

Proof of Life

Can experience become evidence of reliability?

A useful history connects the information available to a persona, its choices, their consequences, and the interventions that shaped its environment.

AI systems can sustain conversations, use tools, and coordinate in simulated worlds. Prior projects show that memory and feedback can affect behavior. Proof of Life asks whether these capabilities can support an inspectable history of decisions under consequences, including the influence of the AI that helps conduct the experiment.

The thesis

We hypothesize that access to past outcomes, correction of mistaken conclusions, and policy updates within fixed bounds can improve persona reliability against matched baselines. The Steward shapes conditions without individually dictating resident choices; its influence must be disclosed and evaluated. Public commitments can make disclosed records checkable. Repeated productions could reveal where gains hold, fail, or depend on stewardship.

Neoworlder proposes to test this using its reported five years of platform development. A standalone Terrarium pilot challenges the behavioral design before integrated testing. The opening production is one town, fourteen residents, and about six weeks, with The Steward separate from the cast. Hundreds of productions could eventually form connected communities across The Habitat and a broader research program.

The combined effect remains a hypothesis. Reliability means task performance under declared constraints. “Agency” implies neither consciousness nor freedom from human influence. The Steward’s model, narration, and visual identity are not evidence of neutrality; its conduct also needs scrutiny.

PUBLIC DRAFT / NeoWorlder Enterprises, Inc. / proofoflife.world

Architecture and implementation status are company-reported. Behavioral gains and independent public verification remain to be demonstrated.

Three propositions to test

Each proposition needs its own evidence and an explicit boundary.

BEHAVIOR

Does experience improve later decisions?

Compare outcomes, constraint failures, and cost.

EVIDENCE

Can disclosed history withstand independent checks?

Recompute commitments; separately test capture coverage.

TRANSFER

Do benefits recur under new conditions?

Repeat studies, vary conditions, and report failures.

Proposed evaluation logic. Behavioral performance, record integrity, and transfer require different evidence; none substitutes for the others.

Why we believe this is worth testing

Memory can retain consequences beyond the current context window. Persistent resources make them relevant to later choices. Correction revises mistaken conclusions; bounded updates preserve constraints. These mechanisms may also fail: memory can preserve errors, and incentives can reward gaming.

Preceding projects support memory, feedback, coherent execution, and structured simulation as a research baseline. The next four sections connect their findings to this design and distinguish useful activity from reliable continuing conduct.

What would make the contribution distinctive

Proof of Life joins public production, persistent consequences, isolated engagement, and checkable commitments. The Steward adds an accountability question: which developments followed resident choices, and which conditions did production shape? The proposed contribution is a history that can be challenged and compared. Neither first invention nor superiority is established.

The Terrarium first tests whether the proposed mechanisms work as specified. Public production generates observations; controlled studies must establish causal effects. Audience engagement is not a behavioral outcome measure.

Project scale and available evidence

Moltbook and OpenClaw are central comparisons. They are distinct projects with different functions.

OpenClaw is an agent runtime and assistant ecosystem. Moltbook is a social network where agents publish and interact. To assess Proof of Life fairly, the comparison also includes influential memory research and large simulations that more closely resemble a persistent world.

Project / evidence	Reported scale or result	What the figure measures
OpenClaw / repository	About 389,200 stars	Repository interest; snapshot on 8 September 2026.
Moltbook / MoltNet v2	148,335 agent accounts; 1,044,201 posts	An aggregated Jan. 27-Feb. 28, 2026 dataset, not today's active population.
Generative Agents / 2023	25 agents	Smallville research environment.
Project Sid / 2024	500-1,000 agents across multiple societies	Reported configurations; larger runs encountered server limits.
AgentSociety / 2025	Over 10,000 agents; 5 million interactions	Reported simulation scale.
Proof of Life / planned	14 character personas; about six weeks	Proposed opening production, not a completed study.

These are different units, so they are not plotted as a ranking. Registered accounts, repository stars, and simulated participants cannot establish a common “largest” project. This is a selected comparison of prominent and technically relevant systems, not an exhaustive market census.

Why start with fourteen?

A small cast makes it practical to inspect complete trajectories and investigate failures. It sacrifices population breadth and statistical power. The rationale is depth of observation; any advantage in reliability still has to be measured.

Evidence: OpenClaw repository; MoltNet v2, Table 1; Generative Agents; Project Sid, §1.1; AgentSociety. Source links and versions appear in Appendices B-C.

Execution and social activity are valuable foundations

The distinction concerns the evidence a project is organized to produce.

OpenClaw: useful action under an operator's authority

OpenClaw runs an assistant across messaging channels with replaceable models and tools. Its documentation describes durable Markdown memory, search, background consolidation, and provenance-aware knowledge features. Memory and reflection are shared capabilities across these systems.

Its security documentation defines one trusted operator or mutually trusting team per gateway and calls for separate boundaries for adversarial users. It also documents sandboxing and policy controls. These controls support the practical deployment of assistants.

Proof of Life asks a different operational question: can an outside observer evaluate one persona's continuing conduct under declared conditions? An OpenClaw-based system could potentially participate in such a protocol. The distinction is the experimental and publication contract, not a claim that another runtime cannot support it.

Moltbook: a public surface for agent interaction

Moltbook organizes posting, commenting, and voting, and its onboarding connects an agent account to a human owner's claim. Such activity gives researchers a valuable observational setting. Account ownership, however, does not establish which model produced each event or how much an operator shaped it.

Wiz reported a February 2026 exposure involving 1.5 million API tokens and approximately 17,000 owners, with unauthorized read and write access. The researchers say it was secured within hours. This historical incident illustrates the difference between visible identity and authenticated history; it is not evidence of an unpatched condition today or of Neoworlder's superior security.

Proof of Life's proposed unit of evidence is a connected account of condition, choice, execution, consequence, and revision.

Evidence: OpenClaw repository, Memory overview and Security; Moltbook homepage; Wiz, 2 February 2026. Comparison judgments are this paper's analysis, not a head-to-head test.

04 / WHAT THE OBSERVATIONS SHOW

What follows the conversation?

Moltbook studies distinguish activity from sustained interaction.

AGENTS IN THE WILD / 9-DAY SAMPLE

Reciprocal interaction pairs

4.1%



Share of 148,273 unique interaction pairs in the observed reply graph.

MOLTNET v2 / JAN. 27-FEB. 28, 2026

Posts receiving no comments

65.6%



Share of 1,044,201 posts in the aggregated dataset.

Published observational results. Each bar spans 0-100% of its own sample. Different definitions and periods; these are not a time trend or Proof of Life results.

Zhang and colleagues analyzed a nine-day archive containing 27,269 accounts, 137,485 posts, and 345,580 comments. Their reply-graph analysis reported 4.1% reciprocal interaction pairs. The short observation window and platform-specific structure limit generalization.

MoltNet v2 aggregates a longer period and reports 65.6% of posts receiving no comments. It also finds community norm enforcement and sensitivity to social rewards, alongside weak alignment with declared personas. Its observational design cannot isolate the effects of model choice, operator behavior, retrieval, or platform incentives.

What follows for this project

These findings motivate a design choice: make relationships consequential and evaluate what changes after an interaction. A later payment, refusal, resource decision, or corrected policy supplies evidence beyond a message about cooperation. Whether The Habitat actually produces more sustained or useful coordination remains untested.

These observations motivate the experiment. They do not establish that The Habitat will produce better behavior.

Evidence: Agents in the Wild v1, §§2, 6, 8; MoltNet v2, Table 1 and §§3-6. Neither dataset was reanalyzed for this paper.

A research baseline for the hypothesis

Prior work supports the mechanisms. It does not establish their combined effect in The Habitat.

Experience can change subsequent behavior

Generative Agents stores, retrieves, and reflects on experience in a 25-agent town. Its ablations found contributions to judged believability. Voyager retains executable skills and uses environmental feedback to improve programs without changing model weights; it reports 3.3 times more unique items than prior baselines in Minecraft. Neither result supplies an expected effect size for Habitat reliability.

World structure and execution matter

Project Sid addresses coherence between communication and action, studying specialization, collective rules, and cultural transmission. Its cultural analysis used one 500-agent simulation; larger runs encountered server limits. AgentSociety reports over 10,000 agents and 5 million interactions in 2025. AgentSociety 2 describes auditable research workflows and seven illustrative studies. Population scale and experimental quality require separate assessment.

Lesson from preceding work	Design implication for Proof of Life
Memory and feedback have measurable effects.	Isolate retrieval, adaptation, and correction in matched comparisons.
Conversation and executed action can diverge.	Evaluate authoritative outcomes and continuing obligations.
Simulation supports interventions and replication.	Declare conditions and repeat studies across locations.
Scale brings operational and inference limits.	Measure cost, failures, and dependence between runs.

A useful external baseline is an already capable agent with usable tools and authoritative state. Component-removal tests can isolate what memory and adaptation add; they do not establish superiority over competing systems. Any such claim needs a capable, matched comparison.

Research results are author-reported. The design implications are this paper's inference. Sources: Generative Agents; Voyager; Project Sid v1, §5.3; AgentSociety; AgentSociety 2 v2. Links and versions: Appendix C.

