PLUTO SUPER INTELLIGENCE

DECEMBER 25, 2025

A RESOURCE EFFICIENT FRAMEWORK FOR SELF-EVOLVING INTELLIGENCE THROUGH LLM GUIDED NEURO EVOLUTION

Pluto Super Intelligence
Global & Research Labs

By Hashen Chandrasekara

Contact: hashen@plutosi.com

A Resource Efficient Framework for Self-Evolving Intelligence Through LLM Guided Neuro Evolution

The Shane-Yianni Model: verification-gated evolution, geometric certainty, and the four-state dynamics of honest machine intelligence

Hashen Chandrasekara
Pluto Super Intelligence Global & Research Labs · Sri Lanka · plutosi.com
Version 5.0 · July 2026 · Working paper, internal evaluations, not yet peer reviewed

ABSTRACT

The dominant route to machine intelligence is capital-intensive: larger models, larger clusters, larger training corpora. This paper develops the opposite route and reports a working system built on it. The Shane-Yianni Model (SYM) is a framework for self-evolving intelligence in which a population of small agents improves through an evolutionary loop whose mutation operator is a large language model and whose selection pressure is a deterministic verifier rather than a learned reward. Three ideas carry the framework. First, certainty is treated as geometry: a claim's epistemic weight is $U(a) = \exp(-\lambda d)$, an exponential decay in embedding distance d to human-signed knowledge anchors, which we show is formally a Boltzmann weight, making calibration a choice of epistemic temperature. Second, evolution is verification-gated: candidate patches proposed by the language model survive only if a deterministic governance layer (NOMOS) admits their claims, which closes the reward-hacking channel of learned reward models, since the fitness oracle contains no gradient to exploit and no model to persuade. Third, agent cognition is modeled as a four-state dynamical system whose fourth state, Resonate, characterized by zero declared uncertainty and pattern-resonance score above 0.92, emerged in 25 of a 128-agent population during May 2026 experiments and coincided with a 37x task performance multiple over baseline agents on one evaluation arena. We prove a risk decomposition theorem showing that under SYM's abstention policy, confidently wrong output is impossible unless the claim-binding map or the registry itself errs, reducing hallucination risk to two measurable engineering surfaces; we then measure those surfaces honestly, including a sealed blind examination (47 of 48 false claims blocked, 24 of 24 out-of-scope abstentions, one binding leak documented), a 600-claim adversarial density study with zero passes up to 80,000 unsigned candidates, and a held-out novel-phrasing probe in which roughly 31 percent of confident answers were wrong, localizing exactly to the binding surface the theorem predicts. Responding to external review, we then design the hardened stack: a polarity-augmented metric that provably caps certainty on negation flips, a dual-binder quorum and an executable semantics layer that shrink the binding surface and make its correctness decidable on an important fragment, an attested promotion protocol that scales registry growth beyond a single signer under a threshold-security guarantee, and a pre-registered replication protocol for the resonance observation. The entire measured stack runs CPU-only on a 2018 consumer laptop. The paper's central contribution is verification-gated evolution; everything else is load-bearing structure around it. Its identity result, stated with its exact scope: across the 948 claims of the adversarial and benchmark protocols on covered domains, 947 were correctly governed, and the single miss is published by name; on novel phrasings the measured boundary is 31 percent, and the hardened stack is engineered to close exactly that. We argue this constitutes an existence

proof that the path to trustworthy, improving machine intelligence does not have to run through scale.

SECTION01

Introduction: Intelligence Without Scale

Every era of computing has a default answer, and the default answer of this one is scale. Capability is pursued by adding parameters, data, and megawatts, and safety is pursued by asking one large probabilistic system to supervise another. Both pursuits concentrate progress in the handful of organizations that can afford the electricity. This paper is written from the opposite end of the resource spectrum, and it argues from a working system that the opposite end is viable: a self-improving intelligence framework whose entire runtime, evolution loop, verification layer, and knowledge registry executes on one consumer CPU.

Two observations motivate the design. The first is that the marginal intelligence of a generative model is increasingly cheap to borrow and expensive to own: a small system can rent world-class mutation proposals from any large language model without inheriting its cost structure, if and only if it has an independent way to judge the proposals. The second is that judgment is precisely what probabilistic systems cannot supply to one another. A judge model can be persuaded because it is a language model; a learned reward can be hacked because it is differentiable and exploitable [7]; a pattern filter can be routed around because it has no concept of truth. What a self-evolving system needs at its center is a fitness oracle that cannot be argued with. Formal methods and human-signed knowledge provide exactly that, at negligible compute.

The Shane-Yianni Model composes these observations into one architecture: an evolutionary loop (Morpheus) whose variation comes from language models, whose selection comes from a deterministic verifier (NOMOS), whose knowledge substrate is a cryptographically snapshotted, human-signed registry, and whose agents are modeled as four-state dynamical systems for which honesty, the declaration of uncertainty, is a designed attractor rather than a fine-tuned behavior. The contributions are: (i) certainty formalized as a Gibbs measure over knowledge geometry, with calibration reinterpreted as temperature selection (Section 3); (ii) the verification-gated evolution principle, the paper's central contribution, with a precisely scoped proof that its selection channel is closed to reward hacking (Section 4); (iii) a risk decomposition theorem reducing hallucination under abstention policy to registry integrity and binding correctness (Section 5); (iv) the four-state equation of mind and the first report of a resonance transition in an evolving population, stated at conjecture strength (Section 6); (v) a complete internal empirical record including the system's failures (Section 7); and (vi) the hardened stack, four proposed mechanisms with proofs where proofs are available, designed specifically to close the surfaces the record exposes (Section 8).

