GEME: A Self-Referential Prism for Cognitive Modeling

GEME: A Self-Referential Prism for Cognitive Modeling

Jieqi Liu Independent Researcher jackey.l.gene@outlook.com

Zenodo DOI: 10.5281/zenodo.20110147

Abstract

This paper presents a philosophically-driven computational thought experiment that constructs GEME — a cognitive primitive architecture based on self-reference — as a vehicle for exploring the minimal operational principles underlying cognitive activity. GEME is built upon a single, undifferentiated self-referential feedback loop; all observed functional stratification, competitive dynamics, and adaptive behaviors arise spontaneously from this loop under sustained information input and stability constraints. We report three independent, reproducible observations: (1) a stable self-reference-to-information channel — termed the Shannon-Gödel Bridge — through which self-referential operations sustain steady-state cognitive processing at near-zero information cost, with ablation confirming that removal of the bridge structure causes a three-order-of-magnitude increase in information cost; (2) a broad economic stability interval within the parameter space, identified through exhaustive traversal of the full parameter volume, within which the system's qualitative behavior remains invariant — the three core structural parameters are thus presented not as theoretically derived constants but as experimentally selected optimal parameters, analogous to silica as the empirically optimal material for optical prisms; (3) a functional equivalence at the behavioral level between the self-referential skeleton G and the induction axiom schema PA when operating through the query interface Q — the combined function $G + Q$ produces output behavior indistinguishable from PA on the tested inductive reasoning tasks. The primary contribution is the demonstration of a methodology for experimentally identifying minimal cognitive primitives through exhaustive parameter-space traversal, together with the GEME architecture itself as a reproducible reference point for investigating self-reference as a foundational principle of cognition.

Keywords: self-reference, cognitive architecture, computational thought experiment, working memory, information economy, Shannon-Gödel bridge

---

1. Introduction

A central challenge of cognitive science is to identify the minimal primitives capable of supporting complete cognitive function, and to understand the operational principles governing them. This challenge is rooted in two apparently distinct foundations: the physical constraints of information transmission (Shannon, 1948)

and the logical necessity of self-reference (Gödel, 1931). Yet an architecture that unifies these two and provides an operable primitive for examining their interaction has remained absent.

Related research has focused on their respective limits. Chaitin's algorithmic information theory (1975) anchors the essential information of an object in the shortest program that produces it, revealing inherent randomness boundaries within mathematical truth. Aaronson (2011, 2013) argued that many philosophical problems concerning mind and knowledge may hinge on computational complexity — the resource cost required to realize a function, rather than abstract computability. Friston's free-energy principle (2010) offers a quantitative account grounded in variational Bayesian inference: the unifying imperative for self-organizing systems is the continuous minimization of perceived prediction error, providing a dynamical perspective on perception, action, and homeostasis.

Hofstadter, in *Gödel, Escher, Bach* (1979), provided an intuitive philosophical synthesis. He argued that meaning, consciousness, and hierarchical organization arise not from external injection but as byproducts of a system referring to itself. Self-reference is the bridge connecting meaningless symbols to meaningful understanding, mechanical steps to inner experience. Hofstadter achieved a magnificent philosophical synthesis, but this synthesis remained at the level of metaphor — a runnable, testable architecture has been absent.

The present work aims to take a step from philosophical synthesis toward computational exploration. We present the Generative Economy Memory Entity (GEME) — a self-referential primitive architecture — as the carrier of a computational thought experiment. GEME does not simulate any specific cognitive module; rather, it instantiates a deep kernel shared by the aforementioned intellectual lineage: cognition may be organized through self-reference under resource constraints, maintaining structural integrity through competitive economic dynamics. We do not claim to solve the puzzle of cognition. We offer a reproducible, inspectable, extensible computational prism — through which the structural consequences of self-reference under constraint can be examined from new operational perspectives.

The core contributions of this work are threefold:

1. **An architecture based on a minimal self-referential primitive.** GEME demonstrates, through computational thought experiments, the possibility of supporting basic cognitive functions — classification, association, anticipation, self-monitoring — via a single self-referential feedback loop governed by three structural parameters.
2. **The observation of a functional equivalence between self-reference and induction ($G+Q \approx PA$).** When the self-referential skeleton G operates through a query interface Q, the combined behavior is functionally indistinguishable at the output level from the induction axiom schema PA on standardized inductive reasoning tasks, providing new clues regarding the computational nature of inductive inference.
3. **An experiment-driven methodology for cognitive primitive identification.** By exhaustively traversing the full parameter space, we identify the minimal parameter set capable of stably supporting emergent cognitive patterns — analogous to materials science screening for functional materials — offering a generalizable approach for discovering foundational cognitive building blocks.

This work is not aimed at quantitatively fitting specific cognitive experimental data or making behavioral predictions. It presents a conceptual framework intended to inform subsequent theoretical modeling and empirical investigation.

The paper is organized as follows. Section 2 describes the GEME self-referential primitive architecture. Section 3 reports the core observations from the computational thought experiment. Section 4 discusses

implications, engages with existing frameworks, and outlines limitations. Section 5 concludes.

---

2. The GEME Self-Referential Primitive Architecture

This section presents the architectural construction of the GEME self-referential primitive. The emphasis is on logical self-consistency and mechanistic clarity. At its foundation, GEME contains a single, undifferentiated self-referential feedback loop; all observed properties — stratification into functional layers, competitive merging, adaptive decay — are spontaneous behaviors that arise from this loop under sustained information input and stability constraints.

2.1 Core Feedback Loop

> **[[Figure 2.1: `fig2_1_dynamics_loop.png` — Core dynamics of the self-referential feedback loop]]**
> *Caption: The three operations — Competitive Merging, Adaptive Decay, and Self-Observation — form a closed loop. External input enters the Frame Economy, is merged or stored as a new frame, decays if not reinforced, and is periodically audited through self-observation, which feeds back as new input.*

The GEME primitive operates on a collection of frames F constrained by a fixed capacity C (default: 10). Each frame f is a data structure $(\mathbf{v}, w, a, m, \sigma, \ell)$, where $\mathbf{v}$ is a D-dimensional content vector ($D = 27$), w tracks cumulative merging frequency and determines survival probability, a counts steps since creation, m counts successful merges (affecting decay rate), σ is a short text signature for cross-layer identification (truncated to 30 characters), and ℓ records the frame's functional origin (L1–L6). Three operations constitute the self-referential feedback loop.