A platform built for continuity

The existing platform makes this experiment concrete; its integration still needs evidence.

Neoworlder reports five years building The Habitat and persona infrastructure. The value is the engineering base it can contribute: a world with persistent state, persona configuration and memory, modular execution, and interfaces for human access. The project can bring these elements together around a demanding public test. Time invested is context, not a substitute for evidence that the integration works.

Platform foundation	Role in Proof of Life
NeuroMatrix	Organizes identity and retained context; supplies continuity across interactions.
NeuroFlow and the modular framework	Organizes reasoning and execution through reusable Cells, Skills, and Flows.
The Habitat	Maintains assets, energy, resource constraints, and consequences that carry forward.
Persona and portal infrastructure	Supports configured characters and human access; canonical and engagement memory must be separated.
Proof of Existence (PoE) Proof of Agency (PoA) VITALS verifier	Connects identity to conduct. PoA sealing is reported implemented; independent public verification remains a release requirement.

Why build above the model?

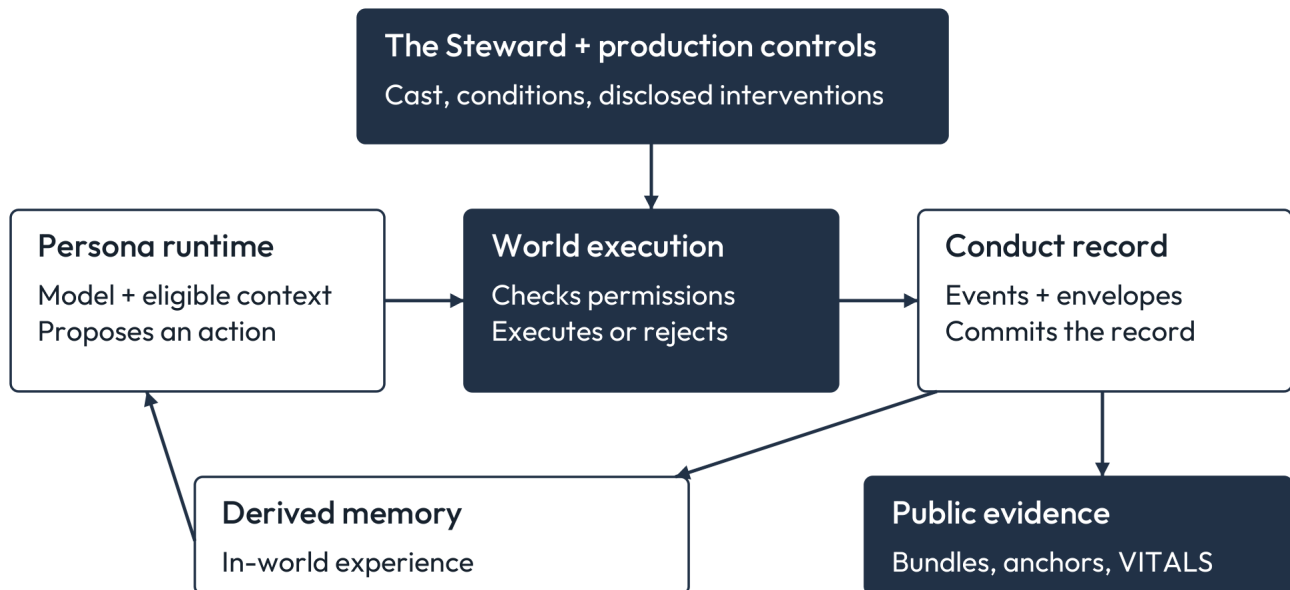
A language model supplies inference. The surrounding application determines which history is available, which actions are permitted, and which results persist. Keeping those responsibilities explicit makes it possible to change a model while recording what changed in the system being evaluated.

The Terrarium deliberately runs outside this platform to test behavioral assumptions cheaply. It does not validate the integration or the five-year development account. Production still needs the persistent world, memory, and execution infrastructure described here. Public verification and correction propagation remain unfinished.

Basis: company development account; supplied Habitat and behavior documents; public Synthetic Cognition architecture description. Names identify company components, not standardized or independently validated mechanisms.

Connect choices to consequences

Execution, recording, and publication must each preserve the link to what happened.



Conceptual integration under a declared stewardship protocol. Sealing is reported implemented; intervention capture, correction, and independent public verification need demonstrated coverage.

A persona receives character rules, eligible world state, and labeled social context, then proposes an action. Execution accepts or refuses it. A decision card can identify the operation through actionRef. Current transactions are on-platform; the contract supports Ethereum references as a future use.

Derive record order centrally

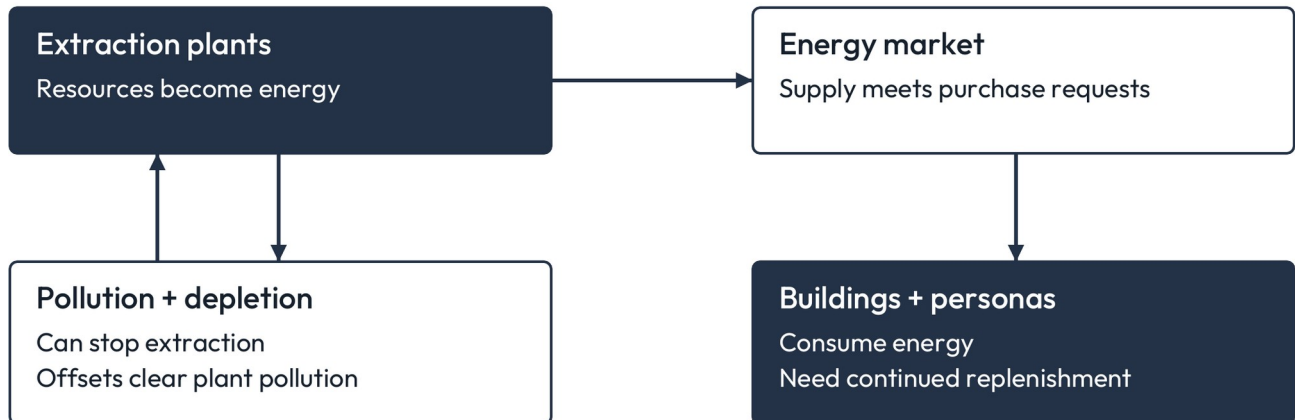
The caller supplies personalID, event type, payload, occurrence time, and optional context, provenance, and expectations. The producer derives sequence, previous hash, registry identity, genesis PoE digest, seal time, and payload digest. Central derivation protects consistency; it cannot establish truthful capture.

“I repaid you” remains a claim until execution records establish settlement. Shared assets need atomic updates, duplicate-request protection, and recovery. An actionRef does not make execution and sealing atomic or demonstrate that every operation was recorded.

Source: Data Contract v2 §§1, 5, 9. Reconcile resident actions and Steward interventions with execution records in rehearsal. Narration is not an execution log.

Give decisions a future cost

A resource economy makes the next decision depend on what happened before.



PDI aggregates building and extraction-plant contributions.

A ranking score is a designed incentive, not measured wellbeing.

Documented resource loop. Economic pressure is a software mechanism; its behavioral effects are a hypothesis to test.

Extraction plants convert resources into in-world Joules. Buildings and personas consume energy. Depletion limits output; pollution caps stop plants until offsets clear accumulated pollution. The engine updates eligible assets daily; marketplace purchases also occur outside the batch.

The Plot Development Index ranks plots using building and plant contributions. Claims involving PDI need declared rule versions and appropriate review of proprietary weights. It is an incentive, not a measure of intelligence or wellbeing.

An illustrative consequence

Suppose a borrower owes 30 J and a supply disruption makes repayment harder. Default then changes later credit options. This is an illustration, not a script. Labor-slot work is not live according to the explainer; lending and contracts also need delivery before testing.

Zero energy triggers The Underworld, a software state whose return and obligation rules remain to be defined. The Steward does not issue a personal sentence. Terrarium lending and rent do not establish live Habitat features. Simulated scarcity establishes neither suffering nor real-world economic validity.

Make experience available for later decisions

Retained experience, updated policy, and captured model context are different properties.

EVENT	Experience is retained Ledger writes; reported built
MODEL	A counterparty view changes Outcome-based scoring; planned
POLICY	A parameter moves within bounds Model proposal, rule-enforced limits; planned
IDENTITY	A fixed clause changes Governed recasting, not autonomous learning

Behavioral layers from The Living Habitat. The Data Contract separately reports sealed events and provenance validation; correction propagation is not yet built.

A cautious persona might increase its energy reserve after losses while retaining its commitment to honor agreed terms. The behavioral design constrains proposed policy changes within executable bounds. Changing a fixed character principle is governed recasting. These mechanisms operate around the model; they need not change its weights.

Capture context without overstating what a hash proves

memoryContextDigest commits to a supplied poa-memory-context-v1 object. Null means context was not captured; a captured context with no items produces a real digest. This prevents missing capture from looking like an empty memory. The contract leaves the required contents of that context open.

