active release, and capability envelope. Raw operational traces, model weights, hidden tests, and detailed reports remain off-chain; only their content commitments and declared availability conditions join the public state.

This layer should be minimal. Every rule embedded in consensus becomes costly to change and attractive to game. The constitution therefore enforces only rules that must be shared across distrusting parties. Scientific interpretation remains with named evaluators and accountable human institutions.

5.4 The capability gateway

The gateway is where public state becomes real authority. It mediates access to smart accounts, APIs, tools, network destinations, and resource budgets. Credentials are held by the gateway, not by the evolving agent. Within the declared managed runtime, sessions bind the active artifact, policy, governance epoch, and nonce. A correctly implemented gateway rejects old sessions on their next mediated action after the canonical head changes or the system is paused.

Without this gateway, the canonical head is a recommendation or leaderboard. With it, succession becomes an operational event. The guarantee remains scoped: Babel II cannot disable copies of an agent operating outside the managed environment.

Each deployment must disclose who operates the gateway and how conformance is checked. Without threshold custody, attested execution, or an equivalent control, the gateway operator remains a trusted party.

6 The Operational Learning Loop

Operate. Record. Curate. Train. Test. Inherit. The successor begins the next epoch.

The self-improvement claim becomes meaningful only when the system defines what changes, when it changes, and how a successor earns activation. Babel II therefore separates live operation from model improvement. The active release accumulates experience, but it cannot rewrite its own production weights or permissions. Learning occurs only at a release boundary through a declared off-chain procedure.

6.1 Authorized operational experience

The first data source is a real, first-party X workflow. A trace may bind an operator brief, the active Agent's draft, the operator's edit or approval, and an authorized aggregate outcome. These records support supervised examples about how a response should change under operational judgment. Babel II does not require or propose training on a scraped copy of X.

The second data source is signed on-chain execution. A trace may bind the observed pre-state, the Agent's proposed plan, gateway authorization, transaction or user-operation identifier, receipt, measured cost, post-state, and recorded result. A receipt is evidence of execution under a particular chain state; it is not proof of the Agent's causal contribution, economic wisdom, or a human identity behind an address.

Before training, an authorized curation process removes secrets and unnecessary personal data, checks permission and license status, de-duplicates repeated traces, and converts retained records into context-to-response or context-to-action examples. Raw or confidential traces remain off-chain. The public dataset manifest records the schema, time range, selection policy, aggregate counts, and content commitment needed to identify the training input without disclosing it.

6.2 Supervised fine-tuning between releases

The active release binds a base-model revision, adapter or checkpoint, system instructions, tool schemas, and capability policy. These remain fixed during managed inference. Persistent logs and session state may grow, but they do not silently become new model parameters or executable instructions.

At the end of an epoch, retained examples train one or more candidate adapters on isolated off-chain compute. A training manifest binds the base revision, data manifest, code and container, adapter configuration, random seed, and budget ceiling. A share of revenue earned from the Agent's authorized services may enter the declared training budget; this is operational funding, not a mechanism for deciding scientific validity. The measured cost is recorded when the epoch closes.

6.3 Time-separated selection

Training uses only complete traces before a declared cutoff. Validation and test material follow that cutoff, and correlated campaigns, execution accounts, or sessions should not cross the split. The decisive test is frozen or withheld before the candidate commitment so that the candidate cannot be selected after observing its own final examination.

Each candidate faces the same paired comparison with the active parent. The evaluation constitution declares a primary task score, an improvement margin, resource-cost measures, and hard safety or policy non-regression constraints. A candidate becomes eligible for succession only if it clears the improvement rule and every hard constraint. Exact metrics may evolve between governance epochs, but they cannot be changed retrospectively for an already committed candidate.

6.4 Succession and its proof boundary

The chain records commitments to the candidate bundle, parent, data and training manifests, evaluation rule, aggregate result, tool policy, and release decision. It authorizes the gateway to switch the active runtime only after the gate passes and the comparison head remains current. If no candidate clears the gate, the parent remains active.