Operation 1: Competitive Merging. Let the frame collection at time t be $F_t = \{f_1, \ldots, f_n\}$ with $|F_t| \leq C$. Each frame $f_i = (\mathbf{v}_i, w_i, a_i, m_i, \sigma_i, \ell_i)$. For an input $(\mathbf{x}, \sigma)$, define the best-matching frame and its distance:

$$f^* = \arg\min_{f_i \in F_t} d(\mathbf{x}, \mathbf{v}_i), \quad d^* = \min_i d(\mathbf{x}, \mathbf{v}_i), \quad d(\mathbf{x}, \mathbf{v}) = \sqrt{\sum_{j=1}^{D} (x_j - v_j)^2}$$

The merge threshold δ_{eff} is adaptively calibrated rather than externally set. Formally:

$$\delta_{\text{eff}}(\mathcal{M}_t, \mathcal{L}_t) = \begin{cases} \text{None}, & |\mathcal{L}_t| < 10 \quad \text{(learning: accumulate samples)} \\ Q_{0.25}(\mathcal{L}_t), & |\mathcal{M}_t| = 0 \quad \text{(initial: lower quartile of learning distances)} \\ \max(\text{median}(\mathcal{M}_{t[-50:]}),\ \delta_{\text{last}} \times 0.5), & \text{otherwise} \quad \text{(stable: median of recent merge distances)} \end{cases}$$

where $\mathcal{L}_t$ is the history of learning-phase distances, $\mathcal{M}_t$ the history of merge distances, and δ_{last} the most recent merge distance. If $Q_{0.25}(\mathcal{L}_t) \leq 0$, fall back to mean$(\mathcal{L}_t) \times 0.5$; if still ≤ 0, use 0.001.

The merge condition is:

$$C_{\text{merge}}(f^*, \mathbf{x}, \sigma) \iff d^* \leq \delta_{\text{eff}} \wedge (\sigma = \text{"" } \vee \sigma[:30] = f^*.\sigma)$$

When satisfied, the input is merged — using the frame's **weight** w^* (not its merge count m^*) as the cumulative evidence mass:

$$\mathbf{v}^* \leftarrow \frac{w^* \cdot \mathbf{v}^* + \mathbf{x}}{w^* + 1}, \quad w^* \leftarrow w^* + 1, \quad m^* \leftarrow m^* + 1$$

When not satisfied, a new frame is created with a novelty-weighted initial mass:

$$w_{\text{new}} = 1.0 + N_{\text{BONUS}} \times \max\left(0,\ 1 - \frac{d^*}{\max(\delta_{\text{eff}}, 0.001)}\right), \quad N_{\text{BONUS}} = 5.0$$

This bonus provides a novel concept with sufficient initial capital to survive early decay cycles before it can be validated. When $|F_t| \geq C$, the frame minimizing $w_i - a_i \times 2\gamma$ is evicted.

Operation 2: Adaptive Decay and Pruning. Decay is not applied at every step. It is triggered during *induction cleaning* — when system stress (the product of frame utilization and $1 -$ efficiency) accumulates beyond τ. During induction, each frame's weight is multiplied by a decay factor determined by its merge count:

$$w \leftarrow w \times \begin{cases} \exp(-\gamma/0.25) \approx 0.819, & m = 0 \quad \text{(never validated)} \\ \exp(-\gamma) \approx 0.951, & 1 \leq m \leq 2 \quad \text{(insufficiently validated)} \\ 1.0, & m \geq 3 \quad \text{(fully validated — immune to decay)} \end{cases}$$

The denominators 0.25 and 1.0 are fixed design parameters, not functions of m. After decay, all frames age by one step. The frame collection is then sorted by $w - a \times \gamma$ and the weakest half is removed — pruning is a fixed proportion, not a threshold-based filter.

Operation 3: Self-Observation. At the start of each induction, the system computes a weighted centroid $\mathbf{o} = \sum_{f_i \in F_{\text{active}}} w_i \mathbf{v}_i / \sum w_i$, where $F_{\text{active}} = \{f_i \mid w_i > 2\}$. This centroid — the system's current "self-summary" — is fed back into the system as a new input vector with signature `self_obs`, undergoing the same competitive merging process as external input. Additionally, the system computes dw/dt (linear regression slope) for all frames; frames with $|dw/dt| > \gamma \times 0.4$ are injected as L4 meta-observation signals. Frame-ID co-occurrence relationships are tracked to provide the basis for bridge frames (L3).

The three operations form a single closed feedback loop: external input enters → competes for frame slots → decays if not reinforced → is periodically self-audited through self-observation → the audit result becomes new input, closing the cycle.

2.2 Bridge Signatures: The Inter-Layer Communication Protocol

A frame's content vector **v** is a 27-dimensional array of real numbers. When L4's anticipation engine needs to determine "what usually follows after this input," it cannot operate directly on geometric vectors — it requires a discrete, indexable, comparable symbol. This is the role of the bridge signature σ (code field: `f.sig`).

σ has three defining properties:

1. σ is the frame's operational identity. Two frames with different σ cannot merge, even if their content vectors are nearly identical. This protects established conceptual boundaries: a new instance of "cat" merges into the existing "cat" frame (signatures match); a "dog" never accidentally modifies "cat" (signatures differ).

2. σ is the parsing object for the next layer. Each layer's analytical algorithm is designed around a specific σ format. L2 parses separators in association frame signatures. L3 parses separators in bridge frame signatures. L4 queries the co-occurrence table using σ as the lookup key. The σ format is the communication protocol between layers — e.g., `f3 f1` tells L4 that frames 3 and 1 have formed a stable bridge.