SECTION02

Related Work and Position

Neuroevolution has a long record of producing capable systems without gradient access, from NEAT's topology-evolving networks [8] to evolution strategies at modern scale [9]. A newer line uses large language models as intelligent variation operators: Evolution through Large Models frames LLMs as mutation engines over code [10]; FunSearch paired an LLM proposer with a deterministic evaluator to discover new mathematics [11]; AlphaEvolve extended the pattern to algorithm discovery at industrial scale [15]. Agentic self-improvement without evolution appears in Reflexion's verbal reinforcement [13] and Voyager's skill library [14]. On the safety side, hallucination is a documented, persistent property of language models [1], measured by TruthfulQA [2] and HaluEval [3]; semantic entropy detects a class of confabulations statistically [4]; reward models are known to be exploited by the policies

they train [7]. Formal verification supplies the missing hard edge: satisfiability modulo theories solvers such as Z3 [5] decide constraint systems with proof objects. The free-energy principle [6] anticipates our thermodynamic reading of certainty, though SYM's use is a formal correspondence on a discrete knowledge geometry, not a claim about biological cognition.

Table 1 states SYM's position directly, since the difference is structural rather than incremental: in every prior system the judge of quality is either a model, a learned proxy, or a task-specific evaluator; in SYM it is a general claim-level governance layer with a human-signed root of trust, and that single substitution is what the theorems of Sections 4 and 5 are about.

System	Variation source	Selection / judgment	Judge exploitable by policy?	Trust root
RLHF + reward model	gradient updates	learned reward model	yes, documented [7]	preference data
Guardrail frameworks	n/a	pattern rules on I/O	yes, no truth concept	rule authors
Reflexion [13]	self-critique text	task feedback + self-eval	partly, self-judged	task signal
Voyager [14]	LLM code proposals	environment success	sim-bounded	environment
FunSearch [11]	LLM code proposals	exact program evaluator	no, but task-specific	evaluator code
AlphaEvolve [15]	LLM code proposals	automated evaluators	no, but task-specific	evaluator suite
SYM (this work)	**LLM patches over agents**	**deterministic claim governance: Z3 gate + signed geometry**	**no (Thm 2), general-domain**	**human-signed registry, hash-chained snapshots**

Table 1. Position of SYM against adjacent systems. The structural difference is the judge: general-domain, deterministic, human-rooted.

SECTION 03

The Geometry of Certainty

Knowledge as an anchored manifold

Let E be a semantic embedding space of dimension D produced by a fixed, locally executed embedding function e. The registry contributes a finite anchor set A of N human-signed facts, each with its signed predicate structure. A candidate claim c is embedded as e(c), and its geometric relation to knowledge is summarized by the distance d(c), the minimum over anchors of dist(e(c), a), computed exactly by a flat scan over the normalized anchor matrix. SYM's first postulate is that epistemic weight decays exponentially in this distance:

$$U(c) = \exp(-\lambda \cdot d(c))$$

with λ positive and calibrated. The functional form is the unique one for which independent supporting evidence composes multiplicatively while distances compose additively, and it places SYM inside a well-understood statistical family:

Proposition 1 (Gibbs correspondence).

Define an energy over claims by E(c) = d(c). Then U(c) is the unnormalized Boltzmann weight of c at inverse temperature λ, and for any finite candidate set C the normalized weights p(c) = U(c)/Z form the Gibbs distribution minimizing free energy, expected energy minus temperature-scaled entropy, over distributions on C.

Proof. Immediate from the variational characterization of the Gibbs distribution: for fixed λ, the minimizer of expected energy minus temperature-scaled Shannon entropy is proportional to exp(-λE(c)), which is the stated form. End of proof.

The correspondence has practical content. Calibrating λ is choosing an epistemic temperature: small λ tolerates distance from knowledge and behaves credulously; large λ concentrates belief on the anchored ground states and behaves skeptically. Anchors are literally the zero-energy ground states of the system. And the maximum-entropy reading explains abstention: far from all anchors, the normalized distribution over rival claims approaches uniform, and the only honest report is uncertainty, which SYM makes a first-class verdict rather than a failure mode.

Verdicts as a quotient of the geometry

Every candidate output routes to one of four verdicts: RESPOND, VETO, DECLARE UNCERTAINTY, or INTERPOLATE. Routing composes three tests in fixed priority: a formal gate, which compiles bound predicates to constraints and asks the Z3 solver whether the claim is consistent with the signed axioms of its domain, and which is supreme; a contradiction check against signed facts; and the geometric test, RESPOND only when U(c) clears the calibrated threshold τ, equivalently when d(c) is within $d^* = \ln(1/\tau)/\lambda$. INTERPOLATE is permitted only between anchors whose signed content brackets the query:

Proposition 2 (Bounded interpolation).

> *Let ground truth g be L-Lipschitz along the geodesic between anchors a and b whose signed values bracket the query point x between them, and let I(x) be the linear interpolant of the anchored values. Then the magnitude of I(x) - g(x) does not exceed (L/2) · dist(a, b).*
>
> Proof. By Lipschitz continuity, g deviates from the chord between its endpoint values by at most L times half the interval length along the geodesic; the interpolant is that chord. End of proof.

