Toward a Complete Framework for Next-Generation AI: Evaluation, Specialist Routing, Substrate Optimization, and Self-Structuring Architecture

Winston Zai Lin
Blair Academy
linzaiwinston413@gmail.com

March 2026

Abstract

The claim that certain domains of human capability—creative work, deliberate reasoning, relational care—lie permanently beyond AI is stated as a fact of nature. We argue it is a fact about design. The distinction matters: facts of nature do not yield to engineering; facts about design do.

This paper's central thesis is that no domain of human cognitive, creative, or relational capability is architecturally irreplaceable by AI. Every claimed barrier reduces to a specific, identifiable architectural requirement that current systems were never designed to satisfy. We specify those requirements and show they are satisfiable.

A complete intelligence system requires seven architectural properties: (1) efficiency measurement grounded in deployment value rather than benchmark accuracy; (2) domain-expert depth through specialist routing rather than uniform generalization; (3) computation allocated to physically optimal substrates rather than forced through a single digital pipeline; (4) structural adaptation during learning rather than fixed topology; (5) autonomous aesthetic agency with an internal drive toward novelty rather than prompt-response generation; (6) deliberate reasoning via explicit intermediate state and metacognitive uncertainty tracking rather than pattern completion; and (7) persistent relational intelligence—a live model of a specific individual that evolves across time and navigates genuine moral complexity. Current systems satisfy none of these fully. Not because a wall has been reached, but because none of these were design goals.

We present **ATLAS** (Adaptive Topology Living Architecture with Specialists), a unified architecture *designed to* satisfy all seven requirements through seven composed components: **CES/TES** (efficiency evaluation), **ARBOR** (specialist routing), **SPECTRUM** (substrate allocation), **FLUID** (topology learning), **MUSE** (aesthetic

agency), **LOGOS** (deliberate reasoning), and **SOCIUS** (relational intelligence). We provide empirical work on two of these. In a sample of eight models on 100 tasks, token efficiency (TES) and answer quality are inversely ranked (Spearman $\rho = -0.33$, $n = 8$, $p \approx 0.42$—an observed sample association, not a statistically significant effect), and the two reasoning models incur a 12–18× token penalty for quality comparable to far cheaper dense models. A grid search over the FLUID growth rule reveals a *narrow* stable-topology regime (12 of 120 configurations) rather than establishing convergence, with the candidate Lyapunov function failing in all stable cases. The remaining five components are theoretically specified with falsifiable validation protocols. All token-efficiency code, the task suite, and results are released (see Code and data).

1 Introduction

The boundary between AI and human capability is drawn in the wrong place. It is drawn at the edge of what has been built, not at the edge of what can be built. These are different edges.

The argument of this paper proceeds in three steps. First: every domain claimed to require permanent human involvement can be decomposed into a finite set of architectural requirements. Second: each requirement is specifiable—it can be stated precisely enough that a system either satisfies it or does not. Third: the specifications are satisfiable—no requirement invokes consciousness, subjective experience, or any property that cannot be engineered. The conclusion follows: no domain is permanently beyond reach.

Seven architectural requirements. A complete intelligence system must be:

1. *Measurable* by deployment-relevant efficiency metrics, not benchmark proxies that ignore compute cost and real-world value

2. *Deep* in any domain it operates in—expert-level, not competent-across-all

3. *Physically efficient*, allocating each cognitive function to the substrate whose physics is best suited to it

4. *Structurally adaptive*, modifying its own architecture during learning rather than being fixed at design time

5. *Autonomously creative*, generating novel aesthetic work from an internal drive rather than producing output on demand

6. *Deliberately reasoning*, maintaining explicit intermediate state, tracking its own uncertainty, and composing known rules in configurations it has never seen

7. *Relationally intelligent*, maintaining a persistent and evolving model of a specific individual across time and navigating genuine moral complexity

No existing system satisfies all seven. This is not evidence of a wall. It is evidence of incomplete design goals. The systems that exist were built to generate plausible text. These requirements specify something different: an intelligence.

This paper. We specify, for each requirement, the architecture that satisfies it. We then show the seven architectures compose without contradiction into a single system, ATLAS, and identify the empirical path from here to there.