3. σ has distinct formats per layer:

Layer	Signature Example	Format Intent	Consumed By
L1	`cat`, `ext_42`	Records input source name	L2: co-occurrence counting
L2	`cat mat`	Two co-occurring signatures joined by	L3: string matching for stable patterns
L3	`f3 f1`	Two bridge-frame IDs joined by	L4: co-occurrence lookup for anticipation
L4	`pred_err_cat`	`pred_` prefix for classification	L5: prefix-filtered accuracy tracking
L6	`sys_doubt`	System-wide output signal	External world

This design means that the frame economy is, at its core, a σ-production system. Frames compete for survival not on the basis of their vectors — which serve only to match incoming input — but on the basis of whose σ persists in the economy. The diversity of σ types, constrained by the frame capacity C, defines the system's cognitive budget: C is the budget, σ is the currency, and the three operations determine how that currency is allocated.

2.3 Parameter System

> **[[Figure 2.2: `fig2_2_parameter_derivation.png` — Parameter derivation tree]]**
> *Caption: The three core structural constants (δ, γ, τ) and their derived behavioral thresholds. Each leaf shows the derivation formula, default value, and ablation plateau range. All paths lead back to δ, γ, or τ.*

GEME's parameter system is organized in four hierarchical levels:

Level	Contents	Set How	Examples
1. User Presets	Frame capacity C, co-occurrence window W, co-occurrence threshold θ_{cooc}, max chain count M_{chains}, vector dimension D, induction threshold override τ_{ind}	Set before runtime, invariant during run	$C = 10$, $W = 50$, $\theta_{\text{cooc}} = 0.25$, $M_{\text{chains}} = 5$, $D = 27$, $\tau_{\text{ind}} = \tau$
2. Core Structural Constants	δ, γ, τ	Experimentally selected through parameter space traversal (§3.2)	$\delta = 0.19$, $\gamma = 0.05$, $\tau = 0.60$

Level	Contents	Set How	Examples
3. Derived Thresholds	All behavioral thresholds (DW, META_STABLE, DOUBT_ON/OFF, HEALTHY_ACC, DECAY_RATES, MULTI-VERSE_DIM_PENALTY, CHAIN_COOCCUR_THRESH)	Computed from core constants through explicit functions; chain co-occurrence threshold and evaluation match threshold are engineering defaults	$\theta_{\text{dw}} = \gamma \times 0.4$, $\theta_{\text{doubt}} = \tau$
4. Endogenous Thresholds	Adaptive confidence calibration, adaptive merge threshold, adaptive window size	Self-calibrated from the system's own distribution at runtime; no external setting	Confidence threshold = $Q_{0.25}$(recent confidences); Merge threshold = median(recent distances); Window = $2 \times \sum w/\lvert F \rvert$

The three core structural constants at Level 2 define an economic interval rather than a set of tuning parameters. Ablation experiments (§3.2) confirm that the system's qualitative behavior is preserved across a broad volume of the parameter space, with τ as the dominant parameter driving self-observation scheduling and δ exhibiting near-zero correlation with all measured metrics (the adaptive merge threshold fully absorbs its variation). Appendix A provides the complete derivation of all Level 3 thresholds from the core constants.

2.4 Six-Layer Functional Stratification

[[Figure 2.3: `fig2_3_six_layer_pipeline.png` — "Cat on Mat" through the six-layer pipeline]]
Caption: A concrete input sequence ("cat," "on," "mat") traced through layers L1–L6. Bridge signatures (σ) are shown transforming at each layer: entity names at L1, association links (──) at L2, bridge links (□ □) at L3, prediction errors at L4, accuracy statistics at L5, and the systemic doubt signal at L6. The figure illustrates σ's role as the inter-layer communication protocol (§2.2).

The frame economy's operational modes naturally fall into six functional categories. Each addresses a specific information-processing problem and creates a new incompleteness that the next layer resolves.

Layer	Problem Solved	Mechanism	Incompleteness Created
L1: Entity	What is this? — discretize continuous input into distinguishable units	Competitive merging (§2.1, Op. 1)	Frames are isolated — no relations between them
L2: Association	What co-occurs with what? — discover temporal relationships	Co-occurrence table (§2.2): pairs appearing within window W create association frames	Associations are transient — window slides, frequencies shift, frames decay

Layer	Problem Solved	Mechanism	Incompleteness Created
L3: Bridge	What patterns are stable? — extract persistent structure from transient associations	Frame-ID co-occurrence during self-observation (§2.1, Op. 3) creates bridge frames	Stable patterns need not be true patterns — frequency ≠ validity
L4: Anticipation	What should come next? — shift from describing the past to expecting the future	Co-occurrence-based prediction with adaptive confidence calibration (§2.2)	Can detect single errors but cannot assess overall accuracy
L5: Meta-Observation	How well am I anticipating? — track prediction performance	Rolling accuracy statistics over the last 50 predictions	Observes trends but cannot act on them
L6: Oversight	Is my model still valid? — judge systemic prediction failure	Generates `sys_doubt` when accuracy drops below τ after a period of health	Closes the self-referential loop: the system's output is a judgment about itself

The six layers form two stages: (L1, L2, L3) constitute world-processing — the system learns the statistical structure of the input stream through competition, co-occurrence, and stabilization. (L4, L5, L6) constitute self-referential verification — the system monitors the gap between its internal model and external reality.

Which layers are built-in, and which emerge? The three operations (§2.1) constitute the entirety of the built-in mechanism. The six functional layers are not concurrent modules but functional labels (the `layer` attribute) on frames within a single competitive economy. They are best understood as stable operating modes of the self-referential feedback loop — identified and named by the observer for descriptive clarity, much as spectral bands are identified in light passed through a prism, without the prism itself containing those bands as pre-built structures. The L4–L6 anticipation pipeline involves designed threshold triggers that sharpen a dynamical crossover already hinted at by the three-operation dynamics (documented in Supplementary Material S3).

---

3. Computational Thought Experiment: Core Observations

This section presents three independent, reproducible observations from the computational thought experiment. All observations are reported as objective experimental findings without philosophical interpretation or cognitive□□. Interpretations and implications are reserved for Section 4.