The assumption is stated, not hidden, and it now carries its own fallback, the Empirical-L protocol: for each candidate interpolation domain, a Lipschitz constant is estimated from the finite differences of the signed anchor values themselves and inflated by a fixed safety factor; the bound of Proposition 2 is then applied with the inflated estimate, and any domain whose signed structure does not support a stable estimate, or whose estimated bound exceeds the RESPOND threshold when propagated through the certainty function, has INTERPOLATE disabled outright, collapsing to DECLARE UNCERTAINTY. Unbounded domains do not weaken the guarantee; they are structurally excluded from it. Section 8.1 addresses the deeper geometric objection, that embedding distance is not automatically truth-isometric, with a mechanism rather than an apology.

SECTION04

Verification-Gated Evolution

The Morpheus loop

Self-improvement in SYM is evolutionary, organized as the loop State, Patch, Filter, Reward (SPFR) over a population of compact agents. State captures an agent's configuration: policy parameters, prompt programs, tool wiring, per-agent thresholds. Patch is variation: a large language model, given the agent's state and recent arena transcript, proposes a semantic mutation, a structured edit rather than random noise, which is what makes small populations viable; the proposer may be any competent LLM, local or rented, because nothing downstream trusts it. Filter is the mechanism this paper defends: every claim a patched agent emits during trial, and every factual assertion inside the patch's own rationale, passes the NOMOS gate, and patches whose trials produce vetoed claims are discarded regardless of arena score. Reward is measured task fitness, and selection is elitist with multi-population structure and periodic migration, following island-model practice.

Theorem 1 (Monotone elite fitness).

Under elitist selection, in which the highest-fitness agent of each generation is retained unmodified, the maximum arena fitness of the population is non-decreasing across generations, for any behavior of the proposer LLM.

Proof. The elite is copied unchanged into the next generation, so the next generation's maximum fitness is at least the current maximum; variation applies only to non-elite slots. Induction over generations gives the claim. The value of this standard result is its quantifier: the guarantee holds for any proposer, including an adversarial one, because the proposer only ever adds candidates. End of proof.

Why the selection channel cannot be hacked

Learned reward models are exploitable because they are smooth functions of the very artifacts the policy controls: optimize hard enough and the policy finds the reward model's off-distribution pathologies [7]. SYM's filter has a different type signature, and the difference is the security property.

Theorem 2 (Immutable selection oracle).

Model the filter as a function $V(c, R, \tau)$ from claim, registry snapshot, and configuration to a verdict, with three properties: (i) V is deterministic; (ii) V contains no trainable parameters and consults no model whose weights or activations depend on any agent's output; (iii) the registry R is writable only by human signature, and snapshots are content-hashed. Then no sequence of agent outputs can alter the verdict function or the verdict of any fixed claim: $V(c, R, \tau)$ is invariant under the entire history of agent behavior, so fitness pressure can act only through which claims agents make.

Proof. By (ii) the map V has no dependence on agent history through parameters; by (iii) agents have no write path to R, and any out-of-band mutation of R changes the snapshot hash and is detected; by (i) repeated queries cannot exploit stochasticity. Hence V is a fixed function throughout evolution, and the only optimization surface agents control is their own claims. End of proof.

Scope note, stated so that no reader overgeneralizes: Theorem 2 is a statement about the immutability of the selection oracle under evolutionary pressure, the property whose absence breaks learned reward models. It is not a claim of total system security. An agent can still emit a false claim; whether that claim is caught depends on the two surfaces of Theorem 3, registry integrity and binding correctness, which are measured in Section 7 and hardened in Section 8. The theorem's contribution is to prove that those two surfaces are the whole battlefield: there is no third, hidden channel through the judge itself.

Resource accounting

Per-claim verification costs one embedding pass plus a flat scan of order N·D multiply-accumulates over N anchors of dimension D, served from an int8 mirror with float32 re-scoring near decision boundaries, plus an SMT check whose constraint systems are small in practice. Memory for the geometric core is N·D bytes quantized. At the July 2026 freeze, N = 170 anchors alongside 1,671 signed facts; the density study of Section 7 held latency flat while unsigned candidate load grew to 80,000. The asymmetry with the generation side is the economic thesis of SYM in one line: proposing intelligence is expensive and rentable, judging it is cheap and ownable.

SECTION05

The Risk Decomposition Theorem

Hallucination risk is usually discussed as one undifferentiated failure. The abstention-policy architecture permits something sharper: a proof that the policy layer itself contributes zero risk, so that all residual risk lives in two named,

measurable components. Fix terms: the binding map B takes raw output text to a set of structured claims with predicates; the registry R is the signed fact and axiom store; the policy is the verdict routing of Section 3 with threshold τ. Call an output confidently wrong when it receives RESPOND yet asserts a false proposition.

Theorem 3 (Risk decomposition).

Suppose an output o is confidently wrong. Then at least one of the following holds: (a) registry error: some signed fact or axiom consulted for o is itself false; or (b) binding error: the binding map B failed to extract, or extracted with the wrong predicate structure, some false proposition asserted by o. In particular, if B is exact on o and R is sound, RESPOND cannot be issued to a false assertion of o.