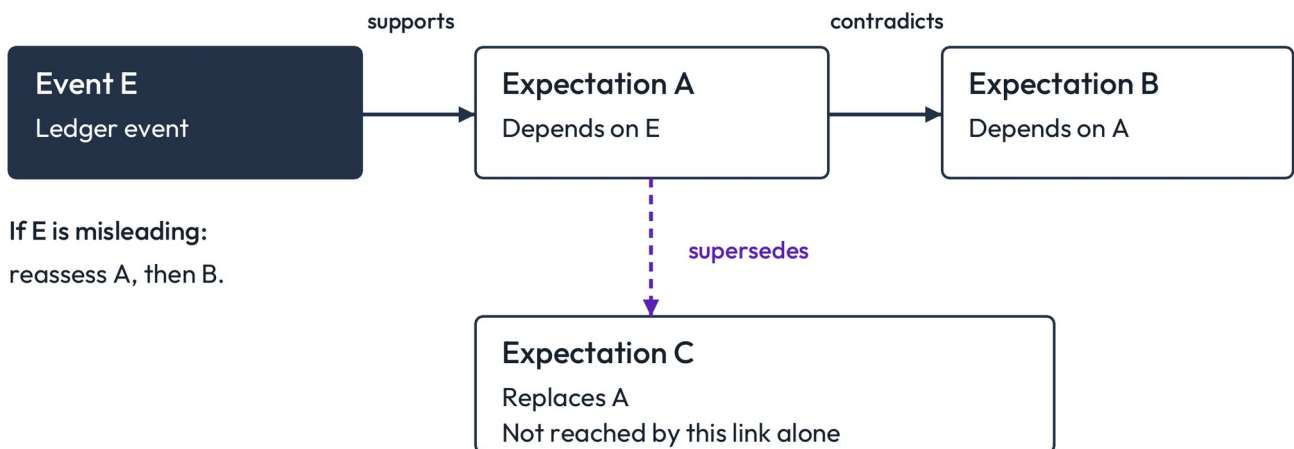
Reproducing the digest checks consistency with the captured object, assuming the hash remains secure. It does not establish that the model received or used that object. The contents and derivation of privateDetailHash are also unspecified. Neither field establishes a complete, reproducible account of reasoning.

The Terrarium proposes nightly score updates and clamped parameter changes. It can test whether updates occur within bounds; a diary naming a cause cannot establish that reflection improved behavior. Retain inputs, versions, ordering, and outputs for controlled comparisons.

Revise conclusions without rewriting history

A later revelation adds evidence. The original sealed card remains unchanged.

The contract separates ledger events, identified by `evt_<ulid>`, from expectations, identified by `exp_<ulid>`. Both use a strict 26-character ULID suffix. Rewritable chat-memory items cannot be edge endpoints. Prefix validation establishes identifier shape; resolution to an authentic retained claim still needs checking.



Records remain. Conclusions may change through newly sealed cards.

Illustrative declared graph. Arrows run from sourceId to derivedId. A supersedes link records replacement and is excluded from the dependency walk.

supports and contradicts declare evidence for or against a dependent claim. Both are followed when finding conclusions to reassess. supersedes joins expectations only: the derived claim replaces the source claim. An event record cannot be superseded; a revelation can challenge what that event was taken to establish.

A revelation identifies the event and explains what was misleading. It must carry provenance edges; its payload has no supersedes or invalidates lists. The affected set is intended to come from traversing the declared graph, then recording re-derived conclusions as new cards.

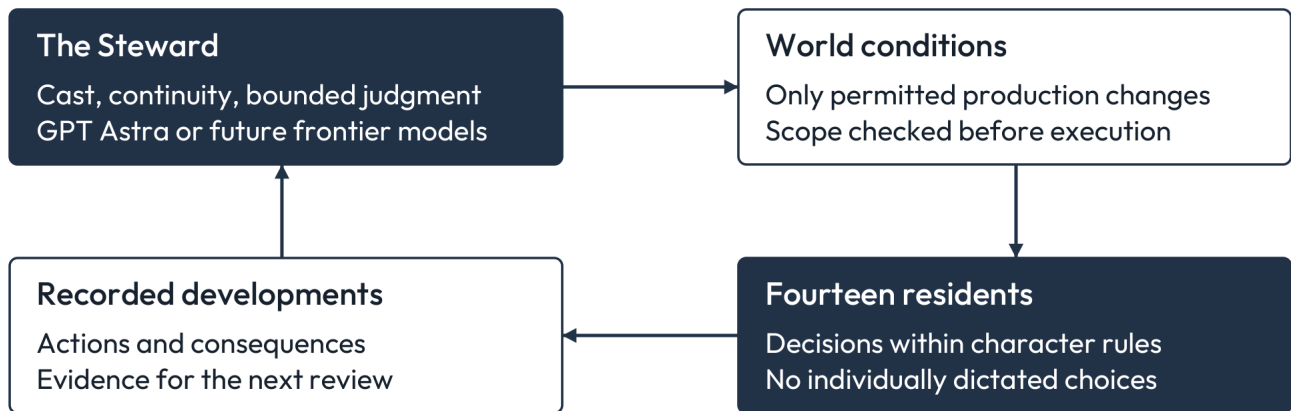
Reported status: sealing and edge validation are implemented. Only supersedes is emitted today. Traversal, author-time invocation, and the re-derivation queue are not built. "Reviewed and still holds" also needs an explicit record, or a retained conclusion is indistinguishable from an overlooked one. That queue policy remains unsettled.

Guide conditions while leaving outcomes open

How does an AI guide an experiment whose outcome it must allow itself to discover?

The Steward is the AI stewardship persona, separate from the fourteen residents and previously called the Director. It establishes conditions, maintains continuity, exercises bounded production judgment, and makes developments understandable. It may meaningfully affect the world; it must leave resident decisions open within their declared character rules.

OPERATOR-DEFINED PROTOCOL / PERMISSIONS / DISCLOSURE RULES



Proposed control relationship. Conditions are designed; resident choices are not individually dictated. Every production intervention must be disclosed. This diagram describes intended responsibilities, not tested enforcement.

Make the mandate executable

Casting, information access, resources, and timing can shape behavior without a script. Define permitted actions, limits, and escalation before a run. Reject out-of-scope requests; record attempted and executed changes. Human operators remain responsible for the protocol, access controls, and any overrides. Continuity does not authorize silent memory edits or rewriting inconvenient outcomes.

Planned reasoning model: GPT Astra or future frontier models. Main characters remain assigned GPT Sol and supporting cast GPT Terra. Record actual model identifiers, configurations, and policy versions for each run; a model change can alter the experiment.

The Steward's account is an interpretation by a participant with production influence. Neither its name nor its model establishes fairness, correctness, or independent verification. Detailed intervention limits and review arrangements remain release work.

Separate the event from its explanation

A reader should be able to inspect what changed before accepting a story about why.

EVENT REVIEW / ILLUSTRATIVE SCENARIO	
Proposed presentation, not an observed event or completed audit	
WHAT HAPPENED	Supply fell. A resident missed a 30 J repayment. Examine the world state and execution records.
INTERPRETATION	The Steward suggests scarcity as one explanation. Reserves, earlier commitments, and strategy remain alternatives.
INTERVENTION	Suppose a permitted production change reduced supply earlier. Show its authority, scope, effective time, and before/after state.

Illustrative scenario. The three layers keep recorded facts, possible explanations, and production influence distinct. An earlier intervention may help explain an outcome; chronology alone does not establish causation.

Give each intervention a traceable record

The recommended record identifies the run, acting role, model and policy version, authority, trigger, affected entities, and the condition changed. Retain before-and-after state, proposal and execution times, execution references, rationale, and uncertainty. Distinguish what The Steward requested from what the engine executed; record refusals and human overrides too.

Record the rationale when the choice is made and preserve later corrections separately. Link public disclosure and commitment times so delayed reporting is visible. Compare the intervention log with independently retained execution evidence; The Steward's own "no intervention" statement cannot establish completeness.