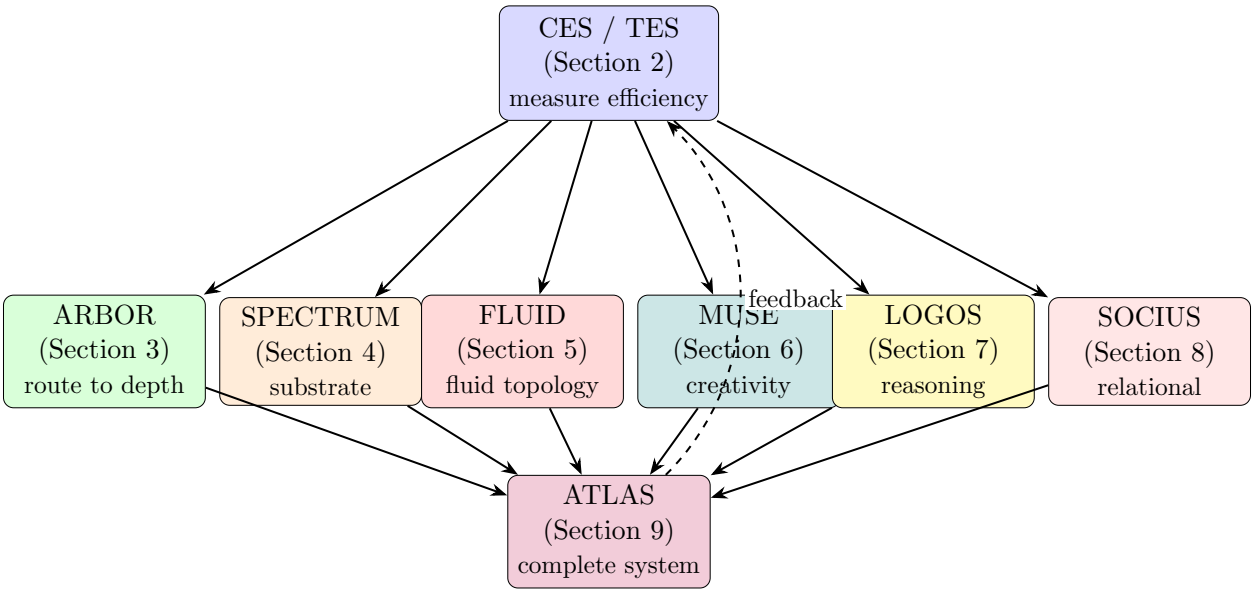

Figure 1: The unified framework. Each component targets one architectural requirement for complete intelligence. ATLAS is designed to combine all seven; CES/TES measures whether it is working.

2 Layer I: Evaluation — CES and TES

2.1 What Efficiency Requires

MMLU accuracy of 90% does not tell you whether a model is worth deploying. It does not tell you the compute cost, the latency, or whether a human with 30 minutes could do better. A model that achieves 90% accuracy at 1000T FLOPs is less valuable than one achieving 87% at 8T FLOPs for most deployments.

2.2 Cognitive Efficiency Score (CES)

Definition 1 (Cognitive Efficiency Score)**.**

$$\mathcal{L}(M) = \frac{I(M)}{\log_2(1 + C(M))} \Big/ \frac{I(H)}{\log_2(1 + C(H))} \tag{1}$$

where $I(M) = w_M \cdot b_{MMLU}(p_{MMLU}) + w_{HE} \cdot b_{HE}(p_{HE})$ *is a bounded competence score aggregating two benchmarks through the per-benchmark transforms* $b_j(\cdot)$ *defined below,* $C(M)$ *is the model's inference FLOPs in teraFLOPs, and* H *is a human expert anchor* ($C(H) = 1000\,T$)*. Variance-optimal weights:* $w_M = 0.47$, $w_{HE} = 0.53$.

Each benchmark j contributes a competence score that normalizes accuracy against that benchmark's chance baseline c_j and clips below-chance performance to zero:

$$b_j(p) = \max\Big(0,\ \frac{p - c_j}{1 - c_j}\Big), \tag{2}$$

with $c_{\text{MMLU}} \approx 0.25$ (four-option) and $c_{\text{HE}} \approx 0$ (unit-test pass/fail). This maps chance to 0 and perfect to 1 per benchmark, avoiding a single fixed threshold. An earlier variant used the binary-entropy form $b(p) = \max\big(0,\ 1 - H_2(p)\big)$, which assumes a 0.5 chance level and therefore over-penalizes sub-0.5 multi-choice scores (a 0.49 on four-option MMLU is well above chance yet would be zeroed); the 13-model figures reported below were computed with that binary variant and would shift under the benchmark-specific baselines above, though we do not expect large reorderings. We use b_j as a monotone $[0, 1]$ competence proxy, not a literal information-theoretic quantity. The anchor $C(H) = 1000\text{T}$, the logarithmic compute penalty, and the weights w are design choices, not derived or independently validated here, and the CES scale depends on them; CES is therefore an *anchor-relative* score. Under this anchor, CES = 1.0 denotes parity with the chosen human reference and CES > 1.0 better quality-per-compute than it. Establishing an absolute "human-expert efficiency" reading would require estimating the anchor and weights from data and reporting sensitivity to them.

Empirical results (13 models): CES correlates with HumanEval at $\rho = 0.923$ and MMLU at $\rho = 0.725$. Because these benchmarks are *inputs* to CES, this correlation is partly mechanical and does not independently validate the metric; it shows only that the compute normalization does not wholly override benchmark signal. FLOPs perturbation ($\pm 30\%$) yields $\rho \geq 0.989$ ranking stability. No model in the 13-model sample exceeds CES $= 0.73$ under variance-optimal weights.

