Principia Cognitia: Axiomatic foundations

Abstract

This paper introduces Principia Cognitia, a comprehensive axiomatic framework that formalizes and extends the MLC/ELM (Metalanguage of Cognition/External Language of Meaning) duality currently under review in Cognitive Science. Building on empirical foundations established in companion work, we present a unified mathematical framework for cognitive processes across biological and artificial systems. The theory centers on a cognitive triad $\langle \mathrm{S}, \mathcal{O}, R_{\mathrm{rel}} \rangle$ and provides substrate-invariant formalization of cognitive operations, thermodynamically grounded processing constraints, and operationalizable metrics for temporal synchronization. We demonstrate applications through analysis of transformer architectures and propose experimental protocols for validation. This framework bridges symbolic and connectionist approaches, offering a mathematical foundation for cross-disciplinary cognitive research with immediate applications to AI alignment and human-machine collaboration.

---

1. Introduction

The history of cognitive science reveals repeated attempts to formalize the nature of thought and language. From the cogito of Descartes to the Computational Theory of Mind and contemporary neural models, each framework has oscillated between metaphysical speculation and empirical grounding. What remains missing is a unified axiomatic foundation: a system of first principles capable of structuring cognition without reliance on implicit metaphors or substrate-dependent assumptions.

The epistemic aim is modest yet exacting: to develop a universal, formally disciplined language for describing cognitive processes, applicable across substrates. Instead of offering a new ontology of mind or defending metaphysical claims about consciousness, Principia Cognitia seeks a grammar for cognition—capable of describing phenomena from chemical signaling in amoebas to primate gesture to LLM conversation. Internal coherence and descriptive power, not ontological classification, are the criteria of success. Ethical prescriptions are deliberately excluded; the framework is an operational tool, not a moral doctrine.

Its philosophical foundation rests on a triad: Turing, Wittgenstein, and Wiener. Together they replace subject-centered metaphysics with an analysis of cognition as algorithmic, linguistic, and regulatory structure—substrate-independent, observable, and modelable. This stance rejects the Cartesian legacy of mind as an inner essence, instead situating cognition in communicative and behavioral patterns.

Principia Cognitia advances three key theses: cognition is a continuous, multidimensional vector process rather than a binary property; language is not a tool of thought but the

medium in which thought occurs; and the resulting cognitive patterns are universal across systems that process information in such media, whether human or machine. By treating meaning as physically instantiated semions in dynamic relation, and by modeling cognitive capabilities as coordinates in a high-dimensional space, the framework prepares the ground for an axiomatic, physically grounded science of mind—one that describes rather than discovers it.

Mid-20th-century giants sketched principles that today's AI research has vividly confirmed. Turing showed that thought could be understood as universal computation, independent of its physical substrate—and judged by behavior rather than metaphysical status. Wiener's cybernetics reframed intelligence as feedback-driven optimization, warning that the real risks lay in speed and scale, not in alien modes of thought, and that underlying control principles are substrate-agnostic. Wittgenstein's later philosophy made meaning a product of use within social "language games," implying that fully developed reason depends on interaction with other minds. Empirical work on hardware invariance, Layer-Zero optimization across architectures, and the narrative-threshold effect in both humans and LLMs aligns with these insights.

Principia Cognitia presents itself as the formal synthesis of that triad. Its axioms $\langle \mathrm{S}, R_{\mathrm{rel}}, \mathcal{O} \rangle$ operationalize Turing-universal cognitive operations, its layered architecture captures Wiener-style optimization across substrates, and its explicit language-environment modules (MLC/ELM) embody Wittgenstein's socially constituted semantics. The framework is not offered as new metaphysics but as a mathematically rigorous, historically anchored architecture for describing cognition across biological and artificial agents—uniting and extending three prophetic visions into a "principia" for the science of mind.

Inspired by the formal rigor of Principia Mathematica (Whitehead & Russell, 1910) and Philosophiae Naturalis Principia Mathematica (Newton, 1687), Principia Cognitia seeks to construct cognition as a formal object, independent of its material realization but open to empirical instantiation. The motivation is not to reduce cognition to a single essence, but to provide a minimal and coherent system of axioms, theorems, and lemmas that allow precise articulation of phenomena such as language, perception, symbolization, and automated processes.

The core principles include:

- [AX-OPER] — cognition as a system of operations and transformations.
- [AX-SUBSTR-INV] — independence of cognitive form from material substrate (as a postulate).
- [AX-DISCR-01] — discretization of continuous phenomena into symbols and operations.
- [LM-SENS-01] and [LM-SYMB-01] — defining the transformation of sensory flux into structured cognition.
- [TH-LANG-01] — duality of languages: Mental Language of Cognition (MLC) and Expositional Language of Manifestation (ELM).

The project does not aspire to metaphysical closure. Rather, it provides a rigorous scaffold from which both empirical studies and philosophical reflection may proceed. By distinguishing clearly between phenomenon and signal, and between cognitive operation and its manifestation, Principia Cognitia rejects informational metaphysics ("the universe as computation") and grounds cognition in material reality, while remaining agnostic about specific physical realizations.

This framework supports interdisciplinary work—from philosophy of language and semiotics to neuroscience, artificial intelligence, and cognitive psychology. Its methodological stance is explicitly anti-reductionist yet formally minimalist: by beginning from axioms, it avoids hidden assumptions and offers testable consequences. In doing so, it seeks to make cognition a precise object of formal inquiry, comparable to how mathematics formalized number, or physics formalized motion.

The theoretical foundations presented here formalize and extend the MLC/ELM duality framework currently being evaluated for publication in *Cognitive Science* [Snow, 2025]. While that work focuses on empirical validation of the dual language hypothesis, the present paper provides the broader axiomatic structure within which such empirical findings can be systematically understood and generalized across cognitive substrates.

---

2. Boundary conditions for applicability in the material world

These boundary conditions define the minimal scope within which the axiomatic system of Principia Cognitia applies "within known reality." They delimit the framework from metaphysical or speculative extensions and ensure formal grounding. They are applicability conditions, not axioms of the theory.

2.1. BC-01: Subject boundary [Axiomatis Subiecti]

- **Statement:** Cognition presupposes a subject X that is materially instantiated and bounded from the environment by a semi-permeable boundary ∂X, allowing exchange of energy, matter, and information.

$$\exists X \subseteq U \text{ s.t. } \partial X = B,\ \mathrm{S}_X \neq \mathrm{S}_{\neg X},\ \text{and } Cognitio(X)$$

- or equivalently

$$\exists \mathrm{S}\ (Subject(\mathrm{S}) \wedge Oper(\mathrm{S}, \Phi, \Sigma, \Lambda))$$

- where U is the physical universe, Φ phenomena, Σ signals, Λ symbols.
- **Rationale:** Without a boundary there is no internal/external distinction; hence no cognition.

2.2. BC-02: Layered architecture with Layer-0 [Postulatum Stratificationis]

- **Statement:** Cognition operates within a layered architecture $\langle \mathrm{L}_0, \mathrm{L}_1, \mathrm{L}_2, \mathrm{L}_3 \rangle$. Layer-0 (L0) comprises physical/continuous effects (quantum/noise/coherence). Higher layers depend on but are not reducible to L0.

$$\forall \mathrm{L}_i \ (i \geq 1 \ \rightarrow \ Dep(\mathrm{L}_i, \mathrm{L}_0) \ \wedge \ \neg Red(\mathrm{L}_i, \mathrm{L}_0))$$

- **Interpretation:** A working decomposition used throughout:
  - **L0:** physical/continuous substrate (quantum, noise, coherence).
  - **L1:** abstract cognitive operations over semions.
  - **L2:** realization dynamics (sensorimotor loops, coding).
  - **L3:** morphology/material form.
- **Rationale:** Separates cognitive analysis from physicalist reductionism while acknowledging material grounding. Asserts compatibility of PC with arbitrary L0 (orthogonality of the core).

2.3. BC-03: Anti-entropy persistence [Postulatum Stabilitatis]

- **Statement:** Cognitive systems maintain operational stability against entropy through structuring, compression, and transformation of information; without this, cognition degrades and boundaries dissolve.
- Persistence condition:

$$\Pr[\partial X \text{ persists on } [t, t+\Delta]] \geq \theta, \ \Delta > 0, \ \theta > 0,$$

- requiring nontrivial energy flow:

$$\dot{F}_X < 0$$

- for an appropriate free-energy functional F_X.

  General form:

$$\forall \mathrm{S} \ (Subject(\mathrm{S}) \rightarrow \exists M \ (Stability(M, \mathrm{S}) \wedge Resist(M, Entropy))).$$

- **Rationale:** Models cognition as active maintenance of order, not passive flow.

---

3. Outline of the formal system

Cognitive triad (Trias Cognitiva)

- **Semion S:** a quantum of meaning; minimal discrete state. Modeled as a vector $\mathbf{s} \in \{0,1\}^n$ with physical realization and exergy.
- **Operation $\mathcal{O}$:** computational/energetic structure inside a subject X acting over S under rules $\mathrm{R}_{\mathrm{rel}}$.

- **Relation** R_{rel}**:** a family of rules $\rho: S \times S \to S$ (and more general maps) that delimit admissible operations and link states, operations, and outcomes.
- **Types of semions:** S_{past}, $S_{current}$, $S_{predicted}$, S_{error}.
- **Temporal petals:** (i) micro-dynamics; (ii) communicative; (iii) narrative.
- **Fundamental theorem (informal):** The unity of S, $\mathcal{O}$, and R_{rel} is necessary and sufficient for cognition.
- **Corollary (analogy):** S ≈ matter; $\mathcal{O}$ ≈ energy; R_{rel} ≈ information.

---

4. The Physical Genesis of Cognitive Primitives

The Cognitive Triad $\langle S, \mathcal{O}, R_{rel} \rangle$ is not merely assumed; it emerges from foundational physical and informational principles. Understanding their origins anchors the axiomatic system within a concrete material framework.

4.1. The Genesis of Semions (S): From Fluctuation to State

Cognition begins not with predefined symbols, but with the discretization of continuous physical flows—whether they arise from external sensory inputs or internal homeostatic signals.

A **semion** S is a physical state that has achieved **exergetic stability**, making it distinguishable from background fluctuations and allowing it to persist long enough ($\tau > \tau_{\min}$) to participate in computational processes.

This act of identifying stable, repeatable states within a stochastic field constitutes the first step of cognition.

> **Lemma *Emergentia Status***
>
> *Ex campo stochasticorum, status cum tempore vitae supra* $\tau_{\min}$ *fiunt semiones.*
>
> *(From a stochastic field, states with a lifetime above the minimum threshold become semions.)*

Formally captured in:

AX-DISCR-01: All cognitive processing begins with the quantization of phenomena into semions.

4.2. The Genesis of Operations ($\mathcal{O}$): A Minimal Computational Basis

Complex cognitive transformations are built from a **minimal set of primitive, physically realizable operations**, a basis postulated to be universal across substrates:

> **AX-OPER-BASIS — *Axiomatis Primitivorum Operandi***
>
> There exists a minimal set of primitive operations:
>
> $$\mathcal{O}_0 = \{\text{cmp}, \text{add}, \text{sub}\}$$

from which all cognitive transformations can be composed.