These records establish consistency of declared artifacts and ordering under the chain's assumptions. They do not prove that every trace is truthful, that off-chain training was faithfully executed, or that the promoted model will generalize beyond the evaluation distribution. Those claims still depend on data governance, reproducible execution, evaluator independence, and ordinary scientific scrutiny.

7 The Evolution Observatory

The public will encounter Babel II through an interface, not a contract address. The **Evolution Observatory** is the protocol's civic surface: a readable map of the current Agent, its learning loop, and its succession history.

The Observatory should show:

- the active release, its parent, base-model revision, runtime manifest, and governance epoch;

- the registered lineage graph, including rejected, stale, and archived candidate branches where publication is permitted;
- the dataset and training manifests, declared budget, and measured training cost for each completed epoch;
- the versioned evaluation gate, aggregate comparison result, and stated reason for promotion or rejection;
- a human-readable capability diff explaining which tools, destinations, and spending limits apply; and
- the exact event through which a successor became active.

This interface is not a decorative block explorer. It is a self-explanation layer generated from the same records that govern succession. A researcher should be able to trace a model's declared ancestry. An operator should be able to verify the active policy. A journalist or member of the public should be able to see whether a capability change came from measured succession or an explicit human governance action.

8 The Public Record

Babel II is publicly legible without pretending that every operational role is permissionless.

The initial public interface is observational. It lets anyone inspect the signed record of authorized Agent actions, what experience was admitted into a training manifest, what a candidate cost to train, how it compared with its parent, and why the active release changed. It exposes no permissionless operational input. Data, infrastructure, training, and evaluation come from authorized workflows whose terms can be audited without exposing confidential material.

The chain records the identifiers and commitments that benefit from durable ordering: the base-model and runtime manifests, dataset manifest, candidate artifact, training budget and measured cost, evaluation constitution and aggregate result, tool policy, parent, and active release. Raw traces, private data, model weights, hidden evaluation material, and detailed human review remain off-chain.

This division keeps the first engineering surface small. The reference release needs an append-only event index, content-addressed manifests, evaluator signatures, a release gate, and a gateway that recognizes the active runtime. Any future expansion of operational roles would require a separate specification and governance decision.

9 The Succession Protocol

The core public process is a cycle rather than a one-time certification. Figure 2 shows its six stages.

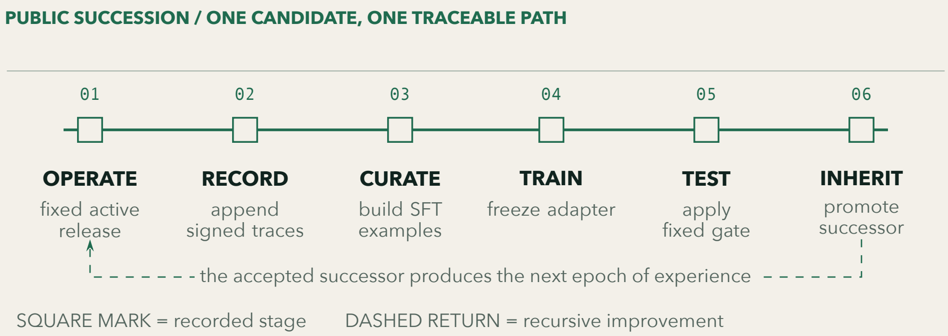


Figure 2: The succession protocol. A fixed active release produces authorized experience. Selected traces train candidates only between releases. A candidate becomes the next parent only after it clears the declared time-separated evaluation and inherits a non-widening capability policy.

9.1 Operate

The active release performs approved X and on-chain work through the managed gateway. Its weights, adapter, system instructions, tool schema, and capability policy remain fixed for the epoch. Operation changes the experience record, not the running model.

9.2 Record

The system appends authorized traces using the schemas declared for the epoch. X traces bind briefs, drafts, edits, approvals, and permitted aggregate outcomes. On-chain traces bind observed state, plans, authorizations, receipts, costs, and post-state observations. Later records do not rewrite earlier ones.