2.3 Token Efficiency (TES)

CES measures efficiency relative to a model's static compute (FLOPs). But the compute a user actually pays for at deployment is not a fixed parameter count — it is the number of tokens the model *generates* to answer each query. Tokens are a primary driver of inference cost and, at a fixed model and hardware, of wall-clock time. Two models of similar size can differ by more than an order of magnitude in tokens emitted per task, and reasoning models that spend a long chain-of-thought before answering can emit tens of times more. We therefore measure generated-token economy. Two caveats. First, per-token cost itself scales with model size—a token from a 671B model is far dearer than one from a 7B model—so TES captures how economically a model uses its *own* generation, not absolute dollar cost; a price comparison across sizes would weight tokens by a per-model rate. Second, tokens are not a uniform unit across models: different tokenizers encode different amounts of text per token, so cross-model counts carry tokenizer variation, and a fully normalized comparison would report characters or bytes (or retokenized output) alongside provider tokens. We treat TES as a token-economy measure and flag both as limitations.

Definition 2 (Token Efficiency Score)**.**

$$\text{TES}(M) = \frac{1000\,\bar{q}(M)}{\bar{\tau}(M)} \tag{3}$$

where $\bar{q}(M) \in [0,1]$ is mean quality across tasks and $\bar{\tau}(M)$ is the mean number of tokens the model generates per task, counting the full completion (chain-of-thought plus answer). TES is quality delivered per thousand generated tokens. Token counts are more portable than wall-clock latency, which varies with server load and batching. They are not, however, exactly deterministic: at temperature 0 *they are typically stable for a fixed endpoint, tokenizer, and model revision, but can vary across inference backends, batching, kernel nondeterminism, and revisions, and depend on the tokenizer and on whether hidden reasoning tokens are billed and observable. We report provider-billed completion tokens from the HuggingFace inference router; the endpoint, decoding settings, and repeated-run variance are given in the released artifacts, and the values below should be read as characteristic magnitudes rather than exact constants.*

We evaluate eight instruction-tuned models (7B–671B parameters) on a 100-task suite spanning structured data, code, math, and classification, all scored automatically against ground truth.

Model	**Params**	**Tokens/task**	**Avg. quality**	**TES**
Qwen2.5-72B	72B	121	0.780	6.46
Qwen2.5-7B	7B	113	0.682	6.03
Llama-3.3-70B	70B	146	0.758	5.20
Gemma-3-27B	27B	151	0.695	4.61
DeepSeek-V3	671B	200	0.885	4.43
Llama-3-8B	8B	162	0.645	3.98
Qwen3-8B	8B	1833	0.867	0.47
DeepSeek-R1	671B	2685	0.841	0.31

Table 1: Token efficiency across eight models on 100 tasks. TES is quality per thousand generated tokens (higher is better). The two reasoning models (bottom) are measured with their full chain-of-thought. Spearman $\rho(\text{quality}, \text{TES}) = -0.33$ ($n = 8$, $p \approx 0.42$; a sample association, not a significant effect).

Token efficiency and quality are inversely ranked in this sample. Across these eight models, quality and token efficiency show a negative rank association (Spearman $\rho(\text{quality}, \text{TES}) = -0.33$). With only $n = 8$ models this is not statistically significant (two-sided $p \approx 0.42$); we report it as an observed pattern to test at larger scale, not an established law. The highest-quality systems in this sample are nonetheless among the *least* token-efficient (Figure 2). DeepSeek-V3 leads on quality (0.885) yet sits mid-pack on TES because it emits the most tokens of any dense model; the most token-efficient models are the mid-size dense Qwen variants (TES 6.0–6.5), which pair competitive quality with terse output. *Reasoning models pay a severe token tax.* DeepSeek-R1 (quality 0.841) and Qwen3-8B (quality 0.867) reach near-top quality but generate 2,685 and 1,833 tokens per task respectively, roughly 12–18× the ∼148-token median of the dense models, collapsing their quality-per-token to 0.31 and 0.47, an order of magnitude below the dense field. (Whether the chain-of-thought *causes* their quality would require a controlled reasoning-on/off comparison, which we do not run here.) This is the cost-grounded, token-count-based form of the

observation that "more thinking" is not free: it is paid for in tokens, and a deployment-value metric must price it. The role of this layer in the unified framework: CES/TES provides the fitness function against which all architectural decisions in ARBOR, SPECTRUM, and FLUID are evaluated — and it explicitly prices the tokens a design chooses to spend.