Physical analogues:

- **Comparison** (*cmp*): Threshold firing in neurons or non-linear activations (e.g., ReLU) in artificial networks.
- **Addition** (*add*): Excitatory summation of synaptic inputs.
- **Subtraction** (*sub*): Inhibitory synaptic inputs.

The complexity of higher-order cognition (e.g., attention, reasoning) does not arise from novel primitives, but from **massively parallel and hierarchical compositions** of $\mathcal{O}_0$.

> **Lemma *de Continuitate Compositionis***
>
> *Novi operatorum generes non oriuntur ex scala per se, sed compositionaliter ex* $\mathcal{O}_0$ *sub vinclis* R_{rel}.
>
> *(New types of operators do not arise from scale per se, but compositionally from the primitive basis, under the constraints of* R_{rel}.*)*

4.3. The Genesis of Relations (R_{rel}): From Interaction to Structure

The relational matrix R_{rel} is the **emergent physical topology of the substrate**, not an abstract table of connections.

It consists of stabilized pathways and constraints that determine which operations are possible or likely, shaped by **error minimization** and **energy optimization**. Over time, pathways leading to predictive success are reinforced.

> **Lemma *Topologia Relationum***
>
> *Ex fluctuationibus et margine systematis emergit structura relationum* R_{rel}*, quae determinat itineraria cognitiva intra subjectum.*
>
> *(From fluctuations and the system's boundary, the structure of relations* R_{rel} *emerges, which determines the cognitive pathways within the subject.)*

4.4. Cyclic Interdependence

The triad $\langle \text{S}, \mathcal{O}, \text{R}_{\text{rel}} \rangle$ forms a complete physical cycle:

1. Stable states S are identified.
2. They are transformed by physical operations $\mathcal{O}$.
3. The history of these interactions shapes R_{rel}.
4. R_{rel} in turn constrains and channels future operations.

---

5. Core axioms

The principles of material cognition are derived from a minimal set of foundational axioms grouped into: physical basis, dynamic/predictive architecture, and communicative grounding. Bracketed tags are used for reference in statements and proofs.

5.A. Physical and thermodynamic foundations

- **[AX-PHYS-01] Axiomatis Materialitatis Semionis (Semionic materiality).**
- Every semion is a physical, exergy-bearing state. No semion exists without a material substrate.

$$\forall \mathbf{s} \in \mathrm{S}:\ \exists\, substrate(\mathbf{s})\ \wedge\ E(\mathbf{s}) \geq E_{\min} > 0.$$

- **[AX-PHYS-02] Axiomatis Pretii Operandi (Operational cost).**
- Every cognitive operation has a thermodynamic cost bounded below by Landauer's limit for irreversible recording/erasure.

$$\forall O \in \mathcal{O}:\ Q(O) \geq k_B T \ln 2.$$

- **[AX-DISCR-01] Axioma Discretisationis (Discretization).**
- Cognition begins with the discretization of sensory/internal flows into semions, yielding a finite-dimensionally parameterizable set of stable, distinguishable states.

$$\Phi\text{:flows} \to \mathrm{S}.$$

5.B. Dynamic and predictive principles

- **[AX-PREDICT-01] Axioma Praedictionis (Prediction).**
- The system continually generates predicted states and minimizes prediction error.

$$\forall t:\ \ Cog(t) = argmin\ Error(Predict(t),\ Reality(t)).$$

- **[AX-HIERARCH-01] Axioma Hierarchiae Temporalis (Temporal hierarchies).**
- Cognitive processes unfold across nested temporal scales with cross-scale feedback (micro ↔ communicative ↔ narrative).

$$Cog = \bigcup_{i=0}^{n} Process_{\mathrm{scale}_i}, \quad \exists \{T_i\}:\ T_i \subset T_{i+1},\ f: T_i \to T_{i+1}.$$

- Scales (indicative):
  - **Micro-scale:** sensorimotor and physiological loops (approximately 10^{-3}–10^{0} s).
  - **Communicative scale:** interaction and synchronization (approximately 10^{0}–10^{2} s).
  - **Narrative scale:** long-term memory, goals, identity (approximately 10^{3}–10^{7} s).

- **[AX-ADAPT-01] Axioma Adaptationis per Errorem (Error-driven adaptation).**
- The relational structure $R_{rel,t} \subseteq S \times \mathcal{O} \times S$ evolves by minimizing predictive error under energetic/informational constraints.

$$R_{rel,t+1} = \text{Update}(R_{rel,t}, \text{Error}_t, \text{Constraints}).$$

5.C. Communicative principle

- **[AX-COMM-01] Axioma Communicationis (Public validity).**
- Cognitive validity and stability are achieved and maintained through communication; stable public representations are integral to persistence.

$$Validity(Cog) \Rightarrow Communicative\ Exchange(Cog).$$

---

6. Postulates (empirically motivated, not axioms)

- **[POS-SUBSTR-INV] Axioma Invariantiae Substrati (Substrate invariance).**
- Cognitive operations $\mathcal{O}$ are substrate-invariant in principle; only scale and efficiency vary with architecture.

$$\forall \mathcal{O} \in \mathcal{O}, \forall \text{Sub}_1, \text{Sub}_2 \colon feasible(\mathcal{O}, \text{Sub}_1) \Leftrightarrow feasible(\mathcal{O}, \text{Sub}_2).$$

- **Comment:** Treated as a postulate supported by cross-substrate evidence; targeted elevation to theorem via [TH-LAYER-ORTH].
- **[POS-OPER-BASIS] Axiomatis Primitivorum Operandi (Primitive set).**
- There exists a minimal set of primitives $\mathcal{O}_0 = \{\text{cmp,add,sub}\}$ from which all cognitive transformations can be compositionally generated.

$$\forall O \in \mathcal{O},\ \exists f(\mathcal{O}_0) \to O.$$

- **Comment:** A constructive postulate to be evaluated in the companion series on the genesis of operations.

---

7. Lemmas

- **[LEM-REFL-EXT-01] Lemma de Reflexione Externa (External reflection).**
- Architectures without intrinsic memory and closed predictive loops can be "closed" only via an external agent/environment (external reflexive loop).
    - **Depends on:** [AX-PREDICT-01], [AX-COMM-01].
- **LEM-PHANTOM-01 Lemma on Semantic Phantoms**
- **Formulation:** *The External Language of Meaning (ELM) can generate syntactically well-formed constructs that have no referent in the Metalanguage of Cognition (MLC).*

- **Description:** The projection $\mu: S \to \Sigma$ may create symbols for which there is no stable semantic mapping in the S-space. Such "phantoms" do not engage long-term linkages in R and are absent from the predictive dynamics. Their emergence is a by-product of the exposure mechanism and can distort cognitive exchange. **LEM-PHANTOM-01** prevents meaningless constructs from contaminating R_{rel}, ensuring that only structurally grounded patterns from the noise reservoir integrate into the cognitive matrix.
- **[LEM-COMP-01] Lemma de Continuitate Compositionis (Compositional genesis).**
- Novel classes of operations arise not from scale per se, but compositionally from $\mathcal{O}_0$ under constraints R_{rel}. Gradients/selection are canonical mechanisms.
    - **Status:** Outlined here; full treatment in ART-GEN-OPS-02.
- **[LM-DISCRET-01] Lemma of discretization.**
- Discrete symbolic elements (semiotic tokens) are necessary for stable operation of MLC; continuous signals alone cannot guarantee reproducibility.
- **[LM-SENS-01] Lemma Sensoria.**
- MLC structures are fed by sensory modalities but transform them into discrete symbolic patterns.
- **[LM-SYMB-01] Lemma Symbolica.**
- MLC requires symbolic recombination capacity to generate higher-order cognition beyond sensory input.
- **[LM-COMPRESS-01] Lemma of cognitive compression.**
- Automatization is achieved through reduction of symbolic weight in MLC, enabling faster but less flexible operations.

---

8. Theorems

T1. Temporal synchronization and the constructed "present"

- **[TH-LAG-01] Theorema Synchronicitatis Tardivae (Delayed synchrony).**
- The "present" is a construction chosen by predictive alignment between delayed input and forecasted futures:

$$O_{\text{present}}(t) = arg\min_{\tau \geq 0} d\big(\mathbf{s}_{in}(t-\delta),\, \mathbf{s}_{pred}(t+\tau)\big),$$

- where $d: X \times X \to \mathbb{R}_{\geq 0}$ is a cognitive-dissonance metric on the space of states, satisfying $d(x,y) = 0 \Leftrightarrow x = y$, symmetry, and the triangle inequality. If a normalized form is used (e.g. $d \in [0,1]$), this is stated explicitly. δ [s] is the input-assimilation latency.

- **Depends on:** [AX-PREDICT-01], [AX-HIERARCH-01], [AX-ADAPT-01].

- **[COR-LAG-W] Corollary (Width of now).**
- The width of the present w_{now} is a function of tolerances in d and adaptation speeds in $\mathrm{R}_{\mathrm{rel},t}$.
  - **Depends on:** [TH-LAG-01].

T2. Narrative extension and thresholds

- **[TH-NARR-LEV-01] Narrativum ut Lever Temporalis (Narrative lever).**
- ELM structures form temporal bridges that extend the effective "present":

$$Narr(\Delta T) = \max_{ELM} P_{survival}(\mathrm{S},\, t + \Delta T), \quad \Delta T \gg 0.$$

- **Depends on:** [AX-COMM-01], [AX-HIERARCH-01], [AX-PREDICT-01].
- **Roadmap:** Full development in ART-NARR-03.
- **[TH-NARR-THRESH-01] Theorema de Limine Narrativo (Narrative threshold).**
- There exists a threshold T such that if $complexity(ELM) > T$, the narrative loop activates and the systemic property reason emerges.
  - **Depends on:** [TH-NARR-LEV-01], [AX-ADAPT-01], [AX-COMM-01].
  - **Roadmap:** ART-NARR-03.

T3. Regulation

- **[TH-REGUL-DUAL-01] Theorema de Dualitate Regulationis (Dual regulation of language activity).**
- Regulation is distributed across two irreducible axes: internal (cognitive compression/expansion in MLC) and external (normative/communicative constraints in ELM). These levels recursively inform each other, balancing mechanism and model.
  - **Depends on:** [AX-PHYS-02], [AX-ADAPT-01], [AX-HIERARCH-01].

T4. Layering and invariance

- **[TH-LAYER-ORTH] Layer-0 orthogonality.**
- For fixed $(\mathrm{L}_1, \mathrm{L}_2, \mathrm{L}_3)$, variations in L_0 change efficiency constants while preserving partial isomorphisms of operations $\mathcal{O}$ and relations R.
  - **Depends on:** [AX-PHYS-01], [AX-PHYS-02], [AX-ADAPT-01], [POS-SUBSTR-INV], [BC-02].
  - **Purpose:** Path toward elevating [POS-SUBSTR-INV] to theorem.