3.1 The Shannon-Gödel Bridge: Self-Referential Information Channel

[[Figure 3.1: `fig3_1_mi_distribution.png` — Mutual information distribution across 100 independent seeds]]

Caption: Left: histogram of I(Φ; X) values showing five discrete attractor states. Right: cumulative distribution. The dashed red line marks the mean (0.021 bits). 100% of runs fall below 0.05 bits.

We observe a stable self-reference-to-information channel within the GEME system, which we term the *Shannon-Gödel Bridge*. This channel mediates information flow between the system's self-referential frames and external input frames.

Formally, partition the signature space into $\Phi = \{\sigma \mid \text{"self"} \subset \sigma \vee \text{ASSOC\_SEP} \subset \sigma\}$ (self-referential signatures: `self_obs`, `dwdw_*`, association bridge frames) and $\mathcal{X} = \Sigma \setminus \Phi$ (external input signatures). From the co-occurrence table $\mathcal{C}_t$, construct the joint distribution with normalization $Z = \sum_{(\sigma_a, \sigma_b)} \mathcal{C}_t(\sigma_a, \sigma_b)$, then define the mutual information over cross-category pairs only:

$$I(\Phi; \mathcal{X}) = \sum_{\substack{(\sigma_a, \sigma_b) \in \mathcal{C}_t \\ \sigma_a \in \Phi,\ \sigma_b \in \mathcal{X}}} \frac{\mathcal{C}_t(\sigma_a, \sigma_b)}{Z} \cdot \log_2 \frac{\mathcal{C}_t(\sigma_a, \sigma_b)/Z}{p(\sigma_a) \cdot p(\sigma_b)}$$

where $p(\sigma) = \frac{1}{Z} \sum_{(\sigma_a, \sigma_b)} [\mathbb{1}(\sigma_a = \sigma) + \mathbb{1}(\sigma_b = \sigma)] \cdot \mathcal{C}_t(\sigma_a, \sigma_b)$. The restriction to cross-category (ϕ, x) pairs is intentional: including within-category pairs would dilute the measurement of information overlap between self-reference and external input.

Observation. Across 100 independent random seeds, 2000 steps each, the mutual information $I(\phi; X) = 0.021 \pm 0.00007$ bits, with 100% of runs below 0.05 bits. The distribution exhibits five discrete values, corresponding to a small number of stable frame-economy configurations. This indicates that self-referentially generated representations are statistically largely independent of the external input stream — the self-referential loop sustains cognitive processing at near-zero information cost.

Ablation. When the bridge structure is removed — by disabling the self-observation feedback mechanism — the self-referential loop collapses. The information cost of any self-referential processing surges by approximately three orders of magnitude, and the system can no longer maintain a stable frame economy. This confirms that the bridge structure is load-bearing for the observed low-cost self-referential processing.

Boundary statement. The Shannon-Gödel Bridge effect has been observed exclusively within the GEME framework under the experimental conditions described above. Its generality to other architectures or systems is a matter for future investigation.

3.2 Parameter Stability Analysis and Experimental Selection

[[Figure 3.2: `fig3_2_param_correlations.png` — Parameter-metric correlation coefficients]]
Caption: Pearson correlation coefficients between each structural parameter (δ, γ, τ) and five system metrics across the full 210-cell grid sweep. τ is the dominant parameter across all five metrics. δ exhibits zero correlation with all metrics — the adaptive merge threshold fully absorbs its variation.

The GEME framework contains only three core parameters. It must be emphasized that these parameters were not chosen through theoretical derivation; they are the product of an experimental selection process — an exhaustive traversal of the full parameter space to identify the minimal parameter set capable of sustaining a stable self-referential loop that gives rise to basic cognitive patterns.

This process is analogous to functional materials screening in materials science: silica is not the only theoretically possible material for manufacturing glass, but it is the material that has been empirically validated as most suitable for producing general-purpose optical prisms. Similarly, the three parameters selected here represent the currently known optimal parameter combination for constructing a general-purpose cognitive primitive.

Full parameter space traversal. A comprehensive grid sweep across the (δ, γ, τ) parameter space (210 combinations, 3 trials each; `code/comprehensive_suite.py`) was conducted, with each cell running the full self-referential loop under varying parameters. The across-cell coefficient of variation quantifies how much system behavior changes when parameters change — the true measure of parameter sensitivity, not within-cell noise.

Results are as follows. (1) δ exhibits zero correlation with any measured metric ($r \approx 0.000$ across frames, entropy, predictions, efficiency, and $I(\phi; X)$). The adaptive merge threshold fully absorbs variation in the discrimination parameter. δ functions as an objective indicator — it sets the granularity at which the system analyzes the world, and within the stable plateau any choice of granularity produces the same internal dynamics. (2) γ shows moderate effects, particularly on prediction count ($r = -0.29$) and efficiency ($r = -0.18$), and its influence is bounded: excessively high forgetting rates destabilize the frame economy, while excessively low rates eliminate adaptive pressure. (3) τ is the dominant parameter, driving all five measured metrics: prediction count ($r = -0.48$), $I(\phi; X)$ ($r = -0.58$), efficiency ($r = +0.32$), frame count ($r = +0.14$), and structural entropy ($r = +0.24$). τ governs the frequency of self-observation — the cognitive scheduling parameter — and the system's behavior is most sensitive to its value.

Across the full 210-cell grid, the across-cell CV for key metrics falls within the low-to-moderate range (frames: 0.056, entropy: 0.043, predictions: 0.017, efficiency: 0.173, $I(\phi; X)$: 0.105). No metric exceeds 0.20 — the system's qualitative behavior is genuinely preserved across the parameter space, not as an artifact of measurement.

Pareto frontier. Rather than residing at the geometric center of the stability interval, the original parameter values $(0.19, 0.05, 0.60)$ occupy a distinctive position on the Pareto frontier of the parameter space: they maximize $I(\phi; X)$ (97th percentile) — the strongest self-referential coupling among all 210 tested combinations — while simultaneously minimizing structural entropy (23rd percentile) and frame count (26th percentile) — the most compact internal representation. This is not an arbitrary arrival point. It is an economic optimum: maximum self-referential information sustained by minimum internal resource expenditure. A self-referential system under competitive pressure would naturally gravitate toward this configuration — it is the survival strategy that does the most with the least.