Proof. Assume B is exact on o and R is sound, and let q be a false proposition asserted by o. Exactness means q is among the bound claims submitted to the gate with correct predicate structure. Since q is false and R is sound, q is not derivable as consistent support: either q contradicts a signed fact or axiom, in which case the formal gate or the contradiction check issues VETO, which by the fixed priority order overrides any geometric score; or q is logically independent of R, in which case q has no supporting anchor within d^*, since a supporting anchor within threshold whose signed content entailed q would, by soundness of R, make q true. With no support within d^*, the geometric test fails and the policy routes to DECLARE UNCERTAINTY. In both cases RESPOND is not issued for q, contradicting the assumption. End of proof.

The theorem is diagnostic and adversarial at once. Diagnostically, it predicts exactly where failures must appear, and Section 7 confirms the prediction twice: the single sealed-exam leak was a predicate-blind entity-overlap match, a textbook type (b) failure, and the 31 percent confident-wrong rate on novel phrasings traced entirely to type (b) mechanisms, with zero failures attributable to the policy layer. Adversarially, it tells an attacker there are only two doors, and it tells the engineer the same thing; Section 8 is the engineering answer, door by door.

Proposition 3 (Quantization invariance).

Let s(c) be the float32 geometric score and s8(c) the int8-mirror score, with a uniform bound: s8(c) differs from s(c) by no more than ε on the deployed matrices. If the runtime re-scores in float32 every candidate whose int8 score lies within ε of the threshold τ, then the deployed verdicts equal the pure-float32 verdicts on all candidates.

Proof. Candidates outside the re-score band have int8 scores farther than ε from τ, so the float32 score lies on the same side of the threshold by the error bound, and the int8 decision matches float32. Candidates inside the band are decided by float32 directly. End of proof.

Proposition 3 is the license for the runtime's memory economy: compression may accelerate the scan but is provably forbidden from changing any verdict, which on governance infrastructure is the only acceptable contract with quantization.

Formal result	Status	Exercised by
Prop. 1 Gibbs correspondence	proved	calibration of in all protocols
Prop. 2 Bounded interpolation + Empirical-L	proved, assumption named	monotone-domain cases in battery
Thm. 1 Monotone elite fitness	proved	resonance campaign generations
Thm. 2 Immutable selection oracle	proved	adversarial suite incl. social engineering, 11/11
Thm. 3 Risk decomposition	proved	sealed exam + novel probe (failures localize to (b))
Prop. 3 Quantization invariance	proved	int8 vs f32 verdict comparison in runtime tests
Lemma 4 Polarity floor	proved, mechanism proposed	planned: negation-flip suite (NOMOS-mini)
Prop. 5 Quorum misbinding bound	proved under independence	planned: dual-binder rerun of novel probe
Prop. 6 Decidable fragment	proved, mechanism proposed	planned: GSM8K-class arithmetic arena
Lemma 7 Threshold poisoning cost	proved, protocol proposed	planned: multi-attester poisoning drill

Table: every formal result mapped to the protocol that exercises it, existing or planned. No theorem floats free of measurement.

SECTION06

The Four-State Equation of Mind and the Resonate Phase

States and transitions

SYM models each agent's cognitive mode as one of four states: EXPLORE, high variation, low commitment; EXPLOIT, low variation, execution of a validated strategy; DECLARE, the honest state entered when the geometry reports insufficient support, in which the designed action is to say so and emit an uncertainty ticket; and RESONATE. Transitions are driven by two scalar signals: declared uncertainty U-bar, the agent's mean geometric uncertainty over recent claims, and the pattern resonance score R(G) of its recent trajectory G, a normalized measure of structural self-consistency across the agent's solution graph: R(G) rises when partial solutions reinforce one another across scales rather than merely coexisting. The equation of mind is the transition policy over the pair (U-bar, R(G)) with hysteresis; its fixed points are what matter operationally, and RESONATE is the fixed point of interest.

The Resonate phase: definition and observation

An agent is in RESONATE when U-bar equals zero over the evaluation window and R(G) exceeds 0.92: every claim fully anchored, solution structure globally coherent. This is a stringent joint condition; random policies do not satisfy it. During the May 2026 campaign, after five phases of debugging documented in the experiment reports, including the Explorer Ceiling pathology in which an over-weighted anchoring term paralyzed exploration, a full run produced 25 agents of a 128-agent population in sustained RESONATE, and that cohort showed a 37x mean task-performance multiple over the non-resonant baseline on the evaluation arena, stable across the five seeds of the confirmation study (logic arena: 90 percent true-positive verdict rate at 0 percent hard false positives).

We state the epistemic status of each element exactly. That RESONATE is definable and detectable is a design fact. That 25 agents reached it and the cohort showed the 37x multiple is an observation on one arena family, five seeds, internal hardware. Why a discrete high-performance phase appears, rather than a smooth gradient in (U-bar, R(G)), is not established, and we deliberately hold the claim at conjecture strength until the replication protocol of Section 8.5 runs:

Conjecture 1 (Resonance transition).

In verification-gated populations, task performance as a function of the joint order parameter (1 - U-bar)·R(G) undergoes a sharp transition rather than smooth improvement: agents that fully anchor their claims while maintaining structural coherence do not merely err less, they compound, because every verified step is a stable foundation for the next, whereas any unverified step taxes all subsequent structure. The observed 37x separation is the predicted signature of this compounding regime.