T5. Persistence and anti-entropy

- **[TH-PERSIST-01] Theorem of persistence.**

- Cognitive patterns persist by maintaining semantic invariants across MLC↔ELM projections and across substrates, provided prediction error is minimized under energy budgets and communicative corrections are maintained.
- Formal/operational core:
    - If a system minimizes $d(\mathrm{S_{in}}, \mathrm{S_{pred}})$ under energy cost Q and maintains communicative correction ([AX-COMM-01]), then ∂X persists over horizon Δ with non-zero probability exceeding background decay.
    - Semantic invariant I_{sem} satisfies $I_{\mathrm{sem}}(x) \approx I_{\mathrm{sem}}(ELM \circ \cdots \circ MLC(x))$ within tolerance ϵ.
    - **Depends on:** [BC-03], [AX-PREDICT-01], [AX-PHYS-02], [AX-COMM-01], [TH-LANG-01] (for invariants).

T6. Language as dual projection

- **[TH-LANG-01] Theorem of dual projection of language.**
- Every linguistic system admits a dual projection: MLC (internal Language of Cognition) and ELM (external Expositional Language Medium). Full cognitive representation requires a bidirectional mapping MLC ⇄ ELM that preserves the structure of relevant relations.
    - **Proof sketch:** If cognition operated only in MLC, private structure would lack public calibration; if only in ELM, exposition would lack generative grounding. Duality ensures computability and communicability via bidirectional mapping with toleranced invariants.
    - **Depends on:** [AX-COMM-01], [AX-HIERARCH-01], [POS-SUBSTR-INV], [POS-OPER-BASIS].
    - **Corollary [COR-LANG-DECOUPLE]:** Partial decoupling of MLC and ELM (inner speech, automatized speech, aphasias) does not destroy the core $\mathcal{O}$; mappings can be re-established within bounds.
    - **Roadmap:** Full proof in ART-LANG-01.
- **TH-NOISE-01 Theorem of Cognitive Noise**
- **Formulation:** *Silentia est matrix potentialis.* — Latent cognitive activity, not actualised in the current vector of attention, forms a structured "noise" that serves as a reservoir of potential patterns.
- **Description:** Any activation that is not a direct generation from already consolidated automatic routes π_0 originates in a reconfiguration of this noise field. Non-linguistic and "dark" neuronal populations (Humphries, *The Spike*) operate as carriers of latent configurations, which can enter $\mathrm{R_{rel}}$ when context shifts. Noise is a necessary precondition for the emergence of insight, creative linkages, and new operator trajectories.
- **From prediction to adaptation: TH-NOISE-01** bridges the predictive core (**AX-PREDICT-01**) and adaptive mechanisms (**AX-ADAPT-01**), formalising where

novelty originates. **From composition constraints**: **LEM-COMP-01** bounds *how* new operators form; **TH-NOISE-01** specifies *where* their raw material comes from.

---

9. Operationalizations

9.1. Present-operators (operational definition and metrics)

- **Definition:** $\mathrm{O}_{\mathrm{present}}$ selects actions/updates from a constructed present rather than an instant:

$$\mathrm{O}_{\mathrm{present}}(t) = \underset{\tau \geq 0}{arg\,mind}\big(\mathbf{s}_{\mathrm{in}}(t-\delta),\, \mathbf{s}_{\mathrm{pred}}(t+\tau)\big),$$

- where d is a dissonance metric (e.g., KL, Hamming over semion bundles), δ is input-assimilation latency and δ' is exposition/actuation latency.
- **Width of the present:**

$$w_{\mathrm{now}} = max\{\tau : d \leq d_{\mathrm{min}} + \epsilon\}.$$

- **Measurement:**
  - **Input latency (δ):** cross-correlate sensed semions with model-state updates.
  - **Output latency (δ'):** correlate model commits with motor/ELM emission.
  - **Width (w_{now}):** span of τ values producing indistinguishable decisions (A/B within ϵ) under fixed context.
- **Predictions:**
  - **Under noise/energy scarcity ([AX-PHYS-02]):** w_{now} narrows.
  - **With improved model fidelity (lower expected d):** w_{now} widens.
  - **Cross-substrate ([POS-SUBSTR-INV]):** functional form is invariant; absolute latencies differ.

9.2. Dual operators (MLC↔ELM over time)

- **Mapping pipeline:**
  1. **Assimilation** A: $ELM \to MLC$.
  2. **Reconstruction** $\mathrm{O}_{\mathrm{recon}}$: $\mathcal{H} \to MLC$ (from memory/history).
  3. **Prediction** P: $MLC \to MLC$.
  4. **Exposition** E: $MLC \to ELM$.
- Composite operator:

$$\mathrm{O}_{\mathrm{td}} = \mathrm{E} \circ \mathrm{P} \circ \mathrm{O}_{\mathrm{recon}} \circ \mathrm{A}.$$

- **Errors:**
  - **Exposition error:** $e_E = d(E(MLC),\, ELM^{\mathrm{target}})$.
  - **Modeling error:** $e_M = d\big(\mathrm{P}(\mathrm{O}_{\mathrm{recon}}(\mathrm{A}(ELM))),\, MLC^{\mathrm{obs}}\big)$.

- **Invariant ([TH-PERSIST-01]):**
- A semantic invariant I_{sem} satisfies $I_{\text{sem}}(x) = I_{\text{sem}}(O_{\text{td}}(x))$ within ϵ.
- **Measurement exemplars:**
    - **Silicon:** A ≈ parsing, E ≈ decoding, O_{recon} ≈ state estimation, P ≈ latent dynamics; estimate e_E, e_M via held-out paraphrase/translation and latent prediction error.
    - **Biology:** A/E ≈ perception/utterance, O_{recon}/P ≈ working memory and predictive coding; estimate via neural decoding and behavioral variance.
- **Trade-offs under energy constraints ([AX-PHYS-02]):**
- Systems trade e_E vs. e_M: stronger compression in reconstruction saves energy but raises e_M. Robust cognition maintains I_{sem} despite bounded errors.

9.3. Metrology (lags, energetics, efficiency)

- **Lag metric:**

  Let

  $$lag\ [\text{s}] = \delta\ [\text{s}] + \delta'\ [\text{s}]$$

  with context-dependent distribution $p(lag \mid \text{task}, \text{noise})$. Here δ is the input-assimilation latency and δ' the exposition/actuation latency. The dissonance function $d: X \times X \to \mathbb{R}_{\geq 0}$ is a metric on the space of states, satisfying $d(x, y) = 0 \Leftrightarrow x = y$, symmetry, and the triangle inequality; if a normalized form is used (e.g. $d \in [0,1]$), this is stated explicitly. Estimate lag by cross-correlation; report $\mathbb{E}[lag]$ [s], variance [s^2], and context effects.

- **Energetics:**

  Logical work W_O [bit] as reduction in uncertainty or predictive loss:

  $$W_O = \Delta I(S)\ [\text{bit}] \quad \text{or} \quad W_O = \Delta KL\ [\text{bit}]$$

  Energy cost E_O [J] measured or bounded in Landauer units

  $$E_L = k_B T \ln 2\ [\text{J/bit}]$$

  If converting between energy and power, the unit factor $1\,\text{W} = 1\,\text{J/s}$ is stated explicitly.

- **Efficiency:**

  $$\eta_O = \frac{W_O\ [\text{bit}]}{E_O\ [\text{J}]/E_L\ [\text{J/bit}]} \in [0,1],$$

interpretable as bits of useful structure gained per Landauer-equivalent of expended energy.

Cross-substrate tests ($POS - SUBSTR - INV$):

Compare η_O across substrates: expect invariant relations with different absolute costs.

- **Predictions:**

  η_O increases when $\mathrm{R}_{\mathrm{rel}}$ matches task statistics ($AX - ADAPT - 01$), decreases with injected noise, and tends to anticorrelate with lag as optimization sharpens prediction.

---

10. Ontological corollaries (explanatory notes)

1. **Materialismus strictus:** sine substracto nulla cognitio; sine pretio energetico nullum O.

   All cognitive states/operations are materially instantiated and energetically priced ([AX-PHYS-01/02]).

   - **Falsifiability:** a genuine instance of cognition with no measurable substrate or energy flow would refute PC.
   - **Implication:** "Pure information" cognition reduces to an implementation claim at L3/L0.
2. **Exponibilitas:** ELM is the condition of public stability; valid cognition is communicable.

   Stable meaning requires externalization and inter-agent calibration ([AX-COMM-01], [TH-PERSIST-01]).

   - **Prediction:** Isolation induces semantic drift—observable as rising e_E and failure to preserve I_{sem}.
3. **Isomorphia micro↔macro:** delayed synchrony (micro) and narrative bridging (macro) manifest the same temporal asymmetry.

   Aggregated micro lags constrain feasible narrative windows ΔT; improving micro prediction widens ΔT at fixed energy.

   - **Measurement:** correlate w_{now} with narrative retention/goal maintenance across tasks and substrates.

---

11. Glossary of symbols

- **S:** set of semions (physically distinguishable, stable states).
- $\mathcal{O}$**:** set of operations (physical transformations over S).
- $\mathrm{R}_{\mathrm{rel},t} \subseteq \mathrm{S} \times \mathcal{O} \times \mathrm{S}$**:** relational structure at time t.
- $E(\mathbf{s})$**:** exergy of semion **s**; $E_{\min} > 0$ minimal threshold for discriminability/stability.

- $Q(O)$**:** energy/heat equivalent of operation $O \in \mathcal{O}$.
- $d(\cdot,\cdot)$**:** metric/divergence of cognitive dissonance.
- w_{now}**:** width of the constructed present.
- L_i**:** layers of architecture, $i \in \{0,1,2,3\}$.
- **MLC / ELM:** internal language of cognition / external expository medium.

---

12. Relationship to Existing Theories

A comparative synthesis situates the Principia Cognitia (PC) triad $\langle \mathrm{S}, \mathcal{O}, \mathrm{R}_{\mathrm{rel}} \rangle$ and the MLC–ELM duality against major symbolic, geometric, computational and representationalist accounts.

What is an "axiomatic theory of cognition"?

Think of it as a blueprint that spells out, in a small set of basic rules, how any thinking system must work – whether it's a human brain, an artificial neural network, or something built in the far future. The "axioms" are those basic rules: they don't change from case to case, but everything else in the theory follows from them.

Why this matters for AI and human minds

By reducing cognition to a handful of testable principles, we can compare very different systems on the same footing. This lets us see where AI truly resembles human thinking and where it differs in kind, not just degree. It also gives a way to design artificial minds with properties we actually want — like reliability, transparency, and the ability to work well with people.

Practical take-aways now

Even without building a full artificial mind, the axioms point to concrete tools: ways to measure how faithfully a system represents and processes information; checklists for what kinds of memory and error-correction it needs; and protocols to validate that two systems, human or machine, really "understand" the same thing. These can be put to work immediately in AI evaluation, cognitive science experiments, and human-AI team design.