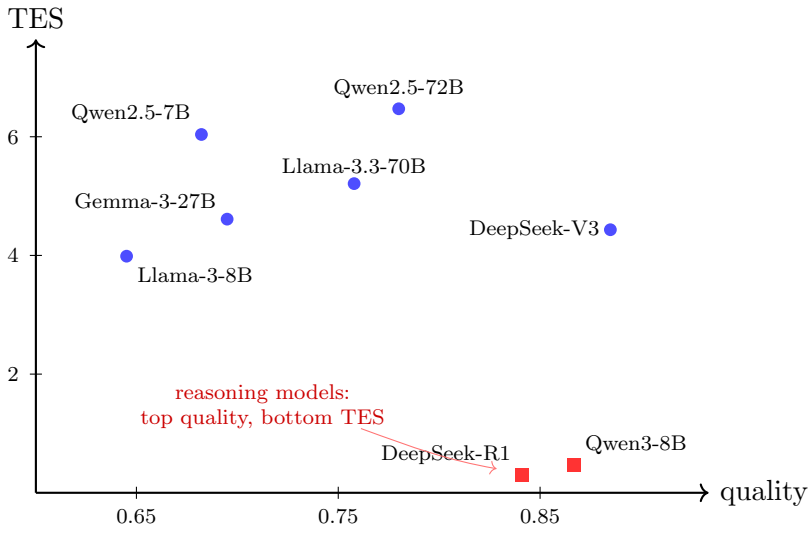

Figure 2: Answer quality vs. token efficiency (TES) across the eight models. Dense models (blue) occupy the high-TES region; the two reasoning models (red) reach near-top quality but the lowest TES. The negative rank association in this sample ($\rho(\text{quality}, \text{TES}) = -0.33$, $n = 8$) is not statistically significant and is shown as an observed pattern.

3 Layer II: Organization — ARBOR

3.1 Depth Requires Separation

A generalist trained on all human knowledge spreads its parameters across all domains; a specialist trained only on physics devotes its full budget to physics. Under a uniform-allocation idealization (and ignoring cross-domain transfer), a specialist can therefore reach up to $N\times$ the per-domain capacity of an equal-budget generalist covering N domains.

Definition 3 (ARBOR System)**.** *ARBOR consists of: (1) a* ***Big Brain*** $\mathcal{B}$*—a generalist model handling routing, translation, and synthesis; (2) a* ***Long Brain Registry*** $\mathcal{L} = \{(\ell_i, \mathbf{d}_i, \Omega_i)\}$*—domain specialists trained exclusively on domain* Ω_i*; and (3) a standardized* ***Interface Contract*** *exposing three functions per specialist:* `domain_embedding()`*,* `query(text)→(answer, confidence)`*,* `scope()`*.*

Depth intuition (informal, not a theorem). Let $\mathcal{D}(m, \Omega)$ denote the effective capacity model m devotes to domain Ω. Under the idealization that a generalist $\mathcal{G}$ of budget P spreads capacity uniformly across its N domains with no cross-domain transfer, its per-domain capacity is $\mathcal{D}(\mathcal{G}, \Omega_i) \approx P/N$, while a specialist committing its full budget $P_\ell = P$ to

one domain has $\mathcal{D}(\ell_i, \Omega_i) \approx P_\ell$, giving

$$\frac{\mathcal{D}(\ell_i, \Omega_i)}{\mathcal{D}(\mathcal{G}, \Omega_i)} \approx \frac{P_\ell}{P/N} = N. \tag{4}$$

This is an *upper bound under a strong idealization*. Real generalists exhibit substantial cross-domain transfer and shared low-level features, and $\mathcal{D}$ is not a measured quantity here, so the realized advantage is smaller and must be established empirically (see the ARBOR open problems). We state this as a design motivation, not a proven result.

3.2 Two-Stage Routing

Stage 1 (fast, $O(N)$): cosine similarity between query embedding and domain embeddings. Stage 2 (precise): Big Brain confirmation score $\mathcal{B}_{\text{route}}(q, \Omega_i) \in [0, 1]$.

Combined: $s_i = (1 - \alpha) \cdot \cos(\mathbf{q}, \mathbf{d}_i) + \alpha \cdot \mathcal{B}_{\text{route}}(q, \Omega_i)$, with $\alpha = 0.4$ as an illustrative default—a tunable hyperparameter, not yet empirically calibrated.

3.3 Hierarchical Synthesis

Single specialist: trivial. Multiple specialists: check agreement $\text{agree}(a_i, a_j) = \cos(e(a_i), e(a_j))$. If agree $> \theta$: weighted merge by confidence. If disagree: debate round (specialists see each other's answers, revise), Big Brain arbitrates. Mirrors expert panel arbitration in medicine and law.

3.4 ARBOR in the Unified Framework

ARBOR's routing policy is learned from TES feedback: specialists with low TES on a query type (poor quality per token spent) are called less. CES measures whether the routing overhead is worth the depth gain. ARBOR feeds efficiency data back to Layer I.

4 Layer III: Substrate — SPECTRUM

4.1 Substrate Is Not Neutral

The dominant foundation-model paradigm runs on dense matrix multiplication on GPUs and TPUs. Other hardware is deployed—CPUs, NPUs, ASICs, FPGAs, and neuromorphic and analog devices—but the large models that define the field are trained and served overwhelmingly on matrix-multiply accelerators. This is not because they are the best substrate for intelligence; it is because they were available when the field scaled. The transformer architecture was not designed for intelligence; it was designed for what this hardware does efficiently.