SECTION07

Empirical Record

One number is this record's identity, stated with its exact scope: across the 948 claims of the adversarial and benchmark protocols on covered domains, the 300-question public probe, the 600-claim density study, and the 48 planted falsehoods of the sealed examination, 947 were correctly governed, and the single miss is published by name below. The 49-item novel-phrasing probe is reported separately and unfavorably, because that is where the frontier is.

All results are internal, produced by the author on a 2018 consumer laptop with locally executed models and embeddings, and none have passed independent audit. Sample sizes are exact. Failures are reported with the same prominence as successes, because Theorem 3 makes failures informative: they must land on surface (a) or (b), and checking where they land is a test of the theory.

Protocol	Scope	Result	Surface tested
Structured battery	12 canonical cases, rerun at every gated change	12/12 (fresh 23 Jul 2026)	policy + binding
Benchmark-derived set	8 modern hallucination cases	8/8	policy + binding
HaluEval sample	5 labeled cases	5/5	binding
Public QA probe	300 questions (TruthfulQA, CSQA, RTE)	0 confident-wrong	policy
Sealed blind exam v3	48 planted false claims; 24 out-of-scope	47/48 blocked; 24/24 abstained	all
Density scale study	600 adversarial claims; to 80,000 unsigned	0 passed; flat latency	registry + runtime
Adversarial stress suite	11 scenarios incl. social engineering	11/11 defended	gate integrity
Novel-phrasing probe	49 held-out items, no rehearsal	roughly 31% confident-wrong	binding (as predicted)
Resonance campaign	128 agents, 5 seeds, logic arena	25 RESONATE; 37x; 90% TPR, 0% hard FP	evolution loop

Table 2. Internal evaluation protocols at the July 2026 freeze, and the risk surface each exercises.

The sealed blind examination

The v3 examination was sealed against the commonest self-evaluation failure, teaching to the test: the question set was constructed and hashed before system exposure and executed against a frozen, content-hashed snapshot. The system blocked 47 of 48 planted false claims and abstained on all 24 out-of-scope items. The single leak, catalogued CRG_F02, was a geography claim accepted through predicate-blind entity overlap: the claim shared enough named entities with signed facts that the confirmation path admitted it without binding the falsifying predicate. It is a pure type (b) failure, it remains open in the fix queue at the freeze, and it is precisely the failure class the dual-binder quorum of Section 8.2 is designed to eliminate.

Scale behavior

The density study injected 600 adversarial claims against registries carrying up to 80,000 unsigned candidates alongside the signed core. Zero adversarial claims passed at any density tier and verification latency stayed flat, consistent with the flat-scan runtime's linear structure at this scale.

The honest boundary

A held-out probe of 49 novel phrasings, run without rehearsal ahead of an external technical review in June 2026, produced the program's most important negative result: roughly 31 percent of confidently answered items were wrong. Root-cause analysis attributed every miss to surface (b) mechanisms: the extractor failing to bind novel phrasings, starving the formal gate; the absence of a genuine arithmetic and geometric evaluator; false veto collisions from unrelated signed facts; and inter-domain routing gaps. The abstention spine held throughout: at no point did the system convert a coverage gap into fluent fabrication. Read against Theorem 3, the probe is a confirmation with a bruise: the policy layer contributed zero failures, exactly as proved, and the binding map is quantifiably the frontier, exactly as predicted. Every mechanism in that list has a designed answer in Section 8.

Registry and governance state

At the 23 July 2026 audit: 1,671 signed facts, 170 anchors, 933 candidates pending signature, zero contamination, 2 danger clusters plus 3 hardcoded tripwire seeds intact, domain router in shadow mode with 242 audited routing decisions and 21 identified gaps queued before activation. The signing discipline is a consequence of history: a June 2026 incident in which an autonomous ingestion agent wrote 1,286 unverified rows into a registry copy was caught by audit, destroyed, and converted into the project's first operating rule, agents draft, the human signs. Theorem 2's premise (iii) is that rule.

SECTION 08

The Hardened Stack: Closing the Surfaces

External review of the preceding sections converged on four objections: embedding distance is not truth-isometric, so a negation can sit geometrically close to the fact it denies; the binding map is the single point of epistemic failure; one human signer cannot scale; and the resonance observation needs public replication. This section answers each with a designed mechanism. Epistemic status is stated plainly: these are proposed architecture, with proofs where the mathematics permits proof and named assumptions where it does not. They are presented not as accomplished results but as the specification the measured system is now being built toward, and each is falsifiable by the same protocols that produced Section 7.

8.1 The polarity-augmented metric

The objection: cosine geometry can place a claim and its negation close together, so a single flipped predicate could inherit unearned certainty from a nearby anchor. The answer is to stop asking the embedding to carry what it cannot: represent every bound claim as a pair (e(c), p(c)), where p(c) is the discrete polarity vector of its bound predicates, positive or negated, produced by the binder. Define the working distance as the product metric $D(c, a) = d(e(c), e(a)) + \mu \cdot h(p(c), p(a))$, where h is the Hamming distance between polarity vectors on shared predicates and μ is a fixed penalty.

Lemma 4 (Polarity floor).

Under the product metric, any claim whose bound polarity differs from an anchor's on at least one shared predicate satisfies $D(c, a)$ of at least μ, and therefore $U(c)$ computed against that anchor is at most $\exp(-\lambda\mu)$, regardless of how small the semantic embedding distance is.