12.1 Comparison Matrix

Theory	Units	Relations	Operations	Strengths	Limitations
Fodor's LOT	Symbols	Syntax	Production rules	Strong compositionality, clear syntax–semantics interface	No account of graded dynamics, learning delegated to ad hoc modules
Gärdenfors' Conceptual Spaces	Regions in $\mathbb{R}^n$	Topological overlap, distance	Geometric transforms	Intuitive mapping to perception, metric structure	Limited operational set, weak temporal dynamics
ACT-R /	Chunks, rules	Slot–filler	Production rules	Mature	Architecture-specif

Theory	Units	Relations	Operations	Strengths	Limitations
Soar		bindings, pattern matching		implementation, task-level modelling	ic, brittle outside trained domains
Enactivism	Actions/situations	Agent–environment coupling	No explicit $\mathcal{O}$	Rich phenomenology, embodiment focus	Lacks formal apparatus for operations or metrics
Predictive Processing	Model elements	Hierarchical prediction–error links	Bayesian updates, gradient descent	Neuro-computational plausibility, unifying principle	No discrete semion layer, ELM projection absent
PC Framework	Semions	Weighted graphs	Physically-realised primitives {cmp, add, sub}	Substrate-neutral, thermodynamically grounded, empirically testable	Requires experimental validation of primitives and invariance

12.2 Symbolic Logic vs. MLC–ELM

Category	MLC	ELM
Unit	Semion (vector $\mathbf{s} \in V$)	Symbol ($\in \Sigma$)
Links	Weighted graphs $R_{rel} \subset S \times S$	Syntax, rules
Representation	Vectorial, substrate-neutral	Discrete, mediative
Communication	Reconstruction on shared MLC	Transmission of symbols
Loss	None (within agent)	Unavoidable in μ-projection

12.3 Evaluation by Coherence and Falsifiability

Theory	Coherence	Falsifiability
LOT	High (symbolic)	Low (rule-list expansion)
Conceptual Spaces	Moderate (geometric)	Moderate (topology tests)
ACT-R/Soar	High (rule-based)	Low (architecture-bound)
Enactivism	Low (narrative)	Low (no protocol)
Predictive Processing	High (computational)	Moderate–high (neuro tests)
MLC–ELM	High (vector+syntax)	High (LLMs, EEG, task accuracy)

12.4 Convergence with Blaise Agüera y Arcas

Concept	PC	*What Is Intelligence?*
Fundamental principle	AX-PREDICT-01: continual prediction	Brain as predictive engine
Cognitive substrate	AX-PHYS-01: material semion	"Computation phase" of matter
Evolutionary driver	LEM-COMP-01: composition, diffusion	Symbiogenesis
Substrate invariance	POS-SUBSTR-INV	Functionalism ("artificial diamonds")
Sociality	Narrative threshold, collective cognition	Theory-of-Mind arms race
Intellectual lineage	Turing, Wiener, Wittgenstein	Turing, von Neumann

The overlap underscores a shared commitment to prediction, material grounding, substrate-neutralism, and social-evolutionary pressures, but PC adds a worked-out minimal operational basis and a formal semion–relation ontology.

12.5 Points of Contact with Prakash Mondal

Mondal's *The Puzzling Chasm…* sets out three "correspondence" conditions between cognitive representations (CR) and linguistic forms (PL). PC reframes these as generative rather than merely matching: both CR and PL emerge from the same primitive operations and relation-graphs, with μ and ν as concrete compilation/projection maps.

Aspect	Mondal	Redundancy	PC bridge	Integration/test
Ontology & aim	Micro→macro dynamics CR↔PL	Wide ontological net	Minimal substrate-neutral ontology: ops, memories, errors	Fix a minimal macrostate type and one micro-dynamic for trial
Linking operator	ξ(CF) ≡ PLR	Carrier unspecified	ν:CF→TPR compiler in WM	Align ξ with ν, compare outputs on same input
Units of analysis	Micro L→macro H; CR/PL as vector macrostates	Macro level descriptive only	Ops/composers, WM/LTM buffers, causal tracer	Operator+arguments+context tags as attractor labels
Dynamics	Symbolic via phase-space partition	No "grid" of measurable transitions	Discrete orchestration parse→CF→ν→TPR→execute	Parallel transition logs for both channels
Mapping condition	Equivalence classes manifest as CR and PL	Abstract, no metric	Dual labels from same trajectory	Mutual information CF↔TPR over states/transitions
Uniformity/stability	Macro-dynamics similar over micro-variants	No tolerances	Two implementations (spiking, RNN) under same scheduler	ε/δ-bounds on attractor distances, KL of kernels
Correspondence	Overlapping/identical attractor basins	No partition procedure	Split trajectories by "operator events" and "symbol snapshots"	Graph bisimulation ≥ τ overlap
Equivalence principle	$F(\langle T_i \rangle) \leftrightarrow P(T_i)$	Risk of hard-coding	Polytype compilation: one op sig→multiple NFs	Check type-sig isomorphism on compile
Compositionality	Extensional superposition + concatenation	Mixes language/thought procedures	PC combinators: sequence, choice, parallel, merge	Complexity growth: tree depth vs. program length
Memory/errors/fact-check	Out of scope	Validity drift risk	Error buffer, fact-check, provenance modules	Maintain equivalence under error injection/correction
Substrate neutrality	Declared micro↔macro	Bound to neuro-lexicon	Same protocol across carriers	Demonstrate ε/δ-invariance on two carriers
Replicability	"Better explains data" claim	No test-set	PC task-set: causative, path, TR–LM, quantifier	Preregister metrics and success thresholds

Mondal's conditions become, in PC, testable invariants with operational definitions, tolerances and proposed measurement protocols.

---

13. Conclusion

Principia Cognitia offers a formal framework for cognition that is independent of metaphysical speculation yet applicable across disciplines and substrates. By articulating a coherent set of axioms, it clarifies the dual nature of language, the role of discretization, and the relationship between symbol and phenomenon. Future work will extend the system toward empirical validation, computational implementation, and integration with ongoing debates in philosophy of mind and cognitive science. The aim is not finality, but the establishment of a generative starting point for cumulative progress. Furthermore, the role of **competition as a fundamental anti-entropic selection mechanism** warrants a separate axiomatic treatment. Preliminary analysis suggests that competition is not merely a social phenomenon but an ontological law governing the persistence of organized structures $\langle S, R_{rel}, \mathcal{O} \rangle$. This will be the subject of a forthcoming work, extending the *Principia Cognitia* framework to evolutionary and game-theoretic dynamics.

---

14. Acknowledgements

The author is deeply grateful to the researchers and engineers behind the development of large language models at OpenAI, Anthropic, xAI, Google, Perplexity AI, Moonshot AI, Alibaba/Meta, and Kunlun Tech. Their groundbreaking work made possible both this paper and the forthcoming *Principia Cognitia*, and inspired many of the formal insights herein. The author thanks the anonymous reviewers and editors at Cognitive Science for their consideration of related work on MLC/ELM duality, which provided valuable context for the present theoretical development.

---

15. References

15.1. Self-reference

1. Snow, A. (2025). The Dual Nature of Language: Metalanguage of Cognition and External Language of Meaning. Cognitive Science. [ART-LANG-01 — Manuscript under review].
2. Snow, A. (2025). The Genesis of Cognitive Operations. [ART-GEN-OPS-02 — in progress].
3. Snow, A. (2025). Defining the Narrative Threshold. [ART-NARR-03 — planned].
4. Snow, A. (2026). Principia Cognitia. [In progress].

15.2. Source of Proofs and Ideas

5. Agüera y Arcas, B. (2024). Computational life: How well-formed, self-replicating programs emerge from simple interaction. *arXiv*. https://arxiv.org/abs/2406.19108
6. Baars, B. J. (1988). *A cognitive theory of consciousness*. Cambridge University Press.
7. Baldwin, J. M. (1896). A new factor in evolution. *The American Naturalist*, *30*(354), 441-451. https://doi.org/10.1086/276408
8. Butz, M. V., & Kutter, E. F. (2017). *How the mind comes into being*. Oxford University Press.
9. Churchland, P. S., & Churchland, P. M. (1998). *On the contrary: Critical essays, 1987–1997*. MIT Press.
10. Clark, A., & Chalmers, D. J. (1998). The extended mind. *Analysis*, *58*(1), 7-19.
11. Deacon, T. W. (1997). *The symbolic species: The co-evolution of language and the brain*. W. W. Norton.
12. Dennett, D. C. (1995). *Darwin's dangerous idea: Evolution and the meanings of life*. Simon & Schuster.
13. Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127-138. https://doi.org/10.1038/nrn2787
14. Hinton, G. E., & Nowlan, S. J. (1987). How learning can guide evolution. *Complex Systems*, *1*, 495-502.
15. Humphries, M. (2022). *The spike: An epic journey through the brain in 2.1 seconds*. Princeton University Press.
16. Landauer, R. (1961). Irreversibility and heat generation in the computing process. *IBM Journal of Research and Development*, *5*(3), 183-191. https://doi.org/10.1147/rd.53.0183
17. Santos, M., Szathmáry, E., & Fontanari, J. F. (2015). A critique of the Baldwin effect. *PLoS Computational Biology*, *11*(8), e1004345. https://doi.org/10.1371/journal.pcbi.1004345
18. Shai, A. S., Marzen, S. E., Teixeira, L., Oldenziel, A. G., & Riechers, P. M. (2025). Transformers represent belief state geometry in their residual stream. *arXiv*. https://doi.org/10.48550/arXiv.2405.15943
19. Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience*, *5*(1), 42. https://doi.org/10.1186/1471-2202-5-42
20. Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, *s2-42*(1), 230-265. https://doi.org/10.1112/plms/s2-42.1.230
21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (Vol. 30).
22. Wiener, N. (1948). *Cybernetics: Or control and communication in the animal and the machine*. MIT Press.
23. Wittgenstein, L. (1953). *Philosophical investigations* (G. E. M. Anscombe, Trans.). Blackwell. (Original work published 1953)

15.3. Inspirational / Context

24. Agüera y Arcas, B. (2025). *What is intelligence? Lessons from AI about evolution, computing, and minds*. The MIT Press. https://mitpress.mit.edu/9780262049955/what-is-intelligence/
25. Descartes, R. (1641). *Meditations on first philosophy*.
26. Diderot, D. (1774). *Entretien d'un philosophe avec la maréchale de***.
27. Eliade, M. (1954). *The myth of the eternal return*. Princeton University Press.
28. Lao Tzu. (n.d.). *Tao Te Ching*.
29. Lévi-Strauss, C. (1955). The structural study of myth. *The Journal of American Folklore*, *68*(270), 428-444.
30. Newton, I. (1687). *Philosophiæ naturalis principia mathematica*.
31. Orwell, G. (1949). *Nineteen eighty-four*. Secker & Warburg.
32. Rousseau, J.-J. (1781). *Essai sur l'origine des langues*.
33. Whitehead, A. N., & Russell, B. (1910–1913). *Principia mathematica*. Cambridge University Press.
34. Wolfram, S. (2002). *A new kind of science*. Wolfram Media.
35. Zhuangzi. (n.d.). *Zhuangzi*.