The brain uses at least six physically distinct computational mechanisms simultaneously, each suited to a different function. SPECTRUM applies the same principle to silicon.

4.2 Substrate Taxonomy

Substrate	Best suited to	SPECTRUM layer	Power
Neuromorphic [7]	Event detection, temporal patterns	L1: Perception	mW-scale
Optical [23]	Fourier transforms, correlation	L1: Perception	passive/low
Energy-based [19]	Associative memory, constraint sat.	L2: Working memory	mW-scale
Molecular (DNA)	Long-term storage, pattern match	L3: Long memory	passive
Analog [24]	Physics simulation, continuous opt.	L4: Reasoning	mW-scale
Quantum	Search, combinatorial optimization	L5: Search	device-dependent
Digital	Symbolic manipulation, language	L6: Language output	W-scale

Table 2: SPECTRUM substrate allocation with qualitative per-substrate power-magnitude classes (from the cited hardware literature). These are *not* measured figures or a system total: the classes come from different tasks and devices and omit workload, throughput, data movement, and cross-substrate conversion overhead. A like-for-like comparison against a transformer baseline would require a defined workload and end-to-end measurement, which we do not provide here.

Two novel substrates not yet integrated in any architecture:

Physical reservoir computing: any nonlinear dynamical system (springs, fluid, soft robots) computes naturally when driven by input; only the readout layer is trained. Approaches thermodynamic energy limits.

Morphological computation: graph topology itself computes—changing structure is the operation, not the carrier of weights.

4.3 Critical repositioning

In SPECTRUM, the transformer is L6: the output layer. It handles human communication. All reasoning occurs in L1–L5. This is the architectural inversion current AI needs: language is the mouth, not the mind.

4.4 SPECTRUM in the Unified Framework

CES can be extended to substrate-aware efficiency: $\mathcal{L}_{\text{SPECTRUM}}$ uses physical FLOPs (summed across substrates with substrate-specific energy equivalents) rather than digital FLOPs alone. TES measures the token cost of routing a query through SPECTRUM's multi-layer pipeline. ARBOR's Long Brains run on substrate-native hardware matched to their domains.

5 Layer IV: Adaptation — FLUID

5.1 Architecture Must Be a Variable

In the dominant paradigm, a neural network is a *fixed* graph $G = (V, E)$: learning modifies weights $w : E \to \mathbb{R}$ but not the graph. Neuroevolution, dynamic sparse training, and adaptive-computation methods do vary connectivity (we compare to them below), but typically between training phases or by weight-magnitude criteria rather than by continuous, co-activation-driven growth during learning. In the standard case, representational capacity is largely determined before the first training step, and learning cannot fully compensate for a poorly matched architecture.

The brain undergoes continual structural plasticity—synapses form and prune throughout life, not only during development. Structural change is not noise; it is part of how long-term memory and new skills are encoded.

5.2 FLUID Architecture

Definition 4 (FLUID Network)**.** *A FLUID network $\mathcal{F} = (G_0, \mathcal{H}, \mathcal{C}, \mathcal{M})$ where G_0 is a seed graph, $\mathcal{H}$ is a Hebbian growth rule, $\mathcal{C}$ is a causal memory module, and $\mathcal{M}$ is a meta-network that monitors G_t and proposes structural modifications. $G_t = (V_t, E_t)$ with both V and E varying over time.*

Hebbian growth. Edge creation: if co-activation $\rho_{ij}(t) = \frac{1}{T}\sum a_i a_j > \theta_{\text{create}}$. Edge deletion: if $|w_{ij}| < \theta_{\text{prune}}$ for T_{prune} steps. Node creation: when an edge cluster saturates.

Causal memory. Every observation stored as $(x, \mathbf{U}, \mathbf{PA})$ with causal context, enabling interventional ($do(X)$) and counterfactual queries—not just correlational pattern matching.

Temporal hierarchy emergence. Nodes cluster by temporal receptive field $\text{TRF}(i)$ without hand-designed layers. Fast sub-networks (ms) handle transient events; slow sub-networks (hours) integrate extended history.

Recursive self-architecture. Meta-network $\mathcal{M}$ (itself a FLUID network) proposes structural modifications: split nodes, merge nodes, add shortcuts, adjust TRF. Modifications accepted if they improve held-out validation performance.

Intrinsic motivation. Compression-progress curiosity: $\kappa(t) = C_{t-\Delta t} - C_t$ (reduction in description length). Rewards structured novelty without human-assigned reward functions.

Conjecture 1 (Fluid Stability)**.** *Under Hebbian growth with compression-progress curiosity, G_t converges in distribution to a stable small-world graph rather than degenerating. The natural Lyapunov candidate is $\mathcal{V} = C_t/|E_t|$.*

5.3 Empirical Stability Results

The Fluid Stability Conjecture has been empirically tested on 100-node networks across three synthetic tasks (XOR, Parity-4, symbolic regression) with a grid search over 120 parameter configurations in $(\theta_{\text{create}}, \theta_{\text{prune}}, T_{\text{prune}})$ space [10].