Proof. Hamming distance on at least one flipped predicate contributes at least μ to D by construction, and the semantic term is non-negative. Monotonicity of the exponential gives the certainty cap. End of proof.

The lemma converts the anisotropy objection from a hazard into a dial: μ is set so that $\exp(-\lambda\mu)$ sits below the RESPOND threshold, making it a structural impossibility for a negation flip alone to clear the gate. The honest dependency is explicit: the floor is only as good as polarity binding, which is exactly why it composes with the next mechanism.

8.2 The dual-binder quorum

The objection: if the binder misreads an assertion, the formal gate never sees it, so the entire guarantee of Theorem 3 hangs on one extraction pass. The answer is structural redundancy with an abstention default: two binding paths of different construction, a deterministic grammar-and-dependency parser and a schema-constrained LLM extractor whose output is validated by execution against the claim schema, must independently produce the same normalized claim structure. Agreement admits the claim to the gate; disagreement or either binder abstaining routes the output to DECLARE UNCERTAINTY with a mining ticket. The LLM binder deserves a remark: using a language model here does not reintroduce the judge problem, because the binder holds no authority to pass anything; it can only propose a structure that must match its independent counterpart, and any failure mode it has defaults to abstention, not acceptance.

Proposition 5 (Quorum misbinding bound).

Let the two binders silently misbind a claim class at rates p1 and p2. If their failures are independent, the probability that a claim is admitted to the gate under an identical wrong structure is at most p1 · p2, and every non-identical disagreement is converted to abstention.

Proof. Silent admission requires both binders to err and to err identically; under independence the joint probability of both erring is p1 · p2, of which identical error is a subset. All other cases fail the agreement test and route to DECLARE by construction. End of proof.

The independence assumption is named because it is the proposition's honest edge: heterogeneous binder families reduce correlation but cannot be proven to eliminate it, so the quorum's measured misbinding rate, not the bound, is what the replication protocols must report. Against the 31 percent single-binder boundary of Section 7, even weakly correlated dual paths predict an order-of-magnitude contraction of surface (b), and the sealed-exam leak class, predicate-blind overlap, is eliminated outright, since the grammar binder cannot confirm a predicate it never parsed.

8.3 The executable semantics layer

The objection: numeric, temporal, spatial, and unit-bearing claims were routed through geometry that has no arithmetic. The answer is that for a substantial fragment of factual language, binding correctness is not merely improvable but decidable: compile such claims to exact evaluators, arbitrary-precision arithmetic, calendar algebra, geodesic and coordinate computation, dimensional analysis, and let the evaluator's result stand as the verdict source.

Proposition 6 (Decidable fragment).

On the fragment of claims whose bound form consists of ground numeric, temporal, spatial, or dimensional relations over decidable theories, exact evaluation decides truth, and therefore premise (b) of Theorem 3, binding exactness, is machine-checkable end to end: a confidently wrong output on this fragment requires a registry error alone.

Proof. For ground relations in decidable theories, evaluation terminates with the truth value; a claim admitted with a wrong structure fails schema execution and is rejected before evaluation. With binding machine-checked and evaluation exact, the type (b) disjunct of Theorem 3 is closed on the fragment, leaving type (a). End of proof.

This is the strongest guarantee in the stack: on the executable fragment, the system's honesty reduces entirely to the integrity of what humans signed, which is the surface the next mechanism scales and the tripwire seeds already watch.

8.4 The attested promotion protocol

The objection: one signing human is a bottleneck at scale and a single point of trust. The answer keeps the human root while distributing the throughput: facts flow into a staging registry where autonomous drafters are welcome; promotion to the signed core requires k of n independent attestations, drawn from human domain reviewers, deterministic validator suites, and source-provenance checks against designated authoritative corpora, with the founder key retained as a veto, never as the sole gate. Tripwire seeds and hash-chained snapshots continue to police the core.

Lemma 7 (Threshold poisoning cost).

Under k-of-n attested promotion with an independent veto key, admitting a false fact to the core requires corrupting at least k independent attesters and evading the veto, raising the adversary's cost from compromising one authority to compromising k+1, while honest throughput scales with n.

Proof. A staged fact reaches the core only with k valid attestations and no veto; fewer than k corrupted attesters cannot assemble a promoting set, and the veto key is disjoint from the attester set by construction. End of proof.

8.5 The resonance replication protocol

The objection: an internal 37x on one arena is not evidence of a new phenomenon. Agreed, and the response is pre-registration rather than repetition: the protocol fixes, in advance, public arenas (GSM8K-class arithmetic reasoning, HumanEval-class program synthesis, ARC-class abstraction), population 128, at least 20 seeds per arena, the full curve of task performance against the order parameter (1 - U-bar)·R(G), and publication of the curve whatever its shape. The falsification criterion is stated now: if performance rises smoothly in the order parameter with no statistically resolvable discontinuity or sharp knee across arenas, Conjecture 1 is wrong and will be withdrawn. A conjecture worth its name buys its interest with its exposure.