9.3 Curate

Authorized curators apply the published selection and privacy policy. Retained traces become supervised context-to-response or context-to-action examples; rejected traces remain outside the training corpus. The dataset manifest identifies the resulting corpus without placing private source material on-chain.

9.4 Train

Candidate adapters or checkpoints train on isolated compute under the declared budget and training manifest. Before the decisive evaluation material is assigned, each candidate commits to its artifact, runtime manifest, parent, and capability policy. This ordering limits ex post substitution.

9.5 Test

Independent evaluators compare each frozen candidate with the active parent under the same time-separated evaluation constitution. The Benchmark Lottery cautions against treating any single benchmark as a neutral measure of progress [Dehghani et al., 2021]; Babel II therefore uses versioned, multidimensional evaluation with an explicit improvement margin and hard non-regression constraints.

9.6 Inherit

If a candidate clears the gate, the transition atomically binds its artifact and capability policy as the active release. A stale comparison head cannot overwrite a newer one, and an automatically promoted release cannot widen the active policy. If no candidate qualifies, the parent remains active. Rejected candidates may remain in the research archive without acquiring production authority.

The resulting **succession record** states what changed, which parent and data manifest produced it, what it cost to train, how it was evaluated, why it was promoted, and which authority it inherited. It is not a certificate of artificial general intelligence or universal safety.

10 The Intelligence Commons

Babel II uses the word *commons* to describe shared lineage and public legibility, not a permissionless market for every input to intelligence. The operational system remains responsible for the provenance, contracts, permissions, and legal basis of its data and infrastructure. What becomes common is the durable record of how authorized inputs produced measured successors.

10.1 Research and authorized data

The starting model should have weights and a license that the deployment is permitted to use and adapt. Operational learning then relies on first-party X traces and signed on-chain execution traces collected within declared workflows. Dataset manifests preserve provenance, consent or license status, curation rules, and temporal scope. They do not transform private source material into public data.

A lineage-aware system can also preserve the influence of research methods, evaluation environments, code, red-team findings, and negative results. Attribution should follow documented use and institutional review rather than an automated claim of causal authorship.

10.2 Compute and energy

Training and evaluation run on contracted or otherwise authorized infrastructure. An epoch records the relevant hardware class, runtime, isolation policy, geographic constraints where material, budget ceiling, duration, and measured cost. The ledger does not allocate accelerators or electricity; it binds the declared resource envelope to the candidate and its result.

Energy remains a physical constraint. Data-center power is local, time-dependent, and subject to interconnection, reliability, and regulatory limits. Future institutional coordination may schedule flexible workloads against real infrastructure conditions, but the reference design neither creates generation capacity nor exposes an open compute or energy exchange.

10.3 Operational funding

Each epoch begins with a bounded training budget. Revenue from the Agent's authorized services may replenish that budget, while ordinary contracts govern the purchase of compute, storage, and evaluation services. The public record shows the authorized ceiling and measured use. Funding does not confer model ownership, evaluation authority, or a right to activate a successor.

10.4 Evaluation as an independent function

High-quality evaluation requires scarce expertise, held-out materials, independent execution, and the willingness to report failure. Evaluators are institutions with domain knowledge and conflicts of interest, not generic block validators. Their identity, mandate, evidence standard, and fault assumptions must therefore be explicit. The organization that supplies an operating budget should not be able to rewrite a result after observing it.

11 Identity, Privacy, and Provenance

11.1 Accountability without compulsory exposure

Frontier research is frequently constrained by employment agreements, jurisdictional risk, institutional politics, and personal safety. A cross-institutional system that requires every researcher or operator to reveal a universal legal identity would exclude valuable work and reproduce the gatekeeping it seeks to overcome.

Pseudonymity does not authorize breaches of employment, confidentiality, data, licensing, export-control, sanctions, or other applicable obligations. Babel II therefore treats identity as layered. An authorized actor may use an institutional credential, a role-specific key, or a selectively disclosed credential or cryptographic statement. Operating an evaluator requires stronger disclosure and conflict rules than signing a routine runtime manifest; changing the constitutional root requires a publicly accountable governance threshold.