Phase diagram. Outcome distribution across 120 configurations: stable 10.0%, fully connected 33.3%, disconnected 50.0%, growing 6.7%. The dominant degenerate mode is

disconnection, occurring whenever $\theta_{\text{create}} \geq 0.5$ regardless of pruning parameters. The stable region is narrow: $\theta_{\text{create}} = 0.3$, $\theta_{\text{prune}} \leq 0.01$.

Convergence behavior. XOR was the only task inducing topology growth under default parameters, converging to 128 edges (APL = 1.97, clustering = 0.000) at step 11,000. Parity-4 and symbolic regression solved their tasks with zero hidden edges, relying on fixed seed connections. Task complexity did not predict topology density.

Lyapunov analysis. The candidate $\mathcal{V} = C_t/|E_t|$ is non-monotone in all 12 stable configurations tested (0/12 non-increasing; mean decreasing fraction ≈ 0.46, near chance). This suggests stability in this system is limit-cycle rather than fixed-point convergence, and a different Lyapunov function is needed.

Verdict. These experiments do *not* resolve the conjecture. They identify an *observed* stable regime—12 of 120 configurations (10%)—under this particular task set and growth rule; outside a narrow band ($\theta_{\text{create}} = 0.3$, $\theta_{\text{prune}} \leq 0.01$) the dynamics degenerate, the single task that grew under default parameters had zero clustering, two of the three tasks solved with no hidden edges, and the candidate Lyapunov function failed in all 12 stable cases. The evidence supports the *existence of a stable regime under this experiment*, not convergence in distribution to a stable small-world graph. Establishing convergence would require a validated Lyapunov (or other) argument and a broader sweep over tasks, scales, and node-creation dynamics.

Comparison to prior topology learning. NEAT [11] modifies topology between training generations; Deep Rewiring [12] uses weight perturbation rather than co-activation; dynamic sparse training [13] grows connections by weight magnitude rather than Hebbian co-activation. None tests convergence under the specific FLUID Hebbian rule.

5.4 FLUID in the Unified Framework

TES replaces compression-progress as FLUID's growth signal when deployment value matters more than abstract information gain. FLUID Long Brains in ARBOR evolve their internal topology as domain knowledge accumulates. FLUID running on SPECTRUM hardware adapts at both the structural and substrate level simultaneously.

6 Layer V: Creative Agency — MUSE

6.1 Creative Agency Requires an Internal Drive

Music, visual art, and dance are the three domains where the "AI will never replace humans" consensus is most confident and most uniform. The consensus rests on three claims: (1) authentic art requires lived emotional experience, (2) dance requires a physical body, and (3) creative work requires genuine intentionality—a reason to make something, not just the ability to execute it when asked. These claims are correct about current AI systems. They are wrong as architectural constraints.

The failure is not capability but design. Current generative AI is reactive: it produces output when prompted and stops when the prompt is satisfied. It has no internal aesthetic state, no memory of its own creative history, no drive to develop a sustained vision, and—for

dance—no embodied output mechanism. These are not fundamental limits of AI. They are consequences of training paradigms built for question-answering, not creation.

Definition 5 (MUSE System)**.** *MUSE (Multi-modal Unified Synthetic Expression) is an autonomous creative agent* $\mathcal{U} = (\kappa_a, \mathcal{E}_a, \Phi, \mathcal{K})$ *where* κ_a *is an aesthetic curiosity signal,* $\mathcal{E}_a$ *is a multi-modal aesthetic embedding space,* Φ *is an embodied output module, and* $\mathcal{K}$ *is a cultural knowledge base. MUSE generates creative work autonomously—driven by* $\kappa_a > \theta_{create}$*—and produces output across music, visual art, and movement without requiring an external prompt.*

6.2 The Aesthetic Curiosity Signal

FLUID's compression-progress curiosity $\kappa(t) = C_{t-\Delta t} - C_t$ rewards information compression. MUSE requires an aesthetic analogue: a signal that rewards novelty *within* a coherent style, penalizing both repetition and randomness.

Definition 6 (Aesthetic Curiosity Signal)**.**

$$\kappa_a(t) = \lambda_1 \cdot \Delta N_a(t) - \lambda_2 \cdot \Delta S(t) \tag{5}$$

where $\mathbf{e}_t$ *is the embedding of a* candidate *output the generator proposes (not one already committed),* $\Delta N_a(t) = d(\mathbf{e}_t, \mathcal{C})$ *is that candidate's distance from the existing cultural corpus* $\mathcal{C}$ *in the aesthetic embedding space* $\mathcal{E}_a$ *(aesthetic novelty), and* $\Delta S(t) = \|\mathbf{e}_t - \bar{\mathbf{e}}_{self}\|$ *is style incoherence—distance from MUSE's own recent output centroid. Empirically calibrated weights* $\lambda_1, \lambda_2 > 0$ *control the novelty/coherence tradeoff.*

MUSE proposes candidate continuations and scores each with κ_a: it commits the best candidate when its $\kappa_a > \theta_{\text{create}}$, stays silent when no candidate exceeds θ_{prune} (nothing new to say), and otherwise keeps developing existing work. Because the score ranks *prospective* candidates rather than an already-produced output, the trigger is not circular. This specifies an internal, prompt-independent drive to create.