Appendix A. *Axiomata Relationis*

Definitio A.1 (*Relatio Physica*)

Relatio R is a configuration of the physical substrate that determines and constrains transitions between semions through operations:

$$\mathrm{R_{rel}} \subseteq \mathrm{S} \times \mathcal{O} \times \mathrm{S}.$$

Lemma A.1 (*Emergentia Relationum*)

R is not given *a priori*, but instead emerges from a stochastic field of fluctuations:

$$\mathrm{R_{rel}} = \{(\mathrm{S}_i, \mathcal{O}, \mathrm{S}_j) | P(\mathrm{S}_j | S_i, \mathcal{O}) \gg 0 \wedge \tau(\mathrm{S}_j) > \tau_{\min}\}.$$

That is: only those transformations are retained which:

1. Have a transition probability significantly higher than random chance.
2. Produce states whose lifetime exceeds a defined minimal threshold.

Lemma A.2 (*Conditio Subiecti*)

For the stability of $\mathrm{R_{rel}}$, a **system margin** is required:

- The margin separates the system from the external environment.
- External states can influence transitions, but only those are retained that—within the margin—result in more stable structures.

Theorema A.1 (*Topologia Emergens*)

From stochastic fluctuations, through natural or physico-energetic selection, an **emergent topology of relations** arises:

$$\mathrm{R}_{\mathrm{rel}} = \lim_{t\to\infty} F(\mathrm{S}, \mathcal{O}, env_t)$$

where F is a selection function that retains only the links resistant to entropy.

Corollarium A.1 (*Trias Cognitiva Completa*)

Thus:

- S = discretized physical states (*semiones*),
- $\mathcal{O}$ = physical processes that transform them (*operationes*),
- R = the emergent network of stabilized transitions,

This triad forms the **foundation of material cognition**.

Appendix B. *Axiomata Vectorialia*

Definitio B.1 (*Semion Elementarius*)

Elementary semion s_e is a discretized physical state, distinctly recognizable and energetically sustainable.

Formally:

$$\mathbf{s}_e \in \mathrm{S}_{\mathrm{phys}}, \quad |\mathrm{S}_{\mathrm{phys}}| < \infty.$$

In other words, it belongs to the finite set of all possible physical states S_{phys}.

Definitio B.2 (*Semion Vectorialis*)

Vectorial semion s_v is a mathematical structure in a linear space $\mathbb{S}$, represented as a linear combination of elementary semions:

$$\mathbf{s}_v = \sum_{i=1}^{n} w_i \, \mathbf{s}_{e,i}, \quad w_i \in \mathbb{R}.$$

Thus, $\mathbf{s}_v$ is not a new physical state but rather a **mathematical description of the composition** of multiple $\mathbf{s}_e$ instances.

Lemma B.1 (*Algebra Semionum*)

Complex configurations of cognitive states are conveniently handled within a vector-space framework:

- **Addition**: $\mathbf{s}_v + \mathbf{s'}_v = \sum_i (w_i + w'_i)\mathbf{s}_{e,i}$
- **Scalar multiplication**: $\alpha\mathbf{s}_v = \sum_i (\alpha w_i)\mathbf{s}_{e,i}$

This establishes the linear spaces and operations of vectors in $\mathbb{S}$.

Lemma B.2 (*Transformata*)

For dynamic analysis, vectorial semions can be subjected to analytical operations (for example, Fourier transforms, convolutions, and matrix decompositions):

$$\mathcal{F}(\mathbf{s}_v)(\omega) = \sum_i w_i \, e^{-i\omega t_i}$$

Therefore, while the underlying basis is made of discrete physical states, the analysis itself is carried out in vector space.

Theorema B.1 (*Dualitas Physica–Mathematica*)

Every $\mathbf{s}_v$ can be reduced to an ensemble of $\mathbf{s}_e$, but scientific cognition and practice require the vectorial description in order to manage compositions, projections, and approximations effectively.

Formally:

$$\forall \mathbf{s}_v \in \mathbb{S}, \exists \{\mathbf{s}_{e,i}\}: \mathbf{s}_v = f(\{\mathbf{s}_{e,i}\}, \mathbf{w}).$$

This means: for any given vectorial semion, there exists a set of corresponding elementary semions and weights from which it is constructed.

Corollarium B.1 (*Analogia cum Operationibus*)

Just as every operation $O \in \mathcal{O}$ can be reduced to a composition of primitive operations (compare/add/subtract), so too every vectorial semion can be reduced to a collection of elementary semions.

This analogy establishes a **fundamental symmetry between $\mathcal{O}$ (operations) and $\mathbb{S}$ (states)** within the *Principia Cognitia* framework.

Appendix C. The "KilburnGPT" Gedankenexperiment

Appendix C.1 — Manchester Baby LLM:

Objective.

To demonstrate the *substrate invariance* of the cognitive operation $\mathcal{O}$ underlying transformer inference, we consider an extreme hypothetical: running a small, fixed-weight transformer of the class studied in Shai et al. (2025) on a large parallel cluster of

Manchester Baby computers (Kilburn et al., 1948). By mapping each primitive in the model to the minimal instruction set of the Baby, we ask whether any step is *in principle* uncomputable, and we quantify the performance and resource costs.

C.1.1. Historical substrate

The Manchester Small-Scale Experimental Machine ("Baby") was the first stored-program digital computer. Key characteristics relevant here:

- **Word size:** 32 bits; 32 words of Williams tube storage (≈ 1024 bits ≈ 128 B) per machine.
- **Instruction set:** integer add/subtract, conditional branch, bit shift, store/load.
- **Clock speed:** ~1 kIPS (instructions per second).
- **Implementation:** ≈ 300 vacuum tubes, Williams–Kilburn CRT memory.
- **Power draw:** ≈ 3.5 kW.
- A functional replica is maintained at the Museum of Science and Industry, Manchester.

C.1.2. Target model

We adopt the "toy" transformer from Shai et al.:

- **Architecture:** 4 layers, 1 attention head, $d_{model} = 64$, $d_k = 8$, feed-forward dimension 256, context length $L = 10$, vocabulary size 3.
- **Parameters:** ≈ 1.41×10^5 weights, stored at 16-bit fixed-point precision.
- **Operations per forward pass:** ≈ 1.4×10^6 multiply–accumulates (MACs), plus element-wise nonlinearities (ReLU, Softmax, LayerNorm) and small control overhead.

C.1.3. Mapping to Baby primitives

All model operations reduce to the Baby's native arithmetic and control:

- **Addition/subtraction:** direct.
- **Multiplication:** shift-and-add algorithm on fixed-point words (≈ 80 instructions per 16-bit MAC).
- **Division and reciprocal square root:** iterative methods (restoring division, Newton–Raphson).
- **Exponential:** polynomial or table-lookup approximation with linear interpolation.
- **Comparison/max:** conditional branch and subtraction.
- **Vector–matrix products:** distributed across rows (systolic pattern) for parallel execution.

No model primitive is uncomputable on the Baby instruction set; all non-trivial functions can be emulated, albeit slowly.

C.1.4. Resource estimates

Memory footprint

- **Weights:** 0.283 MiB (≈ 2.26 Mbit) ⇒ ≈ 2210 Babies for storage alone (128 B per machine).
- **Activations/buffers/code:** additional ≈ 50 KiB ⇒ total ≈ 2800–3200 Babies as "RAM nodes".

Compute time

We distinguish two scenarios:

- **S0 (single Baby, purely sequential):**
- 1.4×10^6 MAC × 80 instr/MAC ≈ 1.12×10^8 instructions, plus non-linearities ⇒ ≈ $1.5 - 2.5 \times 10^8$ instructions.
- At 1 kIPS: ≈ $1.7 - 2.9$ days per forward pass.
- **S1 (ideal parallelism across ~900 compute-dedicated Babies, plus memory nodes):**
- Row-parallel multiply–accumulate stages dominate: the feed-forward block's 256×64 and 64×256 multiplications execute in ≈ 280–300 s/layer.
- Four layers plus overhead ⇒ ≈ 24–30 minutes per pass.

Energy and components (S1)

- **Cluster size:** 3500–4000 Babies (compute + memory).
- **Vacuum tubes:** ≈ $1.1 - 1.2 \times 10^6$.
- **Power draw:** ≈ $12 - 14$ MW.
- **Energy per pass:** ≈ $4 - 7$ MWh.

C.1.5. Interpretation

From a computability standpoint, the operation is invariant: every step in the transformer's forward pass can be expressed in the Baby's primitive instruction set. There are **no** cognitive primitives that "break" on this substrate. The differences are purely in quantitative resources:

- **Latency gap:** tens of minutes (S1) or days (S0) versus milliseconds on modern GPUs.
- **Energy gap:** megawatt-hours versus joules per inference.
- **Space gap:** thousands of rack-sized units versus a single chip.

This yields a concrete, falsifiable statement: altering only the substrate from modern silicon to 1948 CRT-and-valve logic changes *efficiency* by many orders of magnitude, without altering the formal structure of the cognitive operation.

Note. In principle, S0 could be physically enacted on the operational Baby replica in Manchester; the runtime in days per pass is prohibitive, but the execution is well within the machine's instruction set and storage constraints when using external media for weight paging.

C.1.6. References

1. Kilburn, T., Williams, F. C., & Tootill, G. C. (1948). Report on the Manchester Small Scale Experimental Machine.
2. Shai, A. S., Marzen, S. E., Teixeira, L., Oldenziel, A. G., & Riechers, P. M. (2025). Transformers represent belief state geometry in their residual stream. *arXiv:2405.15943*.

Appendix C.2 — Reliability-aware cluster operation: "KilburnGPT" under component failures

Objective.

We extend the substrate-invariance argument by modeling an operational transformer inference cluster built from Manchester Baby–class machines when components fail in realistic ways (vacuum tube burn-out, Williams tube drift and phosphor wear). We size hot spares, maintenance capacity, and scheduling policies to sustain target throughput with probabilistic guarantees, and quantify the resulting time/energy overheads. This complements Appendix C.1 (ideal reliability) with an availability-constrained regime.

C.2.1. Failure modes and simplifying assumptions

We model failures at the machine level as the superposition of independent component hazards in the "useful-life" phase (constant hazard per component). The following abstraction captures the dominant mechanisms without overfitting scarce archival data.

- **Baby bill of materials per machine:**
  - **Vacuum tubes:** ≈300 small-signal/power valves.
  - **Williams tube memory:** 1 storage CRT (operationally fragile; subject to drift, re-biasing, and eventual wear).
- **Failure processes:**
  - **Vacuum tubes:** random failure of heater/cathode/grid/plate; treat with constant hazard $\lambda_{valve} = 1/MTTF_{valve}$.
  - **Williams tube:** treat as a random "functional outage" that requires swap-and-retune; hazard $\lambda_{WT} = 1/MTTF_{WT}$.
- **Machine failure rate:** assuming component independence,

$$\lambda_{machine} = 300 \cdot \lambda_{valve} + \lambda_{WT}$$

- **Repair times:** two repair classes with mean times
  - **Valve replacement + checkout:** $MTTR_{valve} \in [0.25, 1]\, h$ (swap, rebias, smoke test).