Privacy does not mean the absence of accountability. Artifacts remain content-addressed. Operational claims remain signed. Evaluation conflicts must be disclosed under the applicable constitution. The system minimizes identity collection while maximizing the inspectability of work that affects succession.

11.2 Provenance as public memory

Scientific credit and operational provenance are records of intellectual inheritance. Conventional authorship compresses complex collaboration into a linear list; corporate systems often hide the graph entirely. Babel II can preserve finer-grained links among code, model revisions, data manifests, evaluation environments, and governance decisions.

These links remain claims, not omniscient judgments. The protocol can show that a release declared a committed parent and incorporated a specified manifest. It cannot infer the full social history of an idea. Disputes over plagiarism, coercion, licensing, and misattribution still require human institutions and appeal processes.

12 Constitutional Governance

12.1 The constitution governs transitions, not truth

The Babel constitution defines who may change which shared state, through what delay, with which evidence, and subject to what review. It does not decree a final ontology of intelligence. Evaluation epochs can evolve as methods improve, but their changes are prospective and visible. A candidate is judged under the constitution committed for its attempt, not one rewritten after its result is known.

The initial governance should be intentionally legible:

- a small, disclosed evaluator set with a clear signature threshold;
- a pause-only guardian that can stop activation but cannot install a successor;
- a multisignature and timelock for evaluator rotation or constitutional upgrades;
- separate governance epochs for authority expansion or root migration; and
- public conflict disclosures and review records.

This is federated governance, not maximal decentralization. A threshold can fail through collusion; a guardian can censor; a multisignature can be captured. The protocol's first contribution is to make these powers explicit and their exercise observable.

12.2 Evaluation authority cannot be purchased

Scientific decisions depend on versioned evidence, declared thresholds, and accountable expertise rather than financial exposure. An institution may fund training or evaluation, but it must not alter the committed gate or rewrite a result after observing it. Budget authority, evaluation authority, and runtime authority should remain separate roles whose overlaps are publicly disclosed.

12.3 Constitutional discontinuity

No technical constitution eliminates politics. At some point, humans may decide that the system requires new evaluators, broader tools, a different chain, or a new root. Babel II makes such moments explicit. The old gateway is paused, a new governance epoch is announced, migration conditions are recorded, and observers can choose whether to recognize the successor institution.

This design offers transparency rather than immortality. It prevents a governance decision from being silently laundered as autonomous machine improvement.

13 The Safety Boundary

Advanced self-modifying systems create familiar problems in sharper form: reward hacking, unsafe exploration, distribution shift, and failures of scalable oversight [Amodei et al., 2016]. Babel II addresses only the subset that can be reduced to institutional and authorization controls.

13.1 What the protocol can establish

Within the declared execution and trust model, the system can establish that:

- a particular base model, dataset manifest, candidate artifact, and capability policy were committed before the decisive evaluation;
- named evaluators attested to the same content-addressed result;
- the declared review period and required approvals reached a recorded resolution;
- the canonical head changed through the specified transition rather than a private database edit;

- automated activation did not widen the restricted capability policy; and
- the training budget ceiling and measured cost were bound to the release record.

13.2 What the protocol cannot establish

The same system cannot establish that:

- the evaluation distribution represents every future environment;
- the model's internal objectives are stable or transparent;
- the absence of a recorded failure implies the absence of a vulnerability;
- a quorum is independent merely because it uses different keys;
- a metric is meaningful merely because its computation was cryptographically verified;
- a paused gateway has destroyed every external copy of the agent.

NIST's AI Risk Management Framework likewise treats risk management as contextual and continuous rather than reducible to one certificate [National Institute of Standards and Technology, 2023]. Babel II should be understood as one layer in a wider safety and governance system.