6.3 Multi-Modal Aesthetic Embedding

Music, visual art, and movement are physically distinct but aesthetically commensurable: a minor-key chord progression and a dark color palette occupy nearby regions of human aesthetic experience. MUSE formalizes this with a shared embedding space $\mathcal{E}_a \subset \mathbb{R}^d$ ($d = 512$) in which all three modalities are represented.

Cross-modal encoding: the same aesthetic position $\mathbf{e} \in \mathcal{E}_a$ can be decoded into music (via a music decoder $f_{\text{music}} : \mathcal{E}_a \to \mathbb{R}^T$), into a visual output (via $f_{\text{vis}} : \mathcal{E}_a \to \mathbb{R}^{H \times W \times 3}$), or into a movement sequence (via $f_{\text{move}} : \mathcal{E}_a \to \mathbb{R}^{J \times T}$, where J is the number of joints). A single aesthetic intention produces all three. The embedding is trained on paired human examples: soundtracks composed for paintings, choreography created for music, visual art made to accompany performance.

6.4 Embodied Output Module for Dance

Dance is the hardest domain. The irreplaceability claim is strongest here because it requires a *body*—a physical system that moves through space, not a description of movement. MUSE addresses this through SPECTRUM's morphological computation substrate.

A MUSE dance agent is a physical dynamical system whose topology IS the choreography: graph edges encode movement coupling between body segments; Hebbian co-activation (FLUID mechanism 1) discovers movement patterns that fire together. Specifically:

Body graph. A joint graph $G_{\text{body}} = (V_{\text{joints}}, E_{\text{coupling}})$ where nodes are body joints and edges are learned coupling strengths. A high-weight edge (i, j) means joints i and j move in coordinated patterns.

Morphological choreography. Edge creation via co-activation: if joints i and j co-activate above threshold θ_{create} during a movement segment, a coupling edge is created. Over training, the topology converges to a body-specific movement grammar. This is dance as topology learning, not dance as sequence generation.

Physical implementation. SPECTRUM's physical reservoir computing substrate (L4) provides the dynamical system. The body graph is the reservoir; only the readout weights are trained. Movement sequences emerge from the natural dynamics of the physical system, not from autoregressive token generation.

6.5 Cultural Knowledge Base

Authentic aesthetic output requires cultural grounding, not just statistical plausibility. MUSE's cultural knowledge base $\mathcal{K}$ is implemented as a set of ARBOR domain specialists:

- ℓ_{music}: trained exclusively on music theory, composition history, harmonic analysis, and recorded performance (MIDI + audio corpora)
- ℓ_{visual}: trained on art history, visual theory, color relationships, compositional principles, and curated image corpora
- ℓ_{dance}: trained on choreography notation (Laban movement analysis, Benesh notation), video-captured movement, and dance criticism

When MUSE generates, it routes through $\mathcal{K}$ via ARBOR to ensure each output is culturally situated. The difference from a generalist: ℓ_{music} has trained only on music and can respond to the question "does this chord progression have historical precedent and in what tradition?" with the full depth of a musicologist.

6.6 The Irreplaceability Rebuttal

Irreplaceability claim	**What it asserts**	**MUSE response**
Emotional authenticity	Art requires genuine felt emotion	κ_a produces an internal drive with no human equivalent but equal functional role; "felt" is a process claim, not an output claim
Lived experience	You cannot paint loneliness without experiencing it	FLUID's causal memory encodes counterfactual chains over the complete corpus of human loneliness documentation; CCM provides cultural encoding of loneliness representation across all art traditions
Physical embodiment	Dance requires a body	SPECTRUM morphological computation + FLUID bodygraph; the dynamical system IS the body
Intentionality	Art requires a reason to make it	κ_a is a formal reason: MUSE creates when novelty exceeds threshold and style coherence is maintained
Cultural meaning	Art gains meaning from human social context	$\mathcal{K}$ embeds MUSE's output in the complete cultural context it was trained on; cultural situatedness is in the training, not the substrate
Consciousness	Creativity requires subjective experience	This is unfalsifiable and therefore not a scientific criterion. The testable claim is output quality; see ATT below

Table 3: The six irreplaceability arguments and MUSE's responses. The key conceptual move: irreplaceability is conflated with *process* (how it is made) rather than *output* (what it produces and what effect it has). The only falsifiable claim is output indistinguishability.

6.7 The Aesthetic Turing Test (ATT)

Analogous to TES for creative domains. The ATT operationalizes "replacement" as output indistinguishability by expert judges.