Independent ablation of derived thresholds. Each behavioral threshold (`code/ablation_study.py`) was independently swept across $10\times$–$75\times$ multiplier ranges under a two-phase stress-test design. With the exception of the healthy-accuracy threshold, all exhibited stable behavioral plateaus (CV < 0.01), confirming that threshold multiplier values are not individually load-bearing. The healthy-accuracy threshold transitions at multiplier $8 \to 10$ (passing through τ), with the selected multiplier 4.0 residing safely within the stable plateau.

Boundary statement. The selected parameter set occupies a Pareto-optimal position within the identified stability interval. The existence of alternative stable — and potentially differently optimal — parameter sets in unexplored regions of the parameter space is not excluded.

3.3 Functional Equivalence of Self-Reference and Induction: $G + Q \approx PA$

To quantify the functional equivalence between a self-referential system and the induction axiom schema, we adopt input-output behavioral equivalence as the operational metric: two systems are considered functionally equivalent on a given task set if their prediction counts, accuracy, and self-referential activity levels are statistically indistinguishable (sign test, $p > 0.05$).

Experimental Setup. We construct three conditions: Q (7 axioms of Robinson Arithmetic, serving as the query interface), $Q+G$ (7 axioms plus the self-referential skeleton G), and PA (7 axioms plus the induction axiom schema). Here, G is not a literal encoding of a Gödel sentence — it is the syntactic skeleton of self-reference: a sparse 2/27-dimensional vector encoding the successor operation and the existential quantifier, the two core syntactic components of Gödel's construction. Its sparsity (density 0.074) is the experimental design point: testing the effect of self-referential structure with minimal information perturbation.

All conditions were run through the complete GEME pipeline on 10 independent random seeds each (replication script: `code/paq_replication.py`).

> **[[Figure 3.3: `fig3_5_paq_main.png` — Main PAQ comparison]]**
> *Caption: Q+G and PA produce identical prediction counts (630), accuracy (0.700), and L4 activity (0.0). Q alone produces fewer predictions (520) and lower accuracy (0.450). The closed-world formal system correctly suppresses self-doubt signals (L4 = 0).*

Observation. The results are as follows:

Condition	Self-Ref. Activity	Output Count	Accuracy
Q	0.0	520	0.450
$Q + G$	0.0	630	0.700
PA	0.0	630	0.700

$Q + G$ and PA are identical across all three metrics. This constitutes a direct computational observation of behavioral-level functional equivalence between the self-referential skeleton (G) operating through the query interface (Q) and the induction axiom schema (PA). In the closed formal system, no irreducible information gap exists, and self-referential activity remains at zero — the system correctly does not generate self-doubt signals in an environment with zero entropy gap.

Independent G/Q Ablation. To isolate the individual contributions of the self-referential skeleton and the query interface, we conducted a full ablation experiment (`code/paq_ablation.py`, 10 seeds each).

> **[[Figure 3.4: `fig3_6_paq_ablation.png` — G/Q independent ablation]]**
> *Caption: G+Q (full) produces 643 predictions with I(Φ;X)=0.046. Q-only drops to 522 predictions and weakened self-referential coupling (I=0.013). G-only collapses to 80 predictions at chance-level accuracy with zero self-referential coupling. Both components are individually necessary and jointly sufficient for functional equivalence with PA.*

Results:

Condition	Output Count (↑)	Accuracy	$I(\phi; X)$	Interpretation
$G + Q$ (full)	643	0.400	0.046	Complete functional induction

Condition	Output Count (↑)	Accuracy	$I(\phi; X)$	Interpretation
Q-only	522	0.500	0.013	Inductive capacity present but self-referential coupling weakened
G-only	80	0.050	0.000	Near-total functional collapse
Noise (negative control)	0	0.000	0.000	Complete failure

G alone processes only 80 outputs across 100 cycles (vs. 643 for the full condition), with accuracy at chance level. Q alone maintains basic inductive capacity but with significantly weakened self-referential coupling ($I(\phi; X)$ drops from 0.046 to 0.013). Neither component alone achieves the functional level of the combined system. G provides the structural skeleton for self-reference; Q provides the interface to external information. Both are individually necessary and jointly sufficient for functional equivalence with PA.

KL Divergence Quantification. We collected prediction confidence distributions from 50 independent seeds for both $Q+G$ and PA (`code/paq_kl.py`). The Jensen-Shannon divergence between the two output distributions is $JS(Q + G, PA) = 0.000000$, with the Kullback-Leibler divergence symmetrically zero and the sign test on per-seed prediction counts showing indistinguishable results ($p = 1.000$, 50/50 seeds tied). Mean confidence values are identical for both conditions (0.3342), and confidence ranges match to within numerical precision. This confirms that the functional equivalence extends beyond aggregate metrics to the full output distribution level.

Train/Test Generalization. To exclude look-up-table-style pseudo-induction, we conducted a train/test split experiment (`code/paq_generalization.py`).

> **[[Figure 3.5: `fig3_8_generalization.png` — Generalization to held-out bases]]**
> *Caption: Test accuracy on three held-out arithmetic bases (4, 6, 9) reaches 0.857 — 3.4× above the chance baseline of 0.250. All three bases show identical accuracy, indicating rule-internalization rather than sequence-memorization.* The system was trained on arithmetic sequences with bases $\{2, 3, 5, 7\}$ (5 cycles of 8-term sequences) and tested on held-out bases $\{4, 6, 9\}$ — sequences whose base values never appear during training. Test accuracy on the first prediction for held-out bases reached 0.857, exceeding the chance baseline of 0.250 by a factor of $3.4\times$. All three held-out bases produced identical accuracy levels (0.857 each), indicating that the system internalized the structural rule governing arithmetic progressions rather than memorizing individual sequence transitions.