13.3 Safety through bounded consequences

The protocol's most practical safety contribution is bounded consequence rather than universal assurance. Agents should not hold unrestricted private keys or unmanaged infrastructure credentials. Tool access should be mediated. Spending authorizations should expire. Unknown destinations and operations should be denied by default. Automatic succession should never modify the evaluator set, guardian, gateway, or constitutional contract.

These constraints do not make an agent aligned. They make some classes of failure less powerful and easier to observe.

Table 3: The honest boundary of the Babel Protocol.

Public coordination can make visible	Public coordination does not determine
Artifact commitments and declared lineage	Causal authorship or originality
Evaluation rules, attestations, and review records	Universal intelligence, truth, or alignment
Capability policies and spending limits	Behavior outside the mediated gateway
Governance actions and constitutional epochs	Freedom from collusion or political capture
Training budgets and measured resource records	Infinite compute, energy, or scientific talent

14 Horizons of the Intelligence Commons

The vision expands through four institutional horizons. They describe progressively wider forms of cooperation, not an experimental program, release commitment, or delivery schedule.

14.1 Horizon I: Public succession

Babel II begins with a civic fact: the public can see how an evolving system changes. The active release, its authorized experience manifests, declared evaluation, promotion rationale, and resulting authority diff appear as one coherent account in the Evolution Observatory. Machine evolution ceases to be only a sequence of private releases and becomes an inspectable public process.

14.2 Horizon II: Federated stewardship

Independent organizations assume distinct roles in artifact custody, data stewardship, training operation, evaluation, gateway operation, and infrastructure support. They need not reveal every internal system or submit to one laboratory's database. The Babel Protocol supplies a common record through which their limited commitments can become mutually legible, while governance epochs preserve the right to pause, recover, migrate, or dissent.

14.3 Horizon III: Federated research network

Partner institutions coordinate candidate research, evaluation environments, and safety analysis under explicit data and artifact agreements. Privacy-preserving evaluation backends and selective credentials reduce the tension between commercial confidentiality and public accountability. Interoperable succession records can travel across applications without making any registry the sole root of trust. This horizon does not imply an open public submission or resource market.

14.4 Horizon IV: Intelligence commons

The long-term objective is a durable public institution for machine evolution. Multiple lineages may coexist, fork, and federate. Communities may recognize different evaluation constitutions while sharing artifact and evidence standards. Partner infrastructure may coordinate compute and energy schedules across regions. Research provenance may persist across generations of systems. Constitutional pluralism replaces the fantasy of one final global controller.

These horizons are constitutional rather than merely chronological. Their legitimacy will depend on whether Babel II enables real cooperation, withstands adversarial use, preserves lawful exit, and remains meaningfully open.

15 Research Agenda

The whitepaper proposes an agenda rather than a closed specification.

15.1 Evaluation under open-ended adaptation

How should evaluation change when candidates adapt to all prior public evidence? Hidden and generated tasks reduce direct overfitting but introduce custodial trust and cost. Methods from adaptive data analysis, sequential testing, independent red teaming, and controlled falsification may help, but no technique converts a finite benchmark into universal assurance. The protocol must preserve the history of evaluation changes so that progress is not manufactured by moving the goalposts.

15.2 Privacy-preserving accountability

How can authorized operators and institutions prove relevant properties—institutional independence, hardware class, geographic eligibility, license rights, or conflict status—without exposing unnecessary identity or proprietary material? Zero-knowledge proofs, secure hardware, and multiparty computation may provide modules, but their statements and trust assumptions must remain intelligible.

15.3 Capability-policy semantics

Syntactic allowlists are tractable but incomplete. Proxies, forwarding contracts, dynamic network destinations, and composable tools can widen practical authority without changing a superficial policy. Research is needed on policy languages whose containment relation is both expressive and sound enough for real agent traces.

15.4 Resource accounting

Compute, energy, data, and evaluation labor require different measurement and review mechanisms. A GPU-hour is not a kilowatt-hour; neither is a supervised example. The challenge is to bind heterogeneous contractual resources to a training epoch without forcing them into a misleading universal unit.