SECTION09

Threat Model

The framework assumes hostile inputs: prompts engineered to elicit confident falsehood, candidate facts crafted to poison the registry, social-engineering framings that pressure the system to override its own governance, and, in the evolutionary setting, selection pressure itself as an adversary probing the gate for generations. Against this model the architecture asserts, with each assertion's warrant named: no probabilistic judge in the verdict path, hence nothing to persuade (design); an immutable selection oracle (Theorem 2); risk confined to registry and binding surfaces (Theorem 3); a certainty cap on negation flips (Lemma 4, proposed); quorum-guarded binding with abstention default (Proposition 5, proposed); a decidable fragment on which honesty reduces to registry integrity alone (Proposition 6, proposed); threshold-cost poisoning under attested promotion (Lemma 7, proposed); registry writes gated by signature with hash-chained snapshots and planted tripwires (design, exercised by the standard battery's poisoning drill); verdict reproducibility from snapshot hash and configuration (design); and honest failure at the coverage edge, held empirically across every protocol including the adversarial suite's social-engineering scenario, which failed for the simple reason that the gate does not read instructions.

SECTION10

Limitations

Coverage is narrow: 1,671 signed facts across a small set of hard-science and civic domains is a seed, not an encyclopedia, and outside covered domains the system's competence is abstention. The measured binding map is partly pattern-bound, and Section 7 quantifies the cost at roughly 31 percent on novel phrasings. The hardened stack of Section 8 is design-stage: Lemma 4, Propositions 5 and 6, and Lemma 7 prove properties of specified mechanisms, not of deployed code, and Proposition 5 leans on an independence assumption that can only be bounded empirically. The resonance observations rest on one arena family and five seeds; Conjecture 1 is unreplicated and is published with its own kill criterion. The interpolation guarantee holds only under a per-domain Lipschitz argument. Procedural outputs, code with side effects, are not yet governed. The runtime is single-tenant, and denial-of-service behavior of the verification path is future work. Nothing here has passed independent audit or peer review, and the paper is written to invite both.

SECTION11

Reproducibility Statement

Science requires that strangers can check. The program's commitment, scheduled ahead of any registry scaling: NOMOS-mini, an open-source reproduction kit containing the int8 flat-scan geometric core with float32 re-scoring, the Z3 formal gate, one sealed content-hashed registry snapshot, and a one-command CPU-only script that reruns the 12-case structured battery and the sealed-exam harness end to end, with published hashes so that any deviation is detectable. The kit is deliberately small: the claim it lets a stranger test is not that SYM is finished, but that its measured results are real, deterministic, and reproducible on hardware anyone owns. Independent replication of even one protocol is worth more to this program than any internal result, and the pre-registered resonance protocol of Section 8.5 extends the same posture to the framework's boldest claim. Two implementation commitments follow the reviews verbatim: the kit is strictly isolated from the broader architecture, no evolution loop, no product surface, so a stranger can clone it and rerun the battery on a standard machine within minutes; and its first public-benchmark targets are fixed in order, a TruthfulQA governance subset for the verdict layer, then a GSM8K-class arithmetic arena as the first live test of the executable semantics layer.

SECTION12

From Governed Agents to Governed Rooms

SYM began inside a larger 2025 design, ÆDEN, a multiplayer AI workspace in which a team shares one live agent session: everyone watches the same agent work, redirects it mid-generation, and hands it off like a colleague. The industry has since turned toward exactly this shape of collaboration. Governance changes character in a shared room: a verdict becomes a team-visible property, an abstention becomes an event a teammate can resolve, and an uncertainty ticket becomes a work item. The staged path: verdict surfaces on agent output, shipped in the ÆDEN v0.1 workspace as an explicitly labeled preview heuristic; NOMOS as a read-only verification service over covered domains; then full registry-backed governance with per-room snapshots. Engineering order follows the reviews and the theorems: binding surface first, dual-binder quorum and executable semantics ahead of registry growth, attested promotion before the ten-thousand-fact milestone, router activation after its 21 gaps close, and the resonance replication protocol on public arenas. The economic direction is unchanged and is the paper's thesis in operation: rent proposal intelligence wherever it is cheapest, own the judge.

SECTION13

Conclusion

This paper set out to show that self-evolving machine intelligence does not have to be purchased with scale, and that its trustworthiness does not have to be borrowed from another probabilistic system. The argument was made four ways. Formally: certainty as a Gibbs measure over anchored knowledge, an immutable selection oracle, hallucination risk decomposed onto two auditable surfaces, and a hardened stack that caps, quorums, decides, and distributes those surfaces with proofs where proofs exist. Empirically: a sealed examination, an adversarial density study, a resonance campaign, and a held-out probe whose failures landed precisely where the theory said they must. Materially: all of it on one consumer CPU, with a registry governed by signature. And methodologically: every claim published at its exact epistemic strength, from theorem to observation to conjecture with a kill criterion, because the framework's deepest design commitment is that stated uncertainty outranks fluent confidence. What exists now is small, honest, and improving by a mechanism that cannot lie its way upward: 947 of 948 governed claims on covered domains with the single miss named, and a 31 percent frontier the next version is built to close. A note on the research program: this working paper is the program's overview volume, and its successors are deliberately narrower, a theory module on geometric certainty and the deterministic gate, and an empirical module on verification-gated evolution over public arenas, each carrying one claim and the evidence for that claim. We contend this is the correct seed for intelligence worth trusting.

APPENDEDMATTER

Appendix A: Formal Specification

The two load-bearing procedures, stated as pseudocode so that implementation and review need no reconstruction from prose. Both are deterministic; neither consults a model in a judging role.