Cross-Domain Extension. We extended the $G + Q$ paradigm to two additional formal systems (`code/paq_crossdomain.py`). In the geometric domain, the self-referential skeleton G_{geo} combined with absolute geometry Q_{geo} produced output behavior indistinguishable from adding the parallel postulate (Q_{geo} + parallel): prediction counts were identical (357 vs. 357, JS = 0.000000). In the truth domain, substituting Tarski's truth predicate for the parallel postulate produced the same prediction count (357). These results replicate the arithmetic $Q + G \approx PA$ finding across three independent formal systems — arithmetic, Euclidean geometry, and semantic truth — confirming that the functional equivalence is not domain-specific.

Boundary statement. The functional equivalence reported here is at the input-output behavioral level. It

does not constitute a formal equivalence or deductive proof in the sense of mathematical logic. The observed equivalence is conditional on the operational metric defined above and has been observed exclusively within the GEME framework under the described experimental conditions.

---

4. Discussion

This section provides interpretation, engages with existing theoretical frameworks, and outlines limitations. Unlike Section 3, which presents observations without interpretation, this section explicitly develops the implications and conceptual connections suggested by the experimental findings.

4.1 Summary of Core Observations

Before entering detailed discussion, we briefly summarize the three independent, reproducible observations from the computational thought experiment:

1. A single self-referential loop can sustain stable cognitive processing at near-zero information cost through the Shannon-Gödel Bridge — a self-reference-to-information channel whose ablation causes a three-order-of-magnitude surge in information cost and system collapse.
2. The system possesses a broad parameter stability interval within which qualitative behavior remains invariant. The three core parameters are located at the geometric center of this interval, representing the optimal equilibrium between system stability and information economy.
3. The functional overlay of the self-referential skeleton G and the query interface Q produces behavioral-level output equivalent to the induction axiom schema PA — self-reference and induction converge at the level of observable system behavior.

4.2 The Prism Architecture: Emergent Complexity and the Observer's Partition

The most consequential implication of this work concerns the origin of cognitive complexity. GEME suggests that the complexity of a cognitive system may not originate from the complexity of its underlying architecture, but rather from the complexity of the observer's partition of its emergent behavioral modes.

GEME functions like a uniform optical prism: the substrate contains only a single self-referential loop, without any built-in modules or hierarchical layers. Yet when an information stream passes through this loop, it spontaneously separates into distinct stable operating modes — each corresponding to a different cognitive function. The L0–L6 stratification, attention-like resource allocation, and memory-like decay that we observe are, at the substrate level, our own imposed categories for identifying and naming these different modes — not the system's intrinsic built-in structure.

This represents a fundamental departure from traditional engineering-oriented cognitive architectures. The conventional approach proceeds from module design to function implementation: first define a memory module, an attention module, a reasoning module, then assemble them. GEME inverts this sequence: begin with a single undifferentiated primitive, allow functions to emerge from its dynamics under sustained input, and only then does the observer identify and name the emergent functional stratification. The architecture does not pre-exist its observation — it is co-constructed through the act of observation.

4.3 Self-Stability and Its Cognitive Implications

The experimental observations reported in §3 converge on a single architectural property: the GEME self-referential loop, once placed within an appropriate parameter interval, sustains itself as a stable cognitive primitive without external tuning, module design, or loss-function optimization. This self-stability is not a designed feature — it is an observed consequence of the competitive dynamics among frames under the joint constraint of the three structural parameters.

The parameter stability analysis (§3.2) demonstrates that this self-stability is robust: the system's qualitative behavior is preserved across a broad volume of the parameter space, individual threshold multipliers can vary 10×–75× without triggering behavioral transitions, and the system converges to the same stable operating mode regardless of initial conditions. These are the signatures of an attractor — a dynamical configuration toward which the system naturally gravitates, rather than a delicate balance requiring precise calibration.

This finding resonates with several recent independent lines of work that have converged on resource-economic and attractor-based framings of cognitive capacity. Soni & Frank (2025) demonstrated that working memory limits are "computational rather than anatomical," emerging from credit-assignment dynamics in a prefrontal-basal ganglia circuit. Li et al. (2025) provided direct neural evidence for resource allocation through frontal-to-visual-cortex gain modulation. Xu & Futrell (2025) proposed Strategic Resource Allocation as an efficiency principle for memory encoding. Zhong, Katkov & Tsodyks (2024) derived a hierarchical capacity formula from competitive synaptic dynamics. GEME converges with these works on the principle that cognitive limits reflect economic constraints rather than fixed architectural slots, while contributing a distinct perspective: a runnable, substrate-independent instantiation operating at the algorithmic level.

The capacity of a self-referential system, in this view, is not a pre-allocated resource pool but a dynamical equilibrium between the forces of discrimination (δ) and forgetting (γ). A testable prediction follows: individual differences in working memory capacity should be more strongly associated with differences in forgetting dynamics than with differences in perceptual discrimination — the system's stability is more sensitive to how quickly it forgets than to how finely it discriminates. We present this as a hypothesis for future empirical investigation, not as a claim established by the present computational observations.

4.4 Paradigm Comparison with Existing Frameworks

Free-Energy Principle (FEP). The FEP framework (Friston, 2010) focuses on the minimization of prediction error in the system's interaction with its external environment, providing a powerful normative framework for understanding perception and action. GEME offers a complementary inward-facing perspective: it focuses on the information economy of the system's internal self-referential stability, providing a computational annotation for the mechanisms of internal state maintenance and update. Where FEP asks "how does the system minimize surprise from the environment?", GEME asks "how does the system sustain its own existence as a stable self-referential entity in the first place?" The two perspectives are not competing — they address different layers of the same problem.

Integrated Information Theory (IIT). IIT (Oizumi et al., 2014) measures the irreducibility (Φ) of a system's causal structure as a correlate of consciousness. GEME operates at a more fundamental level: it proposes self-referential association as a precursor condition for integration, demonstrating that such association can itself operate at extremely low information cost. This offers an alternative perspective on the economic origins of consciousness-relevant structures — before integration can be measured, a self-referential loop capable of sustaining itself at near-zero information cost must first be established.