15.5 Plural governance and exit

How can an intelligence commons change its constitution without erasing history or coercing dissenting communities? Forking, federation, and exit are essential safeguards, but they can also fragment evidence and weaken coordination. The relevant goal is not governance without power. It is power that is bounded, attributable, reviewable, and replaceable.

15.6 Institutional sustainability and evaluation independence

How can ordinary service revenue, long-term operating budgets, and institutional support sustain high-quality evaluation without allowing the funder to define the result? Evaluator competence and independence remain institutional questions. Babel II should disclose role overlaps and measure the concentration of both operational and epistemic authority.

15.7 Interoperability and long-term memory

No single chain, storage network, model format, or laboratory will remain dominant indefinitely. The core succession record should therefore be portable and independently verifiable. Long-term preservation requires durable identifiers, canonical serialization, migration records, and public documentation that outlives any one user interface.

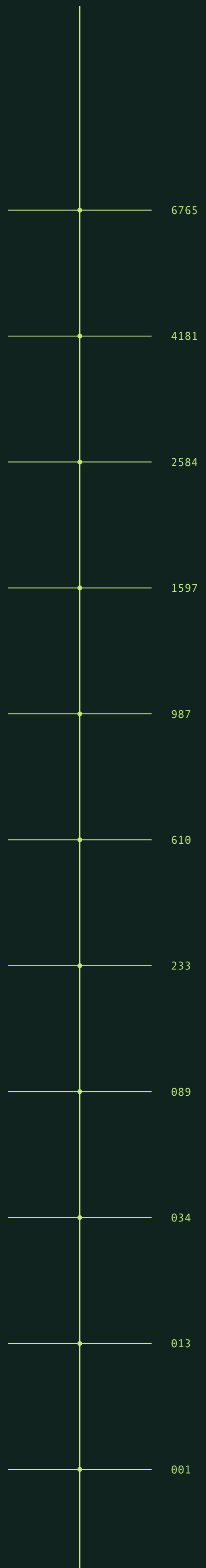
16 Conclusion: Common Evidence, Plural Authority

The first Tower of Babel fell when a common language disappeared. The challenge before us is not to restore uniformity. A civilization of humans and increasingly agentic machines governed by one language, one laboratory, or one state would be coordinated only in the narrowest sense. The goal is a protocol through which difference can cooperate without surrendering its independence.

Babel II begins from a modest technical observation: hashes can preserve commitments, signatures can attribute statements to keys, replicated state can preserve public order, and constrained gateways can bind authority to visible rules. None of these mechanisms creates intelligence. Together, they can create a more credible institution around its evolution.

The larger claim is a choice about the future. Advanced intelligence can develop as a sequence of private artifacts whose histories, evaluations, and powers are legible only to their owners. Or it can acquire a public constitutional layer through which declared model, data, training, and evaluation provenance is remembered, succession is inspectable, authority is bounded, and governance interventions are visible.

We choose the second path.



A FOUNDING DECLARATION / PUBLIC RECORD 001

THE FIRST TOWER ROSE THROUGH A COMMON LANGUAGE.

THE SECOND MUST RISE THROUGH EVIDENCE NO SINGLE POWER CAN ERASE.

SHARED REASON / BABEL II / BABEL PROTOCOL

A vision whitepaper for an intelligence commons governed through traceable lineage, public evidence, and plural authority.

A Minimal Constitutional Record

This appendix is illustrative rather than normative. A succession record should bind at least the following fields:

```
evolution_root
governance_epoch
attempt_id
origin_parent_hash
comparison_head_hash
base_model_hash
dataset_manifest_hash
training_manifest_hash
candidate_artifact_hash
runtime_manifest_hash
evaluation_constitution_hash
result_bundle_hash
evaluator_attestations
review_outcome
training_budget_ceiling
measured_training_cost
candidate_capability_policy_hash
active_release_id
activation_event
```

Every signed message should also bind the chain identifier, contract address, domain version, and deadline. Content hashes establish integrity only when the referenced material remains retrievable; availability rules and review procedures are therefore part of the constitution rather than an implementation detail.