- **Williams tube swap + retune:** $MTTR_{WT} \in [2,\ 6]\ h$ (mechanical replacement, bias/geometry tuning, regeneration checks).
- **Average MTTR (per-machine):**

$$MTTR = p_{valve} \cdot MTTR_{valve} + p_{WT} \cdot MTTR_{WT}$$

- where $p_{valve} = \frac{300\lambda_{valve}}{\lambda_{machine}}, p_{WT} = 1 - p_{valve}.$
- **Availability envelope (scenarios):**
  - **Optimistic:** $MTTF_{valve} = 10{,}000\ h,\ MTTF_{WT} = 2{,}000\ h,\ MTTR_{valve} = 0.5\ h,\ MTTR_{WT} = 4\ h.$
  - **Nominal:** $MTTF_{valve} = 5{,}000\ h,\quad MTTF_{WT} = 1{,}000\ h,\quad MTTR_{valve} = 0.5\ h,\ MTTR_{WT} = 4\ h.$
  - **Pessimistic:** $MTTF_{valve} = 2{,}000\ h,\quad MTTF_{WT} = 500\ h,\quad MTTR_{valve} = 1\ h,\ MTTR_{WT} = 6\ h.$

These ranges reflect the known fragility of Williams tubes and substantial spread in vacuum-tube life depending on de-rating, thermal management, and usage cycles.

C.2.2. Queueing model for spares, staffing, and uptime

We seek to keep a target number M of machines "online" (operational set) while failures and repairs occur. Let N be the total installed machines (operational + in repair), and let $\Lambda = N \cdot \lambda_{machine}$ be the aggregate failure arrival rate. Under a standard infinite-server repair approximation (failures independent; repairs begin immediately and proceed in parallel), the number of machines in repair N_{rep} is Poisson with mean

$$\mathbb{E}[N_{\text{rep}}] = \Lambda \cdot MTTR = N \cdot \lambda_{machine} \cdot MTTR$$

and variance $Var[N_{\text{rep}}] \approx \mathbb{E}[N_{\text{rep}}]$. To achieve an operational set M with high probability $1 - \alpha$, dimension spares via a normal tail bound:

$$N = M + \mathbb{E}[N_{\text{rep}}] + z_{1-\alpha}\sqrt{\mathbb{E}[N_{\text{rep}}]}$$

where $z_{1-\alpha}$ is the standard normal quantile (e.g., $z_{0.999} \approx 3.09$, $z_{0.99999} \approx 4.27$).

- **Technician-hours per hour:** expected labor intensity $L = \Lambda \cdot MTTR$ (machine-hours of repair per hour). With each technician contributing 1 hour of repair per wall-hour, the steady-state staff demand is $\lceil L \rceil$ per shift.
- **Spare parts consumption:** expected component replacements per hour: $300N\lambda_{valve}$ valves/hour and $N\lambda_{WT}$ Williams tubes/hour.

This "birth–death with instantaneous routing" model slightly understates congestion when repair capacity is finite; we therefore report both the mean and a conservative +z-sigma reserve.

C.2.3. Two cluster sizes under failures

We reuse the two inference targets from Appendix C.1: the toy cluster sized to run the Shai-class model in ≈30 minutes per pass, and a larger cluster for a mid-sized 4-layer transformer.

A) Toy "Shai-class" KilburnGPT (target online M ≈ 4 000 machines)

- From Appendix C.1, ≈3.5–4.0 k machines are needed for compute + RAM nodes at ideal reliability. We set $M = 4{,}000$.

Compute the hazard, case by case.

- **Optimistic:**
  - $\lambda_{machine} = 300/10{,}000 + 1/2{,}000 = 0.03 + 0.0005 = 0.0305\ h^{-1}$.
  - Composition: $p_{valve} \approx 0.9836$, $p_{WT} \approx 0.0164$.
  - $MTTR \approx 0.9836 \cdot 0.5 + 0.0164 \cdot 4 \approx 0.566\ h$.
  - With $N \approx M$ at first pass: $\mathbb{E}[N_{\text{rep}}] \approx 4{,}000 \times 0.0305 \times 0.566 \approx 69\ machines$.
  - Dimensioning at $z = 3.1$: $z\sqrt{\mathbb{E}[N_{\text{rep}}]} \approx 3.1 \times 8.3 \approx 26$.
  - **Installed** $N \approx 4{,}000 + 69 + 26 = 4{,}095$ (≈2.4% overhead).
  - **Throughput of failures:** $\Lambda \approx N\lambda_{machine} \approx 125\ failures/h$.
  - **Staffing:** $L = \Lambda \cdot MTTR \approx 71\ tech\text{-}hours/h$ (≈71 technicians per shift).
  - **Spares burn rate:** valves ≈ $300N/10{,}000 \approx 123\ valves/h$; Williams tubes ≈ $N/2{,}000 \approx 2\ tubes/h$.
- **Nominal:**
  - $\lambda_{machine} = 300/5{,}000 + 1/1{,}000 = 0.06 + 0.001 = 0.061\ h^{-1}$.
  - $p_{valve} \approx 0.9836$, $p_{WT} \approx 0.0164$ (same ratio).
  - $MTTR \approx 0.566\ h$ (same times).
  - $\mathbb{E}[N_{\text{rep}}] \approx 4{,}000 \times 0.061 \times 0.566 \approx 138$. Margin $z\sqrt{138} \approx 36$.
  - **Installed** $N \approx 4{,}174$ (≈4.3% overhead).
  - **Failures:** $\Lambda \approx 255\ /h$.
  - **Staffing:** $L \approx 144\ tech\text{-}hours/h$.
  - **Spares:** ≈ $300N/5{,}000 \approx 251\ valves/h$; $N/1{,}000 \approx 4.2\ tubes/h$.
- **Pessimistic:**
  - $\lambda_{machine} = 300/2{,}000 + 1/500 = 0.15 + 0.002 = 0.152\ h^{-1}$.
  - $p_{valve} \approx 0.9870$, $p_{WT} \approx 0.0130$.
  - $MTTR \approx 0.9870 \cdot 1 + 0.0130 \cdot 6 \approx 1.065\ h$.
  - $\mathbb{E}[N_{\text{rep}}] \approx 4{,}000 \times 0.152 \times 1.065 \approx 647$. Margin $z\sqrt{647} \approx 79$.
  - **Installed** $N \approx 4{,}726$ (≈18% overhead).
  - **Failures:** $\Lambda \approx 718\ /h$.
  - **Staffing:** $L \approx 765\ tech\text{-}hours/h$.
  - **Spares:** ≈ $300N/2{,}000 \approx 709\ valves/h$; $N/500 \approx 9.5\ tubes/h$.

Interpretation: even under optimistic life and fast swaps, a 4 k-node Baby cluster requires continuous high-intensity maintenance and a modest (2–18%) hot-spare pool to keep $M = 4{,}000$ online. The dominant operational burden is technician labor and spares logistics, not just energy.

B) Mid-sized "4-layer, d=256" KilburnGPT (target online M ≈ 45 000 machines)

Using the larger model of Appendix C.1, assume $M = 45{,}000$.

- **Optimistic:**
  - $\lambda_{machine} = 0.0305\,h^{-1}$, $MTTR = 0.566\,h$.
  - $\mathbb{E}[N_{\text{rep}}] \approx 45{,}000 \times 0.0305 \times 0.566 \approx 777$. Margin $z\sqrt{777} \approx 86$.
  - **Installed** $N \approx 45{,}863$ (≈1.9% overhead).
  - **Failures:** $\Lambda \approx 1{,}399\ /h$.
  - **Staffing:** $L \approx 792\ tech\text{-}hours/h$.
  - **Spares:** ≈ 1,377 $valves/h$ and 23 WT/h.
- **Nominal:**
  - $\lambda_{machine} = 0.061\,h^{-1}$, $MTTR = 0.566\,h$.
  - $\mathbb{E}[N_{\text{rep}}] \approx 1{,}553$. Margin ≈ 122.
  - **Installed** $N \approx 46{,}675$ (≈3.7% overhead).
  - **Failures:** $\Lambda \approx 2{,}847\ /h$.
  - **Staffing:** $L \approx 1{,}612\ tech\text{-}hours/h$.
  - **Spares:** ≈ 2,801 $valves/h$ and 46 WT/h.
- **Pessimistic:**
  - $\lambda_{machine} = 0.152\,h^{-1}$, $MTTR = 1.065\,h$.
  - $\mathbb{E}[N_{\text{rep}}] \approx 7{,}309$. Margin ≈ 265.
  - **Installed** $N \approx 52{,}574$ (≈17% overhead).
  - **Failures:** $\Lambda \approx 7{,}990\ /h$.
  - **Staffing:** $L \approx 8{,}509\ tech\text{-}hours/h$.
  - **Spares:** ≈ 7,886 $valves/h$ and 105 WT/h.

Interpretation: the hot-spare percentage remains modest, but absolute maintenance demand scales linearly with cluster size and rapidly dominates any throughput gains. In the pessimistic regime, continuous operation becomes logistically implausible.

C.2.4. Scheduling and fault tolerance for inference

Even with hot spares, mid-computation dropouts are inevitable. We therefore adopt three policies that preserve correctness while bounding recomputation:

- **Task tiling with retry:** partition each matrix–vector into tiles that complete in τ seconds. On node loss, only the in-flight tile recomputes elsewhere. Choose τ so that the probability of any node in the tile cohort failing within τ is < ϵ, i.e.,

$$\mathbb{P}[\text{tile failure}] \approx 1 - exp(-n_{tile}\lambda_{machine}\tau) \leq \epsilon$$

- yielding $\tau \leq \frac{-ln(1-\epsilon)}{n_{tile}\lambda_{machine}}$.
- **Erasure-coded partial sums:** within each reduce tree, compute per-row partials with a $(k + r,\ k)$ code; tolerate up to r simultaneous row-worker failures without recomputation.
- **Graceful degradation:** if the hot-spare pool drops below a threshold, reduce model width (disable a fraction of rows in W_1, W_2) and rescale activations to maintain output norms. This preserves $\mathcal{O}$'s structure while trading accuracy for availability.

These policies keep logical $\mathcal{O}$ invariant; only latency and energy per inference change.

C.2.5. Practical notes

- **Calibration downtime (Williams tubes):** beyond random failures, expect scheduled retuning windows (hours to days cadence). Model as planned downtime; it increases $\mathbb{E}[N_{\text{rep}}]$ linearly with calibration duty cycle.
- **Spares logistics:** daily spares demand in the nominal toy case is thousands of valves and a few dozen Williams tubes. Inventory, acceptance testing, and pre-biased subassemblies are essential to keep $MTTR$ low.
- **Staffing model:** the "infinite-server repair" approximation provides a lower bound on technicians. Finite repair bays increase queueing, effectively inflating $MTTR$ and thus the spare pool.
- **Energy footprint:** adding hot spares increases installed base and hence idle draw; for Babies, idle power is close to peak. Expect ≈(2–18)% energy overhead from spares alone, plus pragmatic increases from retries.

C.2.6. Conclusion

- Under realistic component failure rates, a Baby-based inference cluster can, in principle, maintain a fixed online set M with a modest hot-spare fraction and appropriate scheduling. However, the operational burden (spares consumption, technician labor, retries) is enormous and scales linearly with M. The core conclusion survives the harsher conditions: the cognitive operation $\mathcal{O}$ remains substrate-invariant; only efficiency and availability engineering become the story.

Appendix C.3 — Costing "KilburnGPT" in 1948 GBP under component failures

Objective.