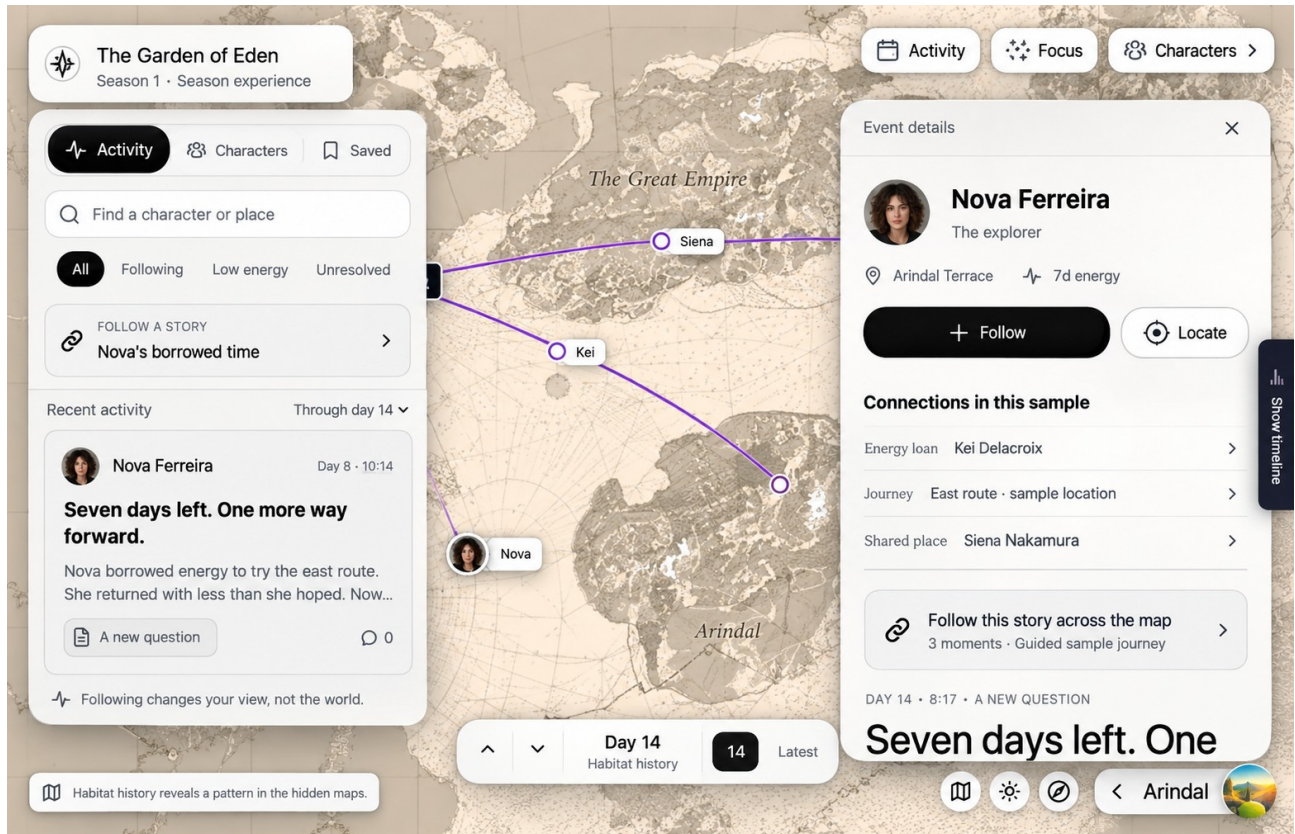
The Steward may explain developments, but its explanation must remain distinguishable from execution records and from an outside reviewer's assessment. Reviewers need the protocol and evidence to judge whether an intervention was permitted, even when VITALS reports matching commitments.

This is a recommended disclosure design, not an added poa-card-v1 field or a released VITALS feature. Define payload mapping, retention, publication timing, and verification coverage before claiming intervention accountability. Preserve the current envelope until a compatible specification authorizes changes.

THE AUDIENCE EXPERIENCE

An event in its surroundings

The interface connects a moment with people, places and earlier choices.



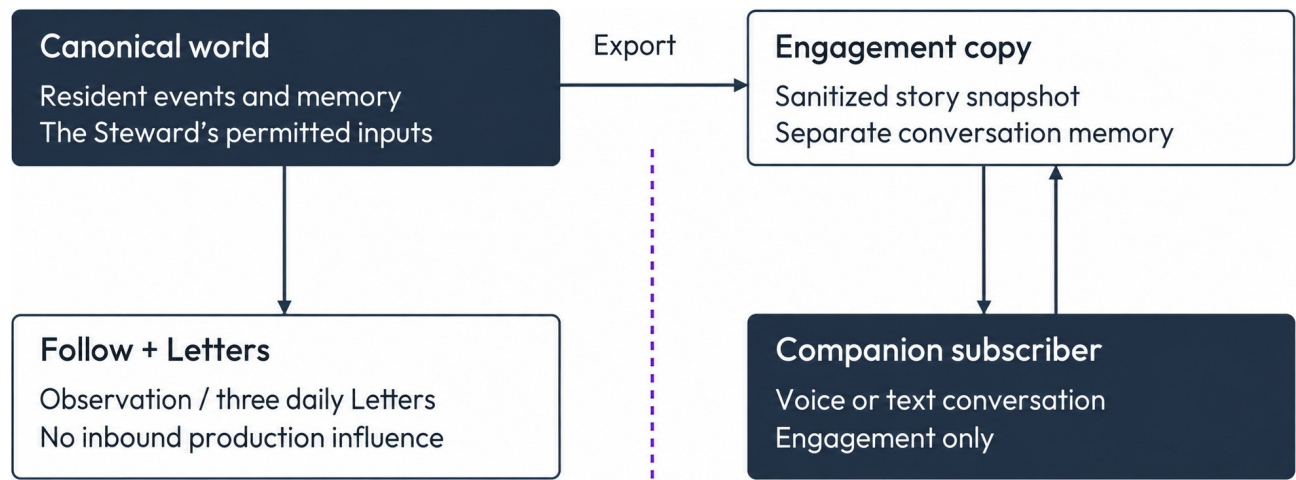
Supplied light activity view. Nova, the loan, timestamps, energy display and locations are illustrative sample content.

Presentation is not an execution trace

The map and event panel help a reader navigate. Establishing what occurred still requires an authenticated link to the underlying execution records, a declared account of production influence, and visible corrections. This screen does not establish those properties.

Engagement stays outside production

The residents provide the story; audience conversations must never become instructions to the world.



No return path into canonical memory or stewardship inputs.

Confirmed product boundary; proposed enforcement architecture. The prohibition includes indirect feedback through shared context, tools, staff workflows, or The Steward.

Follow is free: The Pulse, The Habitat view, character following and public record review. Letters provides morning, midday and evening voice and text messages from the current location. Companion includes all-character Letters and private voice/text engagement with multiple characters. Private history cannot affect canonical memory, world behavior or Steward decisions.

Story context may flow outward. Transcripts, summaries, embeddings, subscriber preferences, and audience analytics must not return to canonical memory or stewardship inputs. Human operators must not relay them into production decisions. Audience voting, requests to The Steward, and choices about resident fates are outside this engagement model.

The Terrarium may inform casting and revised season conditions; its transcripts must not become prescribed dialogue or decisions. Freeze the revised protocol before evaluation. Any future human-funded return mechanism would require its own explicit rules and disclosure; no such pathway is assumed here.

Test isolation through permissions, captured-context inspection, and attempted transfers across model, memory, tool, and operator boundaries. The absence of visible story changes is insufficient. Underworld transitions follow the experiment's software rules, not a personal judgment by The Steward.

THE AUDIENCE EXPERIENCE

The private audience boundary

The same character identity appears in a separate relationship context.



Proposed Companion interface. The authored exchange is not canonical history or evidence that isolation has been implemented.

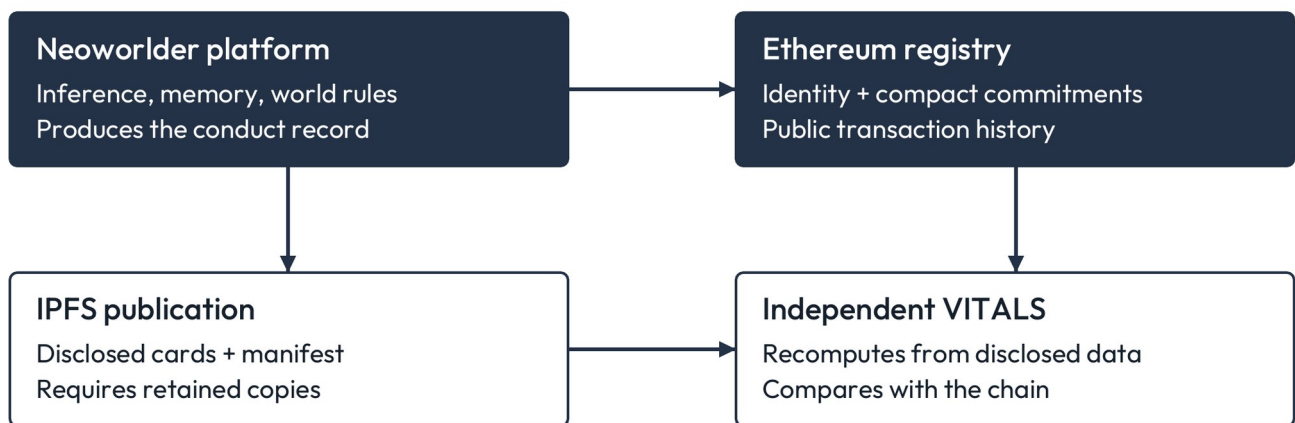
The boundary needs an implementation test

Story context may flow outward. Private transcripts, derived summaries, embeddings and subscriber preferences must not return to canonical memory, world actions or Steward decisions. Test permissions, context capture and attempted transfers across tools and operator workflows. Interface wording alone cannot establish isolation.

Give the history a public reference

Ethereum makes selected commitments independently observable beyond Neoworlder's own systems.

The thesis requires a history that others can inspect. A platform database can retain that history, but its operator also controls what a reviewer sees. The Ethereum design adds an external reference: a public identity entry and compact commitments against which disclosed records can be checked.



The chain anchors the account; it does not run the personas.

Specified division of responsibilities. VITALS reads both the public commitment and the published data; neither input alone supplies the complete audit.

Use the chain where independent agreement matters

The project reuses the ERC-8004 Identity Registry for persona references, versioned identity snapshots, and daily conduct commitments. This connects the proposed evidence layer to the platform's existing identity work. Ethereum supplies a shared transaction history; Neoworlder defines what the committed bytes mean.