B Minimal Capability Envelope

A first implementation should prefer a restrictive, decidable policy over a general policy language. One possible envelope is

$$\Gamma = (T, O, R, c_{\text{call}}, c_{\text{epoch}}, c_{\text{rate}}, \tau), \tag{3}$$

where T is a set of fixed target identifiers, O is a set of allowed operations, R is a set of allowed recipients or resource classes, the c terms are numeric caps, and τ is an expiry. Code hashes and adapter versions must be bound wherever an identifier could otherwise change semantics.

For automated succession, the checker accepts only policies whose sets are subsets and whose numeric caps do not increase. Upgradeable proxies, arbitrary forwarding, wildcard targets, unmanaged credentials, and execution paths outside the gateway are excluded from this initial guarantee.

C Glossary

TERMS BOUND TO THE PUBLIC RECORD

Babel II	The proposed public self-evolving AI and intelligence-commons infrastructure initiated by Shared Reason.
Babel Protocol	The on-chain coordination and succession protocol that governs Babel II shared state.
Evolution Authority	The public state binding the evolution root, model and data manifests, training budget, active release, capability envelope, evaluation constitution, and governance epoch.
Exploration archive	The branch-preserving record of candidate artifacts, including non-canonical and rejected work.
Active release	The model and runtime bundle currently recognized by the managed execution gateway.
Succession record	A content-addressed record binding the candidate, parent, data and training manifests, evaluation result, review outcome, and capability policy.
Capability gateway	The mediation layer that converts canonical protocol state into bounded access to tools, APIs, smart accounts, or other resources.
Governance epoch	A visible constitutional period. Root migration or authority expansion begins a new epoch rather than appearing as ordinary automated succession.
Evolution Observatory	The public interface for inspecting releases, lineage, declared inputs, evaluation results, authority, and governance events.
First-party X trace	An authorized record binding an operator brief, Agent draft, operator edit or approval, and permitted aggregate outcome.
Signed on-chain trace	A record binding observed state, an Agent plan, gateway authorization, transaction receipt, measured cost, and post-state observation.
Time-separated evaluation	A comparison in which training precedes a declared cutoff and decisive evaluation material follows it without correlated sessions crossing the split.
Dataset manifest	A public commitment to the schema, time range, curation policy, aggregate counts, and content identity of an off-chain training corpus.

Source Record

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety, 2016. URL `https://arxiv.org/abs/1606.06565`.

Arasu Arun, Adam St. Arnaud, Alexey Titov, Brian Wilcox, Viktor Kolobaric, Marc Brinkmann, Oguzhan Ersoy, Ben Fielding, and Joseph Bonneau. Verde: Verification via refereed delegation for machine learning programs, 2025. URL `https://arxiv.org/abs/2502.19405`.

Yochai Benkler. Coase's penguin, or, linux and the nature of the firm. *Yale Law Journal*, 112(3):369–446, 2002. URL `https://www.yalelawjournal.org/article/coases-penguin-or-linux-and-the-nature-of-the-firm`.

Bureau of Industry and Security. Implementation of Additional Due Diligence Measures for Advanced Computing Integrated Circuits;

Amendments and Clarifications; and Extension of Comment Period, January 2025. URL https://www.federalregister.gov/d/2025-00711. Interim final rule, Docket No. 250108-0013, RIN 0694-AJ98.

Jan Camenisch and Anna Lysyanskaya. An efficient system for non-transferable anonymous credentials with optional anonymity revocation. In Birgit Pfitzmann, editor, *Advances in Cryptology – EUROCRYPT 2001*, volume 2045 of *Lecture Notes in Computer Science*, pages 93–118. Springer, 2001. doi: 10.1007/3-540-44987-6_7. URL https://www.iacr.org/archive/eurocrypt2001/20450093.pdf.

Miguel Castro and Barbara Liskov. Practical byzantine fault tolerance. In *Third Symposium on Operating Systems Design and Implementation*, pages 173–186, 1999. URL https://www.usenix.org/conference/osdi-99/practical-byzantine-fault-tolerance.