We translate the reliability-aware cluster from Appendix C.2 into 1948-denominated operating costs per generated token. We combine (i) energy, (ii) spares consumption (valves and Williams tubes), and (iii) technician labour during runtime, under the same optimistic/nominal/pessimistic failure envelopes used earlier. All prices are stated explicitly as 1948 assumptions; results are provided as formulas and as worked examples for the toy and mid-sized clusters.

> *Disclaimer.* All historical energy-cost estimates presented here are approximate. They use contemporary prices and performance data for components such as vacuum tubes, which vary significantly across manufacturers and epochs. Ancillary costs — including the reliability of other components, facility rental, cooling requirements, and non-operational overhead — are omitted for clarity, and actual operational costs may have differed substantially.

C.3.1. 1948 price assumptions and symbols

These span credible ranges for U.K. industrial procurement in 1948; users can substitute archival figures without changing the model.

- Energy price per kWh: c_e in £/kWh
- — Working range: 0.003–0.006 (£/kWh) ≈ 0.7–1.5 d per kWh.
- Valve unit price: c_v in £/valve
- — Working range: 0.2–0.5 £ (bulk purchase, common small-signal valves).
- Williams tube unit price: c_WT in £/tube
- — Working range: 200–600 £ (special CRT with pickup plate and drive; Whirlwind-class devices were very costly).
- Technician labour (skilled maintenance): c_L in £/hour
- — Working range: 0.2–0.5 £/h (44-hour week pay bands for skilled technical staff); senior engineer hours, when needed, can be priced separately (e.g., 0.6–1.0 £/h), but maintenance headcount dominates.
- Exchange and accounting: all inputs and outputs in 1948 GBP; no inflation or PPP adjustments.
- Notation carried from C.1–C.2:
- — Power P (MW), runtime T (h), energy E = P·T (MWh).
- — Installed machines N, target online machines M.
- — Failure intensities per hour for valves and Williams tubes: $N\cdot(300/MTTF_v)$ and $N\cdot(1/MTTF\_WT)$.
- — Expected spares per hour (nominal): $S_v = 300N/MTTF_v$, $S\_WT = N/MTTF\_WT$.
- — Technician load per hour L (tech-hours/h) from C.2.

Cost per pass (L-token forward) is:

- Energy: $C_e = E\cdot(1000\cdot c_e)$
- Spares: $C\_sp = T\cdot(S_v\cdot c_v + S\_WT\cdot c\_WT)$
- Labour: $C\_L = T\cdot(L\cdot c\_L)$
- Total per pass: $C\_pass = C_e + C\_sp + C\_L$
- Cost per token (amortized over L tokens): $C\_tok = C\_pass / L$

(When using incremental decoding with KV-caching, per-token cost decreases sublinearly with L; our totals are conservative, dividing pass-level cost by L.)

C.3.2. Reliability envelopes (reused from C.2)

- Optimistic: $MTTF_v$=10,000 h; MTTF_WT=2,000 h; $MTTR_v$=0.5 h; MTTR_WT=4 h.
- Nominal: $MTTF_v$=5,000 h; MTTF_WT=1,000 h; $MTTR_v$=0.5 h; MTTR_WT=4 h.
- Pessimistic: $MTTF_v$=2,000 h; MTTF_WT=500 h; $MTTR_v$=1 h; MTTR_WT=6 h.

Hot-spare sizing modestly increases N above M; we reuse the N values computed in C.2 for each case.

C.3.3. Toy "Shai-class" KilburnGPT (M≈4,000 online; N from C.2)

Configuration recap (Appendix C.1): 4 layers, d_model=64, head_dim=8, MLP=256, L=10.

Runtime and energy (ideal orchestration): P≈12–14 MW, T≈0.4–0.5 h ⇒ E≈4–7 MWh per pass.

For nominal reliability (C.2): N≈4,174; spares per hour S_v≈251 valves/h, S_WT≈4.2 tubes/h; technician load L≈144 tech-hours/h.

Worked costs (use mid-range prices unless noted): c_e=£0.0045/kWh, c_v=£0.30, c_WT=£500, c_L=£0.30.

- Energy: E≈5.5 MWh ⇒ C_e ≈ 5,500 kWh × £0.0045 ≈ £24.8
- Spares per hour: $S_v\cdot c_v$ ≈ 251×0.30 = £75.3; $S\_WT\cdot c\_WT$ ≈ 4.2×500 = £2,100
- — Runtime T≈0.5 h ⇒ C_sp ≈ 0.5×(£75.3+£2,100) ≈ £1,087.7
- Labour: $L\cdot c\_L$ ≈ 144×0.30 = £43.2/h ⇒ C_L ≈ 0.5×£43.2 = £21.6
- Total per pass (L=10): C_pass ≈ £24.8 + £1,087.7 + £21.6 ≈ £1,134.1
- Cost per token: C_tok ≈ £113.4 (1948 GBP)

Sensitivity bands:

- Using optimistic reliability (N≈4,095; S_v≈123/h; S_WT≈2.0/h; L≈71): C_tok ≈ £54–£70 (energy/labour minor; WT failures dominate).
- Using pessimistic reliability (N≈4,726; S_v≈709/h; S_WT≈9.5/h; L≈765; T≈0.5 h):
- — Spares cost ≈ 0.5×(709×0.30 + 9.5×500) ≈ 0.5×(£212.7 + £4,750) ≈ £2,481
- — Labour ≈ 0.5×(765×0.30)=£114.8; energy ≈ £25–£35

- — C_tok ≈ (£2,620–£2,650)/10 ≈ £262–£265 per token.

Interpretation: in 1948 pounds, even the toy model yields tens to low hundreds of pounds per token, overwhelmingly driven by Williams tube attrition in the nominal/pessimistic regimes.

C.3.4. Mid-sized 4-layer model (M≈45,000 online; N from C.2)

Configuration recap (Appendix C.1): d_model=256, head_dim=64, MLP=1024, L=128.

Runtime and energy (ideal orchestration): P≈130–150 MW, T≈16–20 h ⇒ E≈2.2–2.8 GWh per pass.

For nominal reliability (C.2): N≈46,675; S_v≈2,801 valves/h; S_WT≈46 tubes/h; L≈1,612 tech-hours/h.

Worked costs (same prices as above):

- Energy: E≈2.5 GWh ⇒ C_e ≈ 2,500,000 kWh × £0.0045 ≈ £11,250
- Spares per hour: $S_v \cdot c_v$ ≈ 2,801×0.30 = £840.3; S_WT·c_WT ≈ 46×500 = £23,000
- — Runtime T≈18 h ⇒ C_sp ≈ 18×(£840.3+£23,000) ≈ 18×£23,840.3 ≈ £429,125
- Labour: L·c_L ≈ 1,612×0.30 = £483.6/h ⇒ C_L ≈ 18×£483.6 ≈ £8,705
- Total per pass (L=128): C_pass ≈ £11,250 + £429,125 + £8,705 ≈ £449,080
- Cost per token: C_tok ≈ £3,508 (1948 GBP)

Sensitivity bands:

- Optimistic reliability (fewer failures; similar energy): C_tok ≈ £1,800–£2,200.
- Pessimistic reliability (N≈52,574; S_v≈7,886/h; S_WT≈105/h; L≈8,509/h; T≈18 h):
- — Spares ≈ 18×(7,886×0.30 + 105×500) ≈ 18×(£2,365.8 + £52,500) ≈ £988,146
- — Labour ≈ 18×(8,509×0.30)=£45,950; energy ≈ £11–£18k
- — C_pass ≈ ~£1.05M ⇒ C_tok ≈ ~£8,200.

Interpretation: at scale, Williams tube replacements dominate total cost by one to two orders of magnitude over electricity and labour, even with generous energy pricing and low technician wages.

C.3.5. What drives the cost (and how to reduce it)

- Dominant term: C_sp from Williams tubes. Improving MTTF_WT (better derating, improved biasing, scheduled retuning that truly lifts effective life) or lowering c_WT (in-house tube shop, reuse/repair) is the only credible route to sub-£100/token costs, even in small clusters.
- Secondary levers: energy price and labour rates are small contributors at these scales; even halving c_e or c_L shifts totals by only a few percent.
- Architectural lever: reduce Williams tubes per online machine (e.g., offload storage to mercury delay lines or drum memory). This preserves computability but changes

the physical memory technology underlying the Baby—useful as an ablation demonstrating that the cognitive operation's form is substrate-invariant while the economics are not.

C.3.6. Summary statement (1948 GBP)

- Toy cluster (L=10): ~£50–£260 per token across optimistic→pessimistic reliability; nominal ≈ £110/token.
- Mid-sized cluster (L=128): ~£1.8k–£8.2k per token; nominal ≈ £3.5k/token.
- Cost composition at nominal: 90–97% spares (chiefly Williams tubes), 2–8% energy, 1–3% labour.

These magnitudes complete the substrate-invariance argument with a blunt economic coda: in 1948, the cognitive operation is the same, but its price is almost entirely the price of keeping a pre-core-memory substrate alive long enough to finish the thought.

Appendix D: Experimental Protocols for the Validation of Principia Cognitia

This appendix outlines three falsifiable experimental protocols designed to test the core axioms and theorems of *Principia Cognitia* (PC). These protocols move the theory from a formal framework to an empirical research program, providing concrete methods for validating its claims in computational environments.

D.1. The MLC Primacy Experiment (MPE-1)

Objective: To empirically validate the **Theorem of Decoupling of Languages (`TH-LANG-04`)**, which posits that the performance of inter-agent communication is fundamentally bounded by the alignment of their internal **Metalanguage of Cognition (MLC)**, not by the richness of their **External Language of Meaning (ELM)**. This experiment is designed to demonstrate that increasing the descriptive power of ELM yields diminishing returns without a corresponding alignment in the agents' internal world models (MLC).

Methodology:

1. **Environment - The "Asymmetric World":**
    - A 2D physics simulation (e.g., using PyMunk/Box2D) is created. The world contains objects with non-trivial properties (mass, friction, elasticity) and is governed by consistent physical laws (gravity, collisions). The state of the world is defined by the coordinates, velocities, and properties of its objects.
2. **System Architecture:**
    - **Agent A ("Oracle"):** A multimodal neural network (e.g., CNN encoder for visual input + LSTM for dynamics) with complete, real-time visual access to the simulation. Its MLC is a high-fidelity internal model of the world's

physics, learned through direct observation. Its task is to observe the simulation and generate textual descriptions (ELM) of events.

- **Agent B ("Reconstructor"):** A transformer-based language model that has **no** visual access to the simulation. Its only input is the ELM stream from Agent A. Its task is to parse these descriptions and predict a future state of the world (e.g., the coordinates of a specific object after `T` timesteps). This prediction serves as a proxy for successful semion reconstruction.