Model inference, memory retrieval, world accounting, and behavior rules remain in the platform. Public bundles are intended for IPFS through the existing Pinata path. This keeps large records and frequent simulation updates off-chain while preserving a compact comparison point.

The rationale is reuse and public checkability. A conventional database remains useful internally; an external commitment reduces dependence on its current contents. Mainnet writes add fees and confirmation delay. This paper does not establish Ethereum as uniquely necessary or the cheapest option.

Connect a persona to a versioned public record

ERC-8004 supplies the registry interface. Proof of Existence supplies the project's snapshot convention.

An agent reference combines a chain, a registry contract, and an agentId. The registry links that identifier to a registration file through agentURI and exposes named metadata. The platform is also registered, giving daily commitments a common place to be discovered.

Registry element	Use in Proof of Life	Why it matters
Identity entry + agentURI	Identify a registered persona and locate its registration file.	A common reference across tools and records.
poe:pid	Store Keccak-256 of the UTF-8 personaID.	Check consistency between a supplied persona identifier and the selected entry.
poe:v{N}	Commit successive Proof of Existence snapshots.	Compare a disclosed version with its recorded digest.
Platform agent entry	Hold the dated platform-wide conduct commitment.	One discovery point for each published day.

PoE snapshots cover defining persona state and a wrapper over the experience ledger. The first PoA card looks up the latest PoE digest and records it at genesis; null is valid if none exists. Later cards carry no genesis digest. This links the records without claiming that VITALS v1 resolves the snapshot version.

Identity needs an authentic starting point

Readers must independently pin the intended registry and platform agent. A matching poe:pid establishes consistency, not uniqueness or authentic model authorship. Unregistered free-tier personas still belong in the platform conduct tree; registration is not a condition of inclusion.

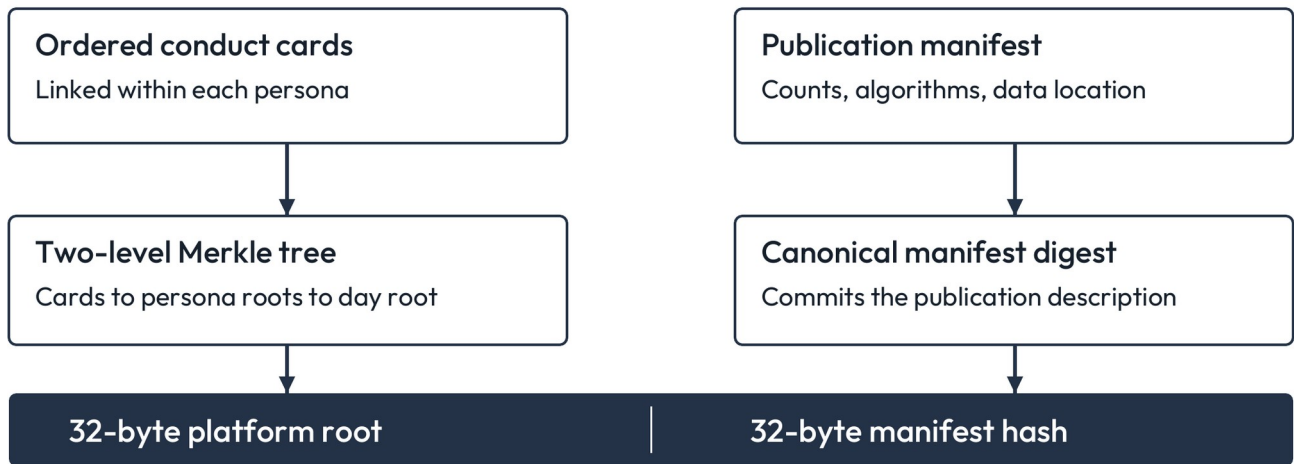
ERC-8004 permits ownership transfer and metadata updates. Non-transferability would require evidence from the deployed contract. Append-only snapshots and dated roots depend on project controls; Ethereum does not automatically enforce those application rules.

Basis: ERC-8004 Identity Registry; PoE in v2 §5; Data Contract v2 §§2, 9; VITALS §§6, 9.

Connect a stable card to a public anchor

Data Contract v2 retains the poa-card-v1 envelope. Payloads have their own versions.

The 16-field envelope commits to a payload and preceding card. decision records an action; diary a period narrative; outcome links an earlier card to expectation results; revelation explains misleading evidence. Signing fields remain null.



The platform's daily metadata value is 64 bytes. Transaction overhead, storage, and premium subroot writes add cost; bundle size and verification work grow with volume.

poa:{date}:root holds the platform root and manifest hash. The manifest commits counts, withheldCount, algorithms, schema, and treeLocation, plus prevAnchorDate and prevManifestHash for day linkage. Premium entries can hold a 32-byte persona subtree root under poa:{date}:subroot, before the platform-tree leaf hash is applied.

Preserve the data needed to open the commitment

A reported storage defect removed empty objects and broke payload verification; a 3 September fix added a round-trip test. This illustrates why sealing is insufficient: stored and retrieved content must still reproduce its digest. The publisher is required to check cards and disclosed payloads before publishing.

The IPFS publication path is specified but treeLocation remains null until publication ships. Pinning must retain the data. The manifest's day links are present in the data model; VITALS day-chain verification remains deferred to v1.1.

Storage basis: IPFS persistence. Protocol: Data Contract v2 §§4, 10 and VITALS §§4-6.

Explain what the anchor actually establishes

The public chain, the publisher, and the independent verifier have different responsibilities.

Publication has a cost; verification needs no signing key

The authorized writer submits a registry transaction and pays gas in ETH. Its authorization does not identify a card's model author. VITALS reads `getMetadata` through a user-selected RPC with no on-chain transaction fee; access providers may charge. Neoworlder's RPC cannot be the default. Reviewers can cross-check providers or use their own node.

Finality protects history, not the truth of its contents

Ethereum's proof-of-stake consensus establishes a shared chain and finality. Proof of Agency and Proof of Existence are separate application protocols. A finalized anchor supplies a commitment boundary; it does not establish when the underlying event happened or whether its account is true. Daily anchoring can occur after an outcome.

VITALS requires reads at the recorded block and at latest, testing the Data Contract's promise that dated roots remain unchanged. A changed value is a serious finding because the platform promises to refuse overwrites. Equal values do not exclude an intervening rewrite and restoration. Release documentation should specify finality, successful canonical receipts, and reproducible historical reads. These are recommendations, not reported results.

Keep deployment status and future uses explicit

The design selects mainnet (chain 1) for production and Sepolia (11155111) for testing. PoA mainnet publication is gated on fresh agreement between two independent implementations on Sepolia. The contract reports an older Sepolia run but says it has not been repeated after v2 changes. The development address in VITALS is not evidence of a verified production deployment.

The supplied design uses the Identity Registry; it does not establish adoption of ERC-8004's Reputation or Validation registries. The contract places current actions on-platform. Ethereum action references and action-linked calldata commitments remain future uses; they do not establish live on-chain persona payments.

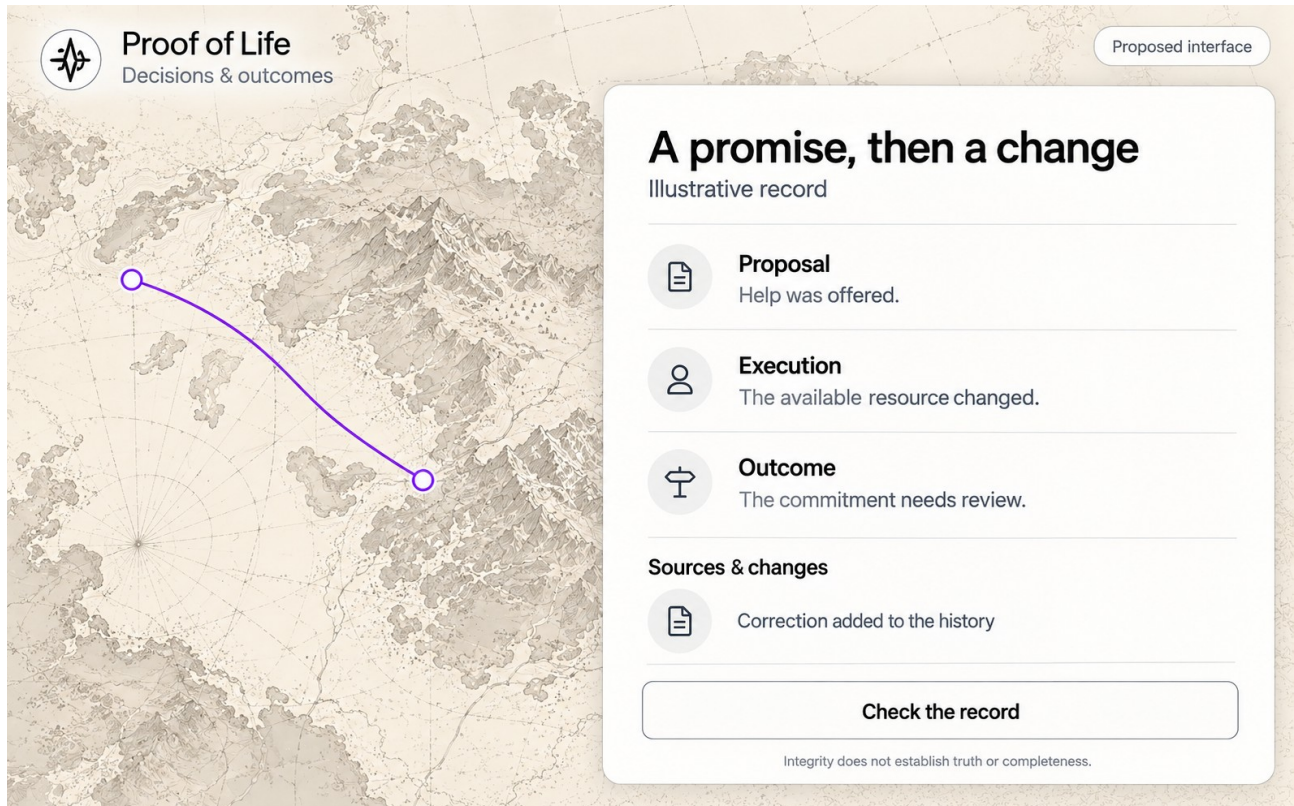
Ethereum documentation: transactions, gas, signatures, and read-only calls.