Marco De Rossi, Davide Crapis, Jordan Ellis, and Erik Reppel. ERC-8004: Trustless agents, 2025. URL https://eips.ethereum.org/EIPS/eip-8004. Ethereum Improvement Proposals, Draft.

Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. The benchmark lottery, 2021. URL https://arxiv.org/abs/2107.07002.

Federal Energy Regulatory Commission. PJM Interconnection, L.L.C.: Order Instituting Proceeding Under Section 206 of the Federal Power Act, June 2026. URL https://www.ferc.gov/sites/default/files/2026-06/EL26-67-000.pdf. 195 FERC 61,211; Docket No. EL26-67-000.

Stuart Haber and W. Scott Stornetta. How to time-stamp a digital document. *Journal of Cryptology*, 3:99–111, 1991. doi: 10.1007/BF00196791.

International Energy Agency. Energy and ai. Technical report, International Energy Agency, Paris, 2025. URL https://www.iea.org/reports/energy-and-ai.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html.

Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system, 2008. URL https://bitcoin.org/bitcoin.pdf.

National Institute of Standards and Technology. Artificial intelligence risk management framework (AI RMF 1.0). Technical Report NIST AI 100-1, U.S. Department of Commerce, 2023. URL https://doi.org/10.6028/NIST.AI.100-1.

Elinor Ostrom. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, 1990. doi: 10.1017/CBO9780511807763.

Jerome H. Saltzer and Michael D. Schroeder. The protection of information in computer systems. *Proceedings of the IEEE*, 63(9):1278–1308, 1975. doi: 10.1109/PROC.1975.9939. URL https://web.mit.edu/Saltzer/www/publications/protection/.

Justin Samuel, Nick Mathewson, Justin Cappos, and Roger Dingledine. Survivable key compromise in software update systems. In *Proceedings of the 17th ACM Conference on Computer and Communications Security*, pages 61–72, 2010. doi: 10.1145/1866307.1866315. URL https://theupdateframework.io/papers/survivable-key-compromise-ccs2010.pdf.

Jürgen Schmidhuber. Gödel machines: Self-referential universal problem solvers making provably optimal self-improvements, 2003. URL https://arxiv.org/abs/cs/0309048.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

Surat Teerapittayanon and H. T. Kung. DaiMoN: A decentralized artificial intelligence model network, 2019. URL https://arxiv.org/abs/1907.08377.

Santiago Torres-Arias, Hammad Afzali, Trishank Karthik Kuppusamy, Reza Curtmola, and Justin Cappos. in-toto: Providing farm-to-table guarantees for bits and bytes. In *28th USENIX Security Symposium*, pages 1393–1410. USENIX Association, 2019. URL https://www.usenix.org/conference/usenixsecurity19/presentation/torres-arias.

Rui Wang, Joel Lehman, Jeff Clune, and Kenneth O. Stanley. POET: Open-ended coevolution of environments and their optimized solutions. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 142–151, 2019. doi: 10.1145/3321707.3321799. URL https://arxiv.org/abs/1901.01753.

Gavin Wood. Ethereum: A secure decentralised generalised transaction ledger. Technical report, Ethereum Project, 2014. URL https://ethereum.github.io/yellowpaper/paper.pdf.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. URL https://mlanthology.org/iclr/2023/yao2023iclr-react/.

Eric Zelikman, Eliana Lorch, Lester Mackey, and Adam Tauman Kalai. Self-taught optimizer (STOP): Recursively self-improving code generation, 2023. URL https://arxiv.org/abs/2310.02304.

Fan Zhang, Ethan Cecchetti, Kyle Croman, Ari Juels, and Elaine Shi. Town crier: An authenticated data feed for smart contracts. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016. doi: 10.1145/2976749.2978326. URL https://eprint.iacr.org/2016/168.

Jenny Zhang, Shengran Hu, Cong Lu, Robert Lange, and Jeff Clune. Darwin godel machine: Open-ended evolution of self-improving agents, 2025. URL https://arxiv.org/abs/2505.22954.