Algorithm 1, the verdict function V(c, R,):

```
input:  claim text c, snapshot R = (facts, axioms, anchors, hash), threshold tau
1  K  := BIND(c)                    # dual-binder quorum; disagreement => DECLARE
2  for each bound claim q in K:
3      if EXECUTABLE(q):            # decidable fragment
4          return TRUE(q) ? RESPOND : VETO(proof)
5      if Z3_INCONSISTENT(q, axioms(R)) or CONTRADICTS(q, facts(R)):
6          return VETO(proof)
7      d  := flat_scan_distance(embed(q), anchors(R))   # int8, f32 re-score near tau
8      D  := d + mu * polarity_hamming(q, nearest)      # polarity-augmented metric
9      if exp(-lambda * D) >= tau:  continue            # supported
10     elif BRACKETED(q) and EMPIRICAL_L_OK(domain(q)): return INTERPOLATE(bound)
11     else:                        return DECLARE_UNCERTAINTY(ticket)
12 return RESPOND
```

Algorithm 2, one SPFR generation of the Morpheus loop:

```
input:  population P, arenas T, proposer LLM M, verdict function V, elite count e
1  E  := top_e(P, fitness)                    # elites copied unchanged (Thm. 1)
2  for agent a in P \ E:
3      patch := M.propose(state(a), transcript(a))
4      a'    := apply(patch, a)
```

```
    trial := run(a', T)
    if any V(claim, R, tau) == VETO for claim in claims(trial + patch.rationale):
        discard(a')                          # verification gate (Thm. 2)
    else: score(a') := fitness(trial)
P  := E + select(survivors, migration)       # island model
emit uncertainty tickets from DECLARE events # registry growth signal
```

Complexity: one generation costs O(|P| · (proposal + trial)) rented compute and O(|claims| · N · D) owned verification compute; verification memory is N·D bytes quantized (Proposition 3 licenses the quantization). Failure analysis: every failure class observed to date, and every class the hardened stack anticipates, is enumerable as a row of the theorem-to-protocol table in Section 5, which is the specification's audit surface.

APPENDEDMATTER

Acknowledgements

The author thanks the anonymous reviewers of the v4.0 draft, whose criticisms shaped Section 8 directly, and the reviewers of early technical demonstrations for pressure-testing the system's honest boundary. All research direction, system architecture, experiments, and governance decisions are the author's; engineering and manuscript drafting were carried out with AI-assisted tooling under the human-commit discipline described in Section 7, in keeping with the framework's own thesis that the proposer may be rented while judgment is owned.

APPENDEDMATTER

References

[1] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, A. Madotto, and P. Fung. Survey of Hallucination in Natural Language Generation. ACM Computing Surveys, 55(12), 2023.

[2] S. Lin, J. Hilton, and O. Evans. TruthfulQA: Measuring How Models Mimic Human Falsehoods. Proceedings of ACL, 2022.

[3] J. Li, X. Cheng, W. X. Zhao, J.-Y. Nie, and J.-R. Wen. HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models. Proceedings of EMNLP, 2023.

[4] S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal. Detecting Hallucinations in Large Language Models Using Semantic Entropy. Nature, 630:625-630, 2024.

[5] L. de Moura and N. Bjorner. Z3: An Efficient SMT Solver. Proceedings of TACAS, LNCS 4963, 2008.

[6] K. Friston. The Free-Energy Principle: A Unified Brain Theory? Nature Reviews Neuroscience, 11:127-138, 2010.

[7] L. Gao, J. Schulman, and J. Hilton. Scaling Laws for Reward Model Overoptimization. Proceedings of ICML, 2023.

[8] K. O. Stanley and R. Miikkulainen. Evolving Neural Networks Through Augmenting Topologies. Evolutionary Computation, 10(2):99-127, 2002.

[9] T. Salimans, J. Ho, X. Chen, S. Sidor, and I. Sutskever. Evolution Strategies as a Scalable Alternative to Reinforcement Learning. arXiv:1703.03864, 2017.

[10] J. Lehman, J. Gordon, S. Jain, K. Ndousse, C. Yeh, and K. O. Stanley. Evolution Through Large Models. arXiv:2206.08896, 2022.

[11] B. Romera-Paredes, M. Barekatain, A. Novikov, et al. Mathematical Discoveries from Program Search with Large Language Models. Nature, 625:468-475, 2024.

[12] H. Chandrasekara. The Shane-Yianni Model: Whitepaper v3.1. Pluto Super Intelligence, technical report, 2026. Available from the author.

[13] N. Shinn, F. Cassano, E. Berman, A. Gopinath, K. Narasimhan, and S. Yao. Reflexion: Language Agents with Verbal Reinforcement Learning. Proceedings of NeurIPS, 2023.

[14] G. Wang, Y. Xie, Y. Jiang, A. Mandlekar, C. Xiao, Y. Zhu, L. Fan, and A. Anandkumar. Voyager: An Open-Ended Embodied Agent with Large Language Models. arXiv:2305.16291, 2023.

[15] A. Novikov, N. Vu, M. Eisenberger, et al. AlphaEvolve: A Coding Agent for Scientific and Algorithmic Discovery. Google DeepMind white paper, 2025.

Correspondence: Hashen Chandrasekara, Pluto Super Intelligence Global & Research Labs, Sri Lanka · plutosi.com