3. **Experimental Procedure:** The experiment is conducted in two parallel branches, using the same set of simulation scenarios.
    - **Branch 1 - ELM Scaling:** The MLC of Agent B is kept naive (e.g., pre-trained on generic text corpora, but not on the specific physics of the Asymmetric World). The complexity of the ELM generated by Agent A is systematically increased across trials:
        - *Level 1 (Basic ELM):* "Red circle moved right."
        - *Level 2 (Detailed ELM):* "The heavy red circle moved right at high speed and approached the blue square."
        - *Level 3 (Hyper-Detailed ELM):* Real-time stream of object properties, coordinates, and vectors.
    - **Branch 2 - MLC Alignment:** The ELM is fixed at a medium level of detail (Level 2). The MLC of Agent B is progressively aligned with Agent A's by pre-training it on the physics of the Asymmetric World. This can be achieved by allowing Agent B to observe and learn from the simulation directly, building its own internal model before the communication task begins. Alignment is varied from 0% (naive) to 100% (trained on the same data distribution as Agent A).

Prediction from Principia Cognitia: * In **Branch 1**, the prediction accuracy of Agent B will rapidly plateau. Beyond a certain point, additional ELM complexity will provide negligible performance gains because Agent B lacks the internal model (MLC) to ground the meaning of the symbols. * In **Branch 2**, the prediction accuracy of Agent B will show a strong, near-linear positive correlation with the degree of MLC alignment.

Metrics and Success Criteria: * **Primary Metric:** Mean Squared Error (MSE) between the predicted future coordinates and the ground-truth coordinates from the simulation. * **Success Criterion:** A successful validation requires demonstrating that the performance improvement (reduction in MSE) per unit of "information gain" is significantly higher in Branch 2 (MLC alignment) than in Branch 1 (ELM scaling) after an initial saturation point.

Falsification Condition: The theorem `TH-LANG-04` would be falsified if the performance of Agent B in Branch 1 continues to improve significantly with increasing ELM complexity, approaching the performance levels achieved through direct MLC alignment in Branch 2.

Required Resources: A multi-GPU compute environment. Python, PyTorch/TensorFlow, and a 2D physics library (PyMunk).

D.2. The Substrate Invariance Test (SIT-1)

Objective: To provide a direct, physical demonstration of the **Axioma Invariantiae Substracti (`AX-SUBSTR-INV`)**. This axiom states that a cognitive operation (()) is an abstract, formal structure (Layer I) whose logical outcomes are independent of the physical substrate (Layer 0/III) on which it is executed. The experiment will show that while the physical costs (time, energy) of an operation vary by orders of magnitude across different substrates, the logical result remains identical.

Methodology:

1. **The Cognitive Operation (()):**
    - A small but non-trivial, pre-trained neural network is chosen as the canonical (). For example, a 2-layer transformer with `d_model=64` and fixed weights, trained on a simple task like character-level text generation. The key is that the weights are frozen; we are testing inference, not learning.
2. **Substrate Implementations:** The exact same logical operation (i.e., the forward pass of the transformer) is implemented on four radically different computational substrates:
    - **Substrate 1 (GPU - Baseline):** A standard, highly optimized implementation using PyTorch/TensorFlow on a modern GPU.
    - **Substrate 2 (CPU - Actor Model):** An implementation on a multi-core CPU cluster using a fundamentally different execution paradigm, such as an actor-based model (e.g., using `Ray` or `Akka`), where matrix multiplications are decomposed into message-passing tasks between parallel processes.
    - **Substrate 3 (FPGA - Hardware Synthesis):** The neural network's architecture is synthesized into a hardware description language (Verilog/VHDL) and "burned" directly into the logic gates of a Field-Programmable Gate Array. This represents a complete translation from software logic to physical circuit configuration.
    - **Substrate 4 (Neuromorphic - Spiking Model):** The network is converted into a Spiking Neural Network (SNN) equivalent and executed on a neuromorphic chip (e.g., Intel Loihi 2). This requires translating continuous activation values into discrete spike trains, representing a different computational model.
3. **Experimental Procedure:**
    - A fixed, canonical validation dataset is created.
    - This dataset is run through the forward pass of the model on each of the four substrates.
    - The logical outputs and physical performance metrics are meticulously recorded for each run.

Prediction from Principia Cognitia: The logical output vectors produced by all four substrates will be equivalent within the bounds of their respective numerical precisions. The physical costs will differ by orders of magnitude, with the GPU being the most efficient

and the CPU Actor Model or Neuromorphic chip showing vastly different performance profiles.

Metrics and Success Criteria: * **Logical Equivalence:** The output vectors from all substrates must be verified as identical using a high-precision comparison (e.g., `torch.allclose` with a tight tolerance). * **Physical Cost:** Latency (wall-clock time per inference) and Energy Consumption (Joules per inference, measured using appropriate hardware/software tools like `pyJoules`). * **Success Criterion:** The experiment is successful if logical equivalence is confirmed while physical costs diverge by at least one order of magnitude between any two substrates.

Falsification Condition: The axiom `AX-SUBSTR-INV` would be falsified if (a) the logical outputs are not equivalent between substrates (beyond explainable precision differences), or (b) any computational primitive of the operation proves to be fundamentally incomputable on one of the chosen substrates.

Required Resources: Access to a GPU, a multi-core CPU cluster, an FPGA development board and toolchain (e.g., Xilinx Vivado), and potentially access to a neuromorphic computing platform.

D.3. The Vectorial Genesis Experiment (VGE-1)

Objective: To test the **Axiomatis Primitivorum Operandi (`AX-OPER-BASIS`)** and the **Lemma de Continuitate Compositionis**. This experiment aims to demonstrate that a complex hierarchy of **vectorial** cognitive operations, analogous to those in modern neural networks, can emerge spontaneously from a minimal basis of primitive vector operations ({`cmp, add, sub`}) through a process of selection and composition. This directly contrasts with symbolic A-Life experiments (e.g., using Brainfuck) by focusing on the genesis of vector-based cognition, the foundation of PC.

Methodology:

1. **Environment - The "Vectorial Soup":**
    - A 2D grid environment is created. Each cell in the grid contains a `d`-dimensional real-valued vector (a primitive semion), e.g., `d=16`. The grid is initialized with regions of vectors with different statistical properties.
2. **System Architecture:**
    - A population of simple "agents" (programs) is initialized. The instruction set available to these agents is strictly limited to three primitive, element-wise vector operations:
        - `add(vec_a, vec_b) -> vec_c`
        - `sub(vec_a, vec_b) -> vec_c`
        - `cmp(vec_a, vec_b) -> mask` (a vector of 0s and 1s based on which elements in `vec_a` are greater than in `vec_b`). This provides non-linearity.

- The agents can read vectors from their local neighborhood in the grid and write results back.

3. **Experimental Procedure:**
   - **Tasks:** The environment presents a series of tasks that require progressively more complex vector processing, e.g.:
     - *Level 1 (Aggregation):* Calculate the average vector in a 3x3 neighborhood.
     - *Level 2 (Boundary Detection):* Move along the border between two regions of vectors.
     - *Level 3 (Feature Extraction):* Identify the presence of a specific vector pattern (a "corner" or "edge") in a local patch.
   - **Evolutionary Mechanism:** A genetic algorithm is used. Programs are represented as sequences of primitive instructions. Their fitness is determined by their performance on the current task. High-fitness programs are selected for crossover and mutation, creating the next generation of programs.
4. **Analysis:** The primary analysis is not on task performance itself, but on the **structure of the evolved programs**. We will analyze the surviving high-fitness programs to identify the emergence of:
   - **Stable Subroutines:** Recurring sequences of primitives that are reused across different programs.
   - **Functional Analogs:** Subroutines that are functionally equivalent to higher-level cognitive operations. For example:
     - *Convolution:* A sequence that performs a series of weighted additions across a neighborhood.
     - *Attention:* A sequence that uses `cmp` to create a mask (attention weights) and then performs a weighted sum.
   - **Hierarchical Composition:** The reuse of simpler, evolved subroutines to build more complex ones.

Prediction from Principia Cognitia: Starting from only `{cmp, add, sub}` and a selection pressure, the system will spontaneously evolve a hierarchy of complex, reusable vector operations. The complexity of these operations will grow continuously and compositionally, demonstrating that the minimal basis is sufficient for the genesis of advanced vector-based cognition.

Falsification Condition: The axioms would be challenged if, despite prolonged evolution, the system fails to produce hierarchical and compositional programs, getting stuck at the level of simple, non-reusable sequences of primitives and failing to solve more complex tasks.

Required Resources: A high-performance computing cluster is recommended, as genetic algorithms are computationally intensive. Python with libraries for parallel execution (e.g., `multiprocessing`, `Ray`).

D.4. Integrated Validation Strategy

The preceding protocols are not isolated tests but constitute an integrated, multi-pronged validation strategy designed to empirically ground the core tenets of *Principia Cognitia*. Each experiment targets a distinct, foundational pillar of the theory, moving from communication to architecture to genesis.

- The **MLC Primacy Experiment (MPE-1)** is designed to validate the theory's model of meaning and communication by directly testing the formal predictions of the MLC/ELM duality (`TH-LANG-04`).
- The **Substrate Invariance Test (SIT-1)** addresses the framework's physical and architectural principles, providing a direct, falsifiable test of the Substrate Invariance postulate (`POS-SUBSTR-INV`) by decoupling the logical form of a cognitive operation from its material implementation.
- Finally, the **Vectorial Genesis Experiment (VGE-1)** provides a constructive proof for the theory's developmental core, testing the claim that the entire hierarchy of complex vectorial cognition can emerge compositionally from the minimal primitive basis (`POS-OPER-BASIS`).

Together, these three experimental avenues systematically target the theory's most unique and powerful claims—its model of meaning, its architectural principles, and its constructive genesis—thereby establishing *Principia Cognitia* as a complete and empirically falsifiable scientific program.

D.5. Experimental outlook

These proposals operationalize Principia Cognitia's core claims with crisp, falsifiable targets and light, modular implementations. They were distilled from prior lightweight script prototypes and are designed for independent replication; complete protocols are available from the author on request.

- **Falsifiability:** Each experiment has a direct pass/fail criterion tied to a pillar of PC—MLC primacy (MPE-1), substrate invariance (SIT-1), and vectorial genesis (VGE-1)—so negative results meaningfully update the theory rather than being shrugged off.
- **Modularity and reproducibility:** Synthetic data generators, deterministic seeds, and small models keep runs cheap and portable. The same workloads can be replayed across substrates (SIT-1) and scaled from desktop to lab clusters without altering the logic.
- **Distinctiveness:** Unlike A-Life demonstrations of symbolic emergence, these tests target PC's unique commitments: vector cognition, thermodynamic constraints, and the MLC↔ELM duality as an operational, measurable interface.
- **Interpretability and diagnostics:** Planned readouts (geometry/probing metrics, bitwise agreement, energy/latency profiles, macro-op lineage) make failure modes

informative—guiding revisions to the primitive basis, alignment procedures, or invariance claims.

- **Collaboration pathway:** The author provides specifications, reference scripts, and analysis templates; experimental partners contribute execution environments (GPUs/CPUs; optional FPGA/neuromorphic) and measurement instrumentation, enabling rapid, transparent replication.

Note from the Author — The author is a theoretical researcher; tools are limited to personal computing (i5 CPU, 64 GB RAM, RTX 4060 8 GB, SSD 3 TB), LLM dialogues, pen and paper, and minimal local ML prototyping (e.g., LM Studio). Full implementation of these protocols is left to experimental collaborators with suitable facilities.