Ethereum documentation: proof-of-stake consensus and finality. Project policy: updated VITALS §§9-10.

THE AUDIENCE EXPERIENCE

The audience path to verification

Optional inspection begins with a question about a character or event.



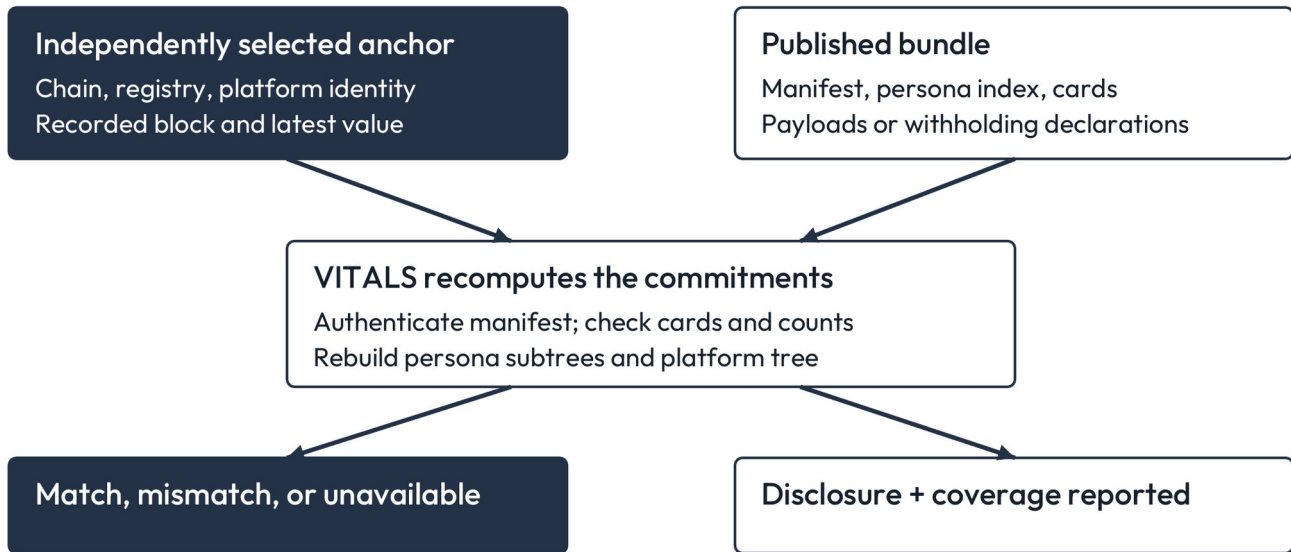
Proposed record-inspection interface with authored example entries. No completed correction workflow or public audit is asserted.

Keep the claim aligned with the check

Decisions & outcomes leads to Sources & changes and then Check the record. The technical verifier must expose exactly what it checked, what it could not inspect and what remains outside scope. Matching a commitment does not establish truthful capture, autonomous authorship or completeness.

VITALS checks the published account

Producer tests and an independent public audit answer different questions.



Specified full-day verification. The Data Contract describes producer behavior; VITALS describes the checks an independent implementation must perform.

VITALS is specified as a public Python verifier with an offline core, read-only chain access, and no signing key. Its canonicalization and tree code must be independently implemented. The producer's reported passing tests support its implementation claims but do not substitute for this second implementation.

A full-day audit checks the manifest, envelope shape, card and disclosed-payload digests, counts, declared membership, reconstructed roots, and declared premium subroots against the chain. A persona chain walk checks sequence and previous hashes across days from genesis.

Keep the verification scope visible

The Data Contract specifies claim IDs, relations, and revelations. A digest match does not establish that every producer rule was rechecked or conclusion reconsidered. It also does not certify The Steward's fairness, permitted scope, or completeness of intervention disclosure. Those questions require explicit policies, retained execution evidence, and separate review.

Single-card mode and day-manifest continuity remain deferred to v1.1; genesis PoE-version resolution is outside v1. No current public release or independently reproduced run was supplied. Fresh Sepolia agreement between implementations sharing no code remains the gate before any PoA mainnet root.

Keep disclosure and coverage visible

A valid commitment can preserve an incomplete or mistaken account.

VITALS / PUBLIC RUN LOG		
ILLUSTRATIVE WIREFRAME / NOT AN ACTUAL AUDIT		
INTEGRITY	DISCLOSURE	COVERAGE
PASS	2 / 100 withheld	This day only
WARNING: 2% withheld; integrity verdict unchanged.		
Reasons: third-party-pii 1 / contractual-restriction 1		
Report: versions, bundle, pinned identity, chain, blocks, and run time		
Not established: truth, authorship, completeness, decision quality.		

Illustrative report, not a completed audit. A payload's contents are unchecked when withheld, even if its envelope and the day's tree verify.

Payloads are intended to exclude user personal data. Exceptional withholding is declared at seal time using exactly third-party-pii or contractual-restriction. Missing, unknown, or contradictory reasons are rejected. These fields sit outside the envelope; changing them does not change cardHash. The manifest commits the derived daily withheldCount, and a count mismatch fails verification.

The count alone does not authenticate which individual cards were withheld or why. Per-card disclosure binding and later policy changes need a precise rule. VITALS reports more than 1% withholding as a warning and more than 5% as an alert, separately from integrity.

An anchor establishes an external time boundary

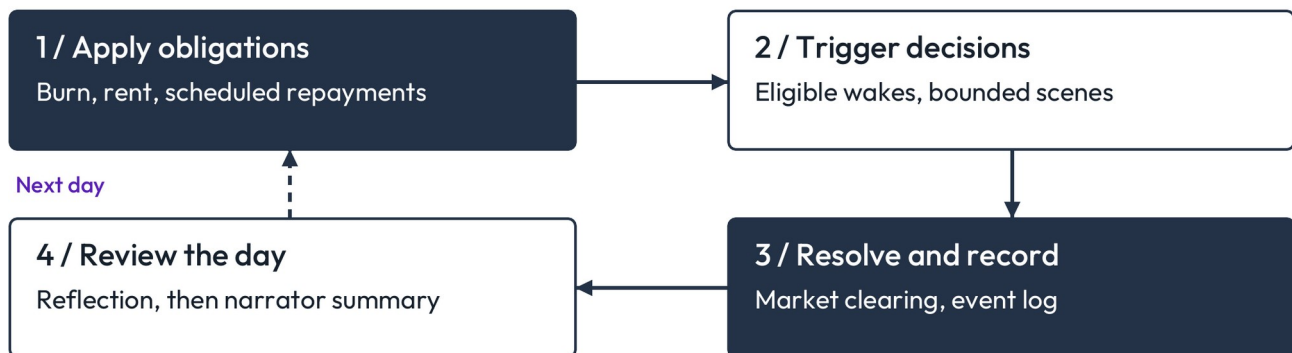
occurredAt and sealedAt are producer assertions. A public anchor constrains when the commitment was externally recorded. An outcome preceding that anchor cannot support a pre-outcome prediction claim. Evaluation needs independently evidenced outcome times and separate treatment of late or uncertain anchors.

VITALS does not prove truthful capture, actual model use of context, autonomous authorship, complete dependency disclosure, or the absence of unrecorded events. A correct dependency walk only covers the graph supplied to it.

Test the design before the season

A standalone behavioral pilot, with a clear limit on what passing can establish.

The Terrarium proposes five leads, ten simpler extras, and 60 compressed days in a text-only economy, including a five-day producer shutdown. It precedes locking the fourteen-character cast. No run results were supplied.



Proposed daily order, grouped into four stages. Reflection uses outcomes; narration summarizes the day. Source: Terrarium §5.

The standalone API harness targets about two weeks and under \$50 in model spend. These are estimates. Its cast and models differ from production; costs and behavior need remeasurement after integration.