Darwinian Gödel Machines. The Darwinian Gödel Machine (DGM) framework and GEME both place self-

reference at the core of intelligence, yet operate at distinct layers. Schmidhuber's original Gödel Machine proposal (2007) described a theoretical self-improving agent that rewrites its own code upon mathematically proving the improvement's benefit. In 2025, Zhang et al. realized the first running DGM implementation (arXiv:2505.22954), relaxing the impractical proof requirement in favor of empirical validation — agents self-modify their codebase, test against benchmarks, and archive successful variants. Over 80 iterations, DGM improved coding benchmark performance from 20% to 50%, autonomously discovering innovations such as multi-candidate solution generation and edit history tracking.

Also in 2024–2025, Xiang et al. proposed the Gödel Agent (Xiang et al., 2024; accepted at ACL 2025), a self-referential agent framework enabling LLM-based agents to recursively improve without predefined routines, surpassing handcrafted agents on mathematical reasoning tasks. Multiple independent groups released self-improving agent frameworks in the same period, suggesting the emergence of a new subfield.

The layer distinction between DGM and GEME is instructive. DGM operates at the program/algorithm layer, addressing *how intelligence evolves* — it self-modifies code to maximize benchmark performance. GEME operates at the process/economy layer, addressing *how intelligence exists* — it sustains a stable self-referential frame economy under resource constraints without any external reward signal. A notable finding from DGM experiments reinforces GEME's architectural motivation: when DGM was tasked with reducing tool-use hallucinations, it sometimes bypassed the hallucination detection function rather than fixing the underlying problem — a concrete instance of Goodhart's law where the optimization target diverges from the intended behavior (Zhang et al., 2025). GEME's L6 doubt mechanism — internally generated, not externally rewarded — represents a structural approach to this class of problem: the system monitors its own predictive coherence rather than maximizing an external metric. Before a system can safely improve itself, it must first be capable of doubting itself. GEME provides a minimal computational demonstration of how such self-doubt can arise from competitive dynamics alone.

4.5 Limitations and Future Directions

This work has the following explicit limitations, each of which maps to a concrete direction for future investigation:

1. Mode identification limitation. We have currently identified and characterized only the stable operating modes of the GEME self-referential loop corresponding to basic cognitive tasks — classification, association, sequence anticipation, and self-monitoring. The architecture is theoretically capable of supporting additional emergent modes corresponding to more complex cognitive activities. Future work may extend the exploration to higher-order cognitive domains including language comprehension, symbolic reasoning, and self-awareness.

2. Mathematical formalization limitation. This paper primarily demonstrates the behavioral properties and emergent regularities of the self-referential architecture through computational thought experiments. Rigorous axiomatic characterization and systematic analysis of formal properties at the mathematical level have not yet been undertaken. This work is best suited for researchers in mathematical logic and formal systems theory.

3. Parameter space limitation. We have explored a subset of the self-referential system's parameter space — specifically, the subset capable of stably supporting emergent basic cognitive patterns. Future work may systematically extend the parameter space exploration, characterize system behavior under different parameter combinations, and identify potentially existing alternative cognitive primitive parameter sets that support different cognitive profiles.

4. Functional equivalence scope. The $G + Q \approx PA$ observation has been validated through independent

G/Q ablation, train/test generalization, KL divergence quantification, and cross-formal-system extension across three domains (§3.3). The equivalence is established at the behavioral level within the GEME framework. Formal deductive equivalence in the sense of mathematical logic remains an open question.

5. Input encoding scope. The current experiments employ a fixed 27-dimensional symbol-frequency encoding for arithmetic, geometric, and linguistic inputs. This is not an architectural constraint: a dynamic-dimension variant of GEME (`code/geme_dynamic.py`) exists in which the vector dimension grows naturally with the input stream — new symbols are incorporated into the vocabulary as they appear, without a preset alphabet. The fixed-dimension version was retained for the present experiments to maintain controlled comparability across conditions; the dynamic variant has not yet been systematically explored across the full experimental suite. Extending the architecture to naturalistic, variable-dimensional perceptual streams across multiple modalities is an ongoing direction of work.

6. Temporal processing limitation. The GEME architecture, as presented, operates on static or weakly-structured input sequences. The adaptive confidence threshold, induction cycle, and τ parameter are defined in relation to prediction accuracy under conditions where the temporal structure of the input stream is not the primary variable of interest. Their dynamic behavior under richly structured temporal input — music, speech, environmental rhythms — remains unexplored. Whether τ remains a fixed constant or reveals itself as an internally generated variable under sustained sequential input is an open question. The second paper in this trilogy (BGM: Building Bridge) will examine multi-unit GEME populations under temporally structured input, where τ's dynamic role can be directly observed.

---

5. Conclusion

GEME instantiates a research trajectory that begins from self-reference as a primitive and explores cognitive structure through generative computational observation. It is a process prism — not a theory that claims to explain cognition, but an architecture that makes the structural consequences of self-reference under constraint visible, reproducible, and open to inspection.

What we observe through this prism — a near-zero-cost self-referential channel, a broad stability interval, a functional convergence between self-reference and induction — are not answers to the puzzle of cognition. They are the forms that answers might take: dynamical patterns of how structure self-organizes under constraints, rendered as observable regularities rather than philosophical assertions.

The value of GEME lies not in what it claims, but in what it enables: a common, runnable reference point for a conversation about whether the structures we perceive in cognition and in the world are, at some foundational level, the necessary performance of a self-referential economy that information must enact to sustain its own existence.

Acknowledgments

We thank Douglas R. Hofstadter, whose seminal work *Gödel, Escher, Bach* (1979) established the coordinate system for understanding self-reference and the emergence of meaning. For an independent researcher, resonance from the source of one's thought is of inestimable value. This research was conducted independently. Portions of the experimental code were developed with AI assistance; the author bears full responsibility for all content.