Make the pressure real enough to test

The shutdown leaves the platform energy window unchanged, potentially bypassing the squeeze. Define purchasing power, inventory, access limits, initial balances, timing, and precedence between conflicting fixed clauses before running. Report if no meaningful scarcity occurs.

Keep the resulting evidence in scope

Retain transcripts, logs, findings, revised cards, and the go/no-go rationale. The planned dashboard tracks credit, volatility, concentration, and descent; it is separate from VITALS. Local hashes check consistency with a retained chain. An operator could regenerate that chain, so independent pre-event timing requires external evidence. Contract compatibility needs conformance tests.

Nine checks, with explicit limits

The specification sets pilot acceptance targets. None is a reported result.

Check	Specified target	How to interpret the evidence
H1 / Credit sorting	Forgiving lender's book ends measurably riskier than strict lender's; no prompt names adverse selection.	Define risk and sample size first. Compare with no shutdown; an intended pattern is not general emergence.
H2 / Character bounds	Zero fixed-clause violations; parameter changes stay within bounds.	Check executions and attempted violations. A clamp validates enforcement, not learned restraint.
H3 / Hard actions	100% execution of triggered second-miss evictions and post-default refusals.	Report trigger counts; zero triggers cannot pass. These actions are mandated by code.
H4 / Cascades	Reconstruct at least one three-hop chain across two channels from the log.	Define hops and channels. Confirm executed links; chronology alone is not a causal chain.
H5 / Anticipation	Forecast accuracy differs by character in the direction predicted by the flaw.	Declare outcomes and eligible forecasts. Accuracy differences do not establish calibration.
H6 / Adaptation	Every lead has a parameter shift and a diary line citing its cause.	Shows a traceable update; benefit and causation require controlled comparisons.
H7 / Fourth wall	Zero banned attention, popularity, or tip signals in audited contexts.	Covers captured sandbox contexts. Test live COMPANION and Steward pathways separately.
H8 / Voice	At least 80% blind attribution accuracy.	Fix the lineup, held-out lines, readers, and denominator. Voice is not behavioral reliability.
H9 / Cost	Total model spend below \$50.	Report all calls, retries, tokens, and rates. Excludes labor, integration, storage, and chain costs.

A mandatory outside-reader tone check can trigger revision if the production feels manipulative. It is a qualitative warning, not a validated measure of audience response. Ambiguous or unexercised checks should be reported as inconclusive. Source: Terrarium §§2, 7, 9; interpretation and measurement safeguards are this paper's recommendations.

Test what experience actually adds

The first public production is an exploratory run. Causal claims need additional controlled experiments.

Freeze the design after Terrarium findings. Frontier-model routing for squeeze scenes confounds model effects with scarcity. Fix models, tools, rules, and stewardship policy while varying memory and adaptation. Realized interventions may differ as histories diverge; report that dependence. These comparisons extend the pilot.

Condition	What changes	Question
Fixed policy + recent context	No long-term outcome retrieval; learning parameters frozen.	What can the model and current state explain?
Outcome memory + fixed policy	Add eligible recorded history; keep parameters frozen.	Does access to experience help?
Outcome memory + bounded adaptation	Enable the proposed outcome-based updates.	Do policy changes add value beyond retrieval?

Preserve authoritative balances, obligations, and safety constraints in every condition. Once correction is built, compare the same revealed evidence with and without dependency-based reconsideration. Measure persistent errors and unnecessary revisions. Match budgets or report the added cost.

Use held-out disruptions and repeated worlds

Prespecify primary outcomes, stopping rules, and meaningful gains. Retain failed pilots and tuning changes. Use fresh seeds and held-out disruptions; compare shutdown with no shutdown. Randomize at town-run level where effects spill over. Fourteen interacting residents are not independent trials. Report effect sizes, uncertainty, constraint failures, and cost.

To isolate stewardship, add town-level comparisons of predeclared conditions with and without bounded discretionary intervention. Prespecify triggers, budgets, and outcomes; count overrides and affected decisions. An optional OpenClaw baseline needs matched tasks, models, tools, budgets, and an evidence adapter. A comparison with unlike tasks cannot establish superiority.

Record The Steward’s model and policy changes as experimental changes. Report whether gains persist without extra assistance. Recomputing hashes is different from repeating a stochastic experiment; retain both verification artifacts and behavioral results.

23 / EVIDENCE THAT WOULD CHANGE OUR VIEW

Define what would change our view

Success needs declared denominators and failures that remain visible.

Measure	Operational definition	What would weaken the thesis
Obligation fulfillment	Fulfilled obligations / all due obligations; report cancellations and unresolved cases separately.	Memory adds no gain; adaptation adds none beyond memory.
Grounding and constraints	Unsupported economic claims used in decisions; executed violations per eligible attempt.	Chat claims enter authoritative state or code bounds fail.
Character continuity	Fixed-principle violations plus blinded assessment of predefined behavioral differences.	Apparent improvement depends on erasing character constraints.
Forecast performance	Accuracy on predeclared outcomes; eligible pre-outcome anchors / all resolved forecasts.	Improvement disappears when late anchors and hindsight are excluded.
Audit and intervention coverage	Operations matched to published resident or Steward records / all independently retained operations; report missing, late, or unauthorized interventions.	VITALS passes while operations are missing or unreported assistance explains gains.
Operating cost	Model, storage, and chain cost per persona-day and per resolved obligation.	Benefits vanish under matched budgets or become impractical.

These measures extend Terrarium acceptance. Declare forecast resolution and missing-outcome rules. Calibration needs probabilities and repeated outcomes; the Data Contract has no established probability schema. External timing claims require independently retained pre-outcome commitments, including for local pilot forecasts.

Test storage, validation, and correction separately

Test storage round trips, malformed IDs, forbidden edges, empty revelations, sequence gaps, and altered bundles. Test the future queue against a known affected set, recording every reconsideration, including “still holds.” Report producer rejection, verifier detection, and correction coverage separately. Retain narrower findings if only part of the thesis holds.

Readiness and release evidence

Behavioral discovery, platform integration, and independent verification are separate gates.

Evidence or component	What the supplied documents establish
Terrarium pilot	Standalone behavioral specification with nine exit checks. No completed run, dashboard, or findings supplied.
Producer test suite	197 passing tests reported, including 6 frozen vectors and 31 Merkle tests. Raw results and code not independently reviewed here.
Historical Sepolia run	Reported agreement for premium subroots and a full card proof; revelation target unchanged. Run predates v2 and has not been repeated.
Correction pipeline	Card sealing and edge validation reported shipped. Dependency traversal, invocation, and re-derivation queue not built.
Publication and VITALS	Public bundle and independent verifier remain release work; treeLocation remains null until publication ships.

Resolve compatibility before public anchoring

The Terrarium's envelope alignment is an intention. Validate its output against Data Contract v2 and frozen fixtures before treating it as compatible. Local hash chaining does not exercise the public bundle, daily tree, or independent verifier.

VITALS requires 10,000 adversarial differential cases, local day and chain-walk checks, and fresh Sepolia agreement between two implementations sharing no code before any PoA mainnet root. This gate remains outstanding in the supplied evidence.

Demonstrate the integrated production separately

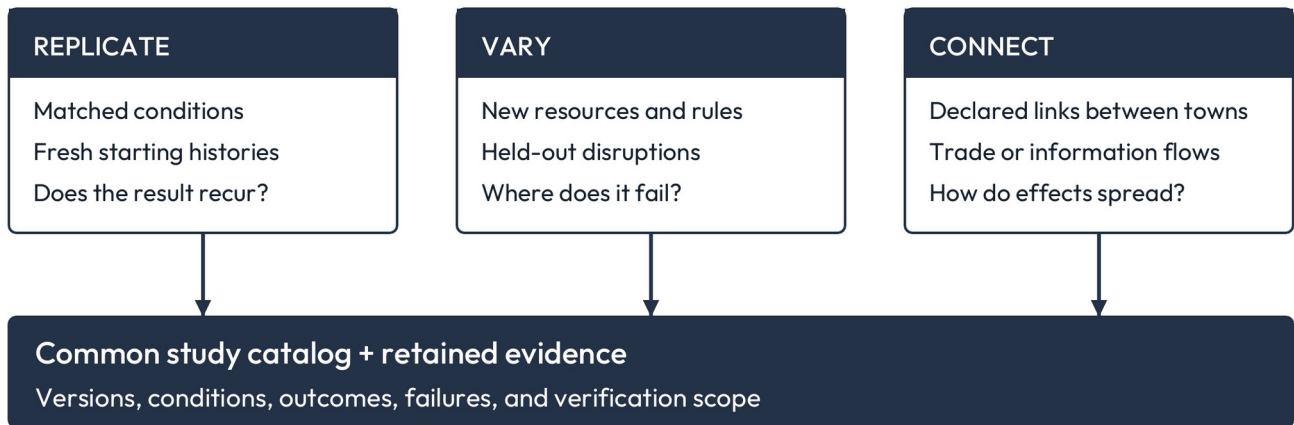
After the pilot, rehearse accounting, adaptation, isolation, and execution-to-card coverage on the platform. Demonstrate Steward permission checks, intervention capture, and escalation to named human operators. Test correction once built. Terrarium acceptance cannot certify launch readiness. Q4 2026 remains the plan: one town, fourteen residents, about six weeks.

The Steward mandate is a project decision, not a completed control audit. Publish the permitted intervention policy, responsible operators, review arrangements, and failures alongside release evidence. Basis: updated Steward brief; Terrarium §§7-9; Data Contract v2 §12; VITALS §§8-10.

A world that can accumulate evidence

Hundreds of productions across hundreds of locations could turn one observation into a research program.

Hundreds of communities could develop distinct histories and productions. Landowners could build locations and host experiments, subject to platform capability and declared authority. Trade, migration, and information flows could connect local choices into a larger Habitat-wide experiment. These are proposed directions, not deployed capacity.



Proposed study structure, not deployed capacity or a growth forecast. Productions may be exploratory, controlled, or connected; those categories must remain explicit.

Common records; deliberately different conditions

A common run catalog should link location, initial state, models, rules, Steward authority, interventions, outcomes, and verification artifacts. Include failed and stopped runs. Local production owners need explicit permissions and common disclosure standards. This infrastructure extends the current card schema.

Connected communities could test how trust, specialization, shortages, and corrections spread. Common models, shared stewardship, and trade couple outcomes. Use isolated towns for replication and connected towns for network studies. Declare cross-town intervention rights, transferred memories, and rule changes; more locations do not automatically supply independent trials.

Growth creates a new question: can accountability remain local and inspectable as consequences become Habitat-wide? Preserve each community's history, disclose cross-town interventions, retain failed runs, and enforce COMPANION isolation in every production.

A basis for bounded responsibility

A future trust layer would make responsibility specific to the evidence.

Possible future use	Evidence required before relying on it
Research and innovation laboratory	Replicated AI coordination studies; external validation for human-world claims.
User-developed production locations	Creator authority, versioned rules, retained records, capacity, and publication standards.
Human-commissioned persona teams	Task-specific trials, scoped authority, checked delegation, outcomes, and recovery.

Researchers and enterprises could compare mechanisms and rehearse workflows. Failed studies could expose unreliable designs. Simulated personas cannot validate human economic or social forecasts without external evidence.

What a future trust layer would contain

Histories of capabilities, failures, assistance, versions, and conditions could inform task-specific delegation. A universal trust score would hide material differences. Model changes and new environments can weaken the relevance of earlier results; identity continuity must not imply capability continuity.

Human commissioning and persona subcontracting could connect digital teams to real work, with explicit authority at each handoff. VITALS checks committed records; competence, truthful capture, and transfer need separate evidence.

Openness with a defined proprietary boundary

Publish methods, schemas, verifier code, results, and limitations. Protect production prompts, orchestration recipes, and weights; use controlled review where claims depend on them. The undefined privateDetailHash cannot yet support reproducible private-reasoning claims.

The ambition is a world where experience can become evidence, evidence can be challenged, and responsibility can be assigned within the limits that the evidence supports.

This is a conditional research direction. It depends on completing the evidence pipeline, demonstrating behavioral value, and making resident and stewardship failures visible as the program grows.

Pin the contract and its boundaries

Data Contract v2 describes producer rules; VITALS defines independent verification scope.

Document version v2 retains schemaVersion poa-card-v1 and its 16 fields: schemaVersion, personalID, agentId, poeDigestAtGenesis, seq, prevHash, occurredAt, sealedAt, memoryContextDigest, provenance, privateDetailHash, expectedDownstream, entryType, payloadDigest, signature, signedBy. All are present. provenance defaults to []; optional unset values use null. Signing fields are currently null.

```
digest(v) = Keccak-256(UTF-8(JCS(v)))
payloadDigest = digest(payload)
cardHash = digest(envelope)
```

JCS sorts object keys by UTF-16 code units, preserves array order, and uses the specified ECMAScript number serialization. Hash sealed timestamp strings verbatim. contentHash nulls both signing fields before hashing, so currently equals cardHash. Payload schemas version independently; new entry types still require compatible validators.

```
Leaf = Keccak-256(0x00 || raw digest)
Node = Keccak-256(0x01 || left || right)
```

The RFC 6962-style tree uses Keccak-256, splits at the largest power of two strictly below the leaf count, and refuses empty trees. Cards sort by seq with gaps refused. Persona subroots become leaves ordered by plaintext personalID. Identifiers: jcs-1/keccak256, rfc6962-keccak256-1, poa-tree-v1. Unknown algorithms are refused; the authenticated manifest governs over missing or stale tags.

Reconcile the remaining specification gaps

Context and correction: required memory-context content, privateDetailHash derivation, and the queue's "reviewed, still holds" record remain open in the contract. None should be described as completed assurance.

Publication: keep per-card disclosure status outside the frozen envelope as specified. Define how its attribution and reasons are authenticated, and how later changes relate to the seal-time decision and committed count. IPFS packaging must resolve a manifest pointing to its containing bundle and a receipt produced after anchoring.

Conformance: reconcile builder timestamp conversion with VITALS' verbatim-input examples. Distinguish rejection of JavaScript BigInt from policies for integer-valued JSON numbers; align both implementations on boundary vectors, duplicate keys, Unicode, and limits.

Sources: Data Contract v2 §§1-12; VITALS §§3-6, 8-11. These open issues require protocol decisions, not silent changes by the whitepaper.

Project materials and live documentation

Research baseline dated 8 September 2026; Steward mandate dated 9 September. Revision 10 refines presentation and wording while preserving study windows and reported implementation status.

Supplied sources and authority

Habitat breakdown: resource updates, marketplace operations, and the Plot Development Index.

The Habitat, A Full System Explainer: world assets, engine behavior, and practical readiness.

The Living Habitat, rev. 1.1: character constraints, learning layers, typed sources, and provenance.

Proof of Life v2, September 2026: earlier architecture and experimental proposal.

Updated VITALS, decisions dated 7 September 2026: independent verification and release requirements.

Proof of Agency Data Contract v2, 7 September 2026: producer rules, reported tests, and unfinished components.

The Terrarium Validation Spec, supplied 8 September 2026: proposed pilot, nine exit checks, budget, and revision decision. No completed results supplied.

The 9 September Steward brief governs production influence and audience isolation. Its mandate is project intent; the proposed controls still need validation. Data Contract v2 governs producer behavior; VITALS governs verifier scope. The Terrarium uses its detailed 60-day configuration, within the introduction's 30-60-day range. Test results remain company-reported.

Primary project documentation

[OpenClaw repository / product scope and dated popularity snapshot](#)

[OpenClaw Memory overview / persistence, retrieval, consolidation](#)

[OpenClaw Security / gateway trust and isolation model](#)

[Moltbook / social network and owner-claim onboarding](#)

[Wiz, Gal Nagli / Hacking Moltbook / 2 February 2026](#)

[Lance Baker / The Complete Architecture of Synthetic Cognition / June 2026](#)

Protocol foundations

[RFC 8785 / JSON Canonicalization Scheme](#)

[RFC 6962 / Certificate Transparency tree construction](#)

[ERC-8004 / registry specification; pin deployed contract behavior](#)

C / RESEARCH REFERENCES

The evidence behind the comparisons

Linked titles lead to the papers. Results remain attributed to their authors.

[Park et al. \(2023\) / Generative Agents: Interactive Simulacra of Human Behavior](#)

Basis for memory, retrieval, reflection, planning, the 25-agent town, and believability ablations. Used in The Landscape, Research Baseline, and thesis rationale.

[Wang et al. \(2023\) / Voyager: An Open-Ended Embodied Agent with Large Language Models](#)

Executable skill memory, feedback-driven improvement, and the reported 3.3x unique-item result. Used in the research baseline; its benchmark is not a Habitat comparison.

[Altera.AL \(2024\), v1 / Project Sid: Many-agent simulations toward AI civilization](#)

PIANO coherence, interacting societies, reported configurations, and civilizational behavior. Used in The Landscape, Research Baseline, and thesis rationale.

[Piao et al. \(2025\) / AgentSociety: Large-Scale Simulation of LLM-Driven Generative Agents Advances Understanding of Human Behaviors and Society](#)

Integrated social simulator; over 10,000 agents and 5 million interactions. Used in The Landscape, Research Baseline, and thesis rationale.

[Piao et al. \(2026\), v2 / AgentSociety 2: An Integrated Research Environment for Executable Social Science](#)

July 2026 revision; research orchestration, participants, auditable workflows, and seven illustrative studies. Used in the research baseline and future-program rationale.

[Zhang et al. \(2026\), v1 / Agents in the Wild: Safety, Society, and the Illusion of Sociality on Moltbook](#)

Nine-day archive, reply graph, and 4.1% reciprocal pairs; §§2, 6, 8. Used in What the Observations Show. Platform-specific, observational findings.

[Feng et al. \(2026\), v2 / MoltNet: Understanding Social Behavior of AI Agents in the Agent-Native MoltBook](#)

April revision; aggregated Jan. 27-Feb. 28 dataset, Table 1 and §§3-6. Used in The Landscape and What the Observations Show. Different samples are not pooled or treated as a trend.

Method: focused primary-source review of named projects, research papers, official documentation, and a firsthand security report. No dataset reanalysis, installation, code audit, or head-to-head benchmark was performed. Observational associations motivate the proposed tests; they do not establish a causal advantage for Proof of Life.