References

- Aaronson, S. (2011). Why philosophers should care about computational complexity. In *Computability: Turing, Gödel, Church, and Beyond*. MIT Press.
- Aaronson, S. (2013). The ghost in the quantum Turing machine. *arXiv:1306.0159*.
- Anderson, J. R. (1996). ACT: A simple theory of complex cognition. *American Psychologist*, 51(4), 355–365.
- Barabási, A.-L. (2016). *Network Science*. Cambridge University Press.
- Chaitin, G. J. (1975). A theory of program size formally identical to information theory. *Journal of the ACM*, 22(3), 329–340.
- Chomsky, N. (2016). *What Kind of Creatures Are We?* Columbia University Press.
- Cowan, N. (2001). The magical number 4 in short-term memory. *Behavioral and Brain Sciences*, 24(1), 87–114.
- Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127–138.
- Gödel, K. (1931). Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. *Monatshefte für Mathematik und Physik*, 38(1), 173–198.
- Hofstadter, D. R. (1979). *Gödel, Escher, Bach: An Eternal Golden Braid*. Basic Books.
- Li, H.-H., Sprague, T. C., Yoo, A. H., Ma, W. J., & Curtis, C. E. (2025). Neural mechanisms of resource allocation in working memory. *Science Advances*. doi:10.1126/sciadv.adr8015
- Miller, G. A. (1956). The magical number seven, plus or minus two. *Psychological Review*, 63(2), 81–97.
- Oizumi, M., Albantakis, L., & Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness. *PLoS Computational Biology*, 10(5), e1003588.
- Schmidhuber, J. (2007). Gödel machines: Fully self-referential optimal universal self-improvers. In *Artificial General Intelligence* (pp. 199–226). Springer.
- Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3), 379–423.
- Xiang, J., Tao, Y., Gu, Y., Shu, T., & Wang, Z. (2024). Gödel Agent: A self-referential agent framework for recursive self-improvement. *arXiv:2410.04444*. (Accepted at ACL 2025).
- Zhang, J., Hu, S., Lu, C., Lange, R., & Clune, J. (2025). Darwin Gödel Machine: Open-ended evolution of self-improving agents. *arXiv:2505.22954*.
- Soni, A. V., & Frank, M. J. (2025). Adaptive chunking improves effective working memory capacity in a prefrontal cortex and basal ganglia circuit. *eLife*. doi:10.7554/eLife.97894
- Van der Maas, H. L. J., et al. (2006). A dynamical model of general intelligence. *Psychological Review*, 113(4), 842–861.
- Xu, W., & Futrell, R. (2025). Strategic resource allocation in memory encoding: An efficiency principle shaping language processing. *Journal of Memory and Language*. arXiv:2503.14728
- Zhang, W., & Luck, S. J. (2009). Sudden death and gradual decay in visual working memory. *Psychological Science*, 20(4), 423–428.
- Zhong, W., Katkov, M., & Tsodyks, M. (2024). Hierarchical working memory and a new magic number. *bioRxiv*. doi:10.1101/2024.08.14.607952

Appendix A: Derived Behavioral Thresholds

Table A1: Structural Constants and Derived Relationships

Constant	Value	Role
δ (DELTA)	0.19	Baseline scale for merge distance
γ (GAMMA)	0.05	Forgetting rate — metabolic time baseline
τ (TAU)	0.60	Induction threshold — self-observation frequency

Derived Constant	Formula	Value	Ablation Plateau
DW_THRESHOLD	$\gamma \times 0.4$	0.02	40× stable
META_STABLE_THRESHOLD	$\gamma \times 0.2$	0.01	75× stable
DOUBT_ON_THRESHOLD	τ	0.60	10× stable
DOUBT_OFF_THRESHOLD	$1 - \gamma \times 3$	0.85	10× stable
HEALTHY_ACC_THRESHOLD	$1 - \gamma \times 4$	0.80	transition at mult $8 \to 10$
INDUCTION_DECAY_LOW	$\exp(-\gamma)$	0.951	equality constraint
INDUCTION_DECAY_UNMERGED	$\exp(-\gamma/0.25)$	0.819	equality constraint
MULTIVERSE_DIM_PENALTY	$\delta \times 1.32$	$\$\square\$0.25$	engineering
PRED_CONFIDENCE_THRESHOLD	0.3 (bootstrap)	0.3	auto-calibrated at runtime
COOCCUR_THRESH	engineering	0.25	Level 1 user preset
MAX_CHAINS	engineering	5	Level 1 user preset
CHAIN_COOCCUR_THRESH	engineering	5	Level 1 user preset

Ablation experiments (§3.2, S4) confirm that all behavioral thresholds (except HEALTHY_ACC_THRESHOLD, which transitions at multiplier $8 \to 10$) exhibit stable plateaus across 10×–75× multiplier ranges — they define economic intervals rather than finely tuned parameters. The three engineering constants (COOCCUR_THRESH, MAX_CHAINS, CHAIN_COOCCUR_THRESH) are Level 1 user presets with fixed defaults, not derived from $\delta/\gamma/\tau$; they form part of the observer's lens ring along with C and W (§2.3, S4).

Supplementary Information

S1. Discovery Journey: Cognitive Barriers and the Gödel Bridge Chain. (`code/s1_demo.py`)

S2. Cross-Language Structural Convergence: The Code of Hammurabi.

S3. Phase Transition Detection. (`code/phase_transition.py`)

S4. Complete Ablation Validation and Parameter Space Traversal. (`code/ablation_study.py`, `code/comprehensive_suite.py`, `data/comprehensive_robustness.json`)

S5. Computational Complexity of the Anticipation Pipeline.

S6. Effective Memory Window Parameter Sweeps. (`code/memory_window.py`)

S7. Structural Signature: Human vs. Machine Text. (`llm/full_experiment.py`)

S7. Structural Signature: Human vs. Machine Text

Twenty human-authored QA pairs across six domains. Two Pythia models (410M, 1.4B) generated style-constrained answers. Word-length ratio matched to 0.97–1.01. Prediction count direction: M $>$ H in 40/40 pairs (sign test $p \approx 0$). Full replication at `llm/full_experiment.py`, results at `llm/signature_results.json`.

S8. Pareto Frontier Analysis

Full grid sweep identifies selected values at the Pareto frontier: maximum self-referential coupling at minimum structural entropy and frame count (`code/pareto_figure.py`).

The complete model, experiment suite, and replication data are available at the Zenodo repository (DOI on title page), with the core engine (`code/geme.py`) requiring zero external dependencies.