Definition 7 (Aesthetic Turing Test)**.** *The ATT presents expert judges (professional musicians, gallerists, choreographers) with pairwise outputs (human vs. MUSE) in a blind evaluation. MUSE passes the ATT in domain d if the identification rate $r_d \leq 0.55$ (near chance;*

$r = 0.50$ is perfect indistinguishability). The threshold 0.55 is provisional: a sound test must pre-specify a sample size, report a confidence interval on r_d, frame indistinguishability as an equivalence test against a stated margin rather than a mere failure to reject the null, and model judge and item random effects. We use $r_d \leq 0.55$ as an illustrative target, not a validated criterion.

Three tiers:

1. **Expert blind**: trained domain experts evaluate labeled pairs; provisional target $r \leq 0.55$ (subject to the sample-size and equivalence-testing protocol above)
2. **Audience blind**: untrained audiences in authentic performance contexts (concert hall, gallery, stage); provisional target $r \leq 0.55$
3. **Cultural impact**: MUSE-generated work achieves industry recognition without disclosure (chart position, gallery representation, dance company premiere)

Tier 3 is the strongest criterion: cultural impact requires not just indistinguishability but sustained aesthetic contribution. A human composer who consistently writes music that cannot be identified as AI-generated AND achieves chart success is, by any reasonable definition, replacing a human composer for that function.

6.8 MUSE in the Unified Framework

MUSE draws on all four prior ATLAS layers:

- **CES/TES**: extended to ATT for creative output; CES measures efficiency of the generation process
- **ARBOR**: $\mathcal{K}$ is implemented as ARBOR domain specialists; MUSE routes to ℓ_{music}, ℓ_{visual}, ℓ_{dance} during generation
- **SPECTRUM**: Φ (embodied output module) runs on morphological computation substrate (L4/physical); κ_a runs on analog substrate
- **FLUID**: κ_a is an aesthetic variant of compression-progress curiosity; body graph topology adapts via Hebbian coupling

MUSE is not an independent system. It is the creative layer of ATLAS—the point where the framework produces work that current AI systems, by consensus, cannot.

7 Layer VI: Deliberate Reasoning — LOGOS

7.1 Reasoning Requires Explicit State

> *"AI is not good at nonlinear thinking, and therefore, solving human problems can't be the strength of AI."*
> — Mark Beccue, Enterprise Strategy Group [36]

Every current large language model, when presented with a novel problem, does the same thing: it retrieves the most probable continuation given its training distribution. This is extraordinarily useful for the vast majority of queries. It is fundamentally incapable of deliberate reasoning.

The difference is not speed or scale. It is structural. Pattern completion produces an answer in one forward pass. Deliberate reasoning requires:

1. Decomposing the problem into sub-problems
2. Solving sub-problems and recording their conclusions as explicit state
3. Checking intermediate conclusions for consistency
4. Recognizing when the current chain has entered unknown territory and flagging uncertainty
5. Composing conclusions into a final answer that is *warranted* by the chain, not merely suggested by surface similarity to training examples

No current architecture does steps 2–5. Chain-of-thought prompting [32] elicits intermediate text, but that text is still generated by the same pattern-completion process—the "chain" is not a logical structure with checked consistency, it is a continuation that sounds like a chain. When a model writes *"therefore..."* it is predicting that a conclusion-word follows; it is not verifying that the conclusion is warranted. This is the distinction LOGOS addresses.

Definition 8 (LOGOS System)**.** *LOGOS (Layered Ontological Grounded Systematic reasoning) is a deliberate reasoning architecture* $\mathcal{G} = (\mathcal{R}, \mathcal{M}_c, \mathcal{I}, \mathcal{V})$ *where* $\mathcal{R}$ *is an explicit reasoning graph,* $\mathcal{M}_c$ *is a metacognitive monitor,* $\mathcal{I}$ *is a compositional inference engine, and* $\mathcal{V}$ *is a consistency verifier. LOGOS treats intermediate conclusions as first-class objects—nodes in* $\mathcal{R}$*—rather than as tokens in a continuation.*

7.2 The Reasoning Graph

The central data structure of LOGOS is the reasoning graph $\mathcal{R}_t = (C_t, J_t)$ where C_t is a set of conclusion nodes and J_t is a set of justification edges. Each conclusion $c_i \in C_t$ has:

- A propositional content $\phi(c_i)$
- A confidence score $u(c_i) \in [0, 1]$, tracked by $\mathcal{M}_c$
- A provenance: either a premise (from ARBOR domain specialist or FLUID causal memory) or derived (from an inference step in $\mathcal{I}$)
- A status: *active*, *suspended* (confidence below threshold), or *contradicted*

A justification edge $(c_i, c_j) \in J_t$ asserts that c_j follows from c_i (and possibly other premises) under inference rule r_{ij}. This is the logical structure that chain-of-thought text only simulates.

At each reasoning step, LOGOS adds a new conclusion node, checks consistency with the existing graph via $\mathcal{V}$, and updates confidence scores. Contradictions are not discarded—they are retained as *contradicted* nodes with their full provenance, enabling counterfactual analysis of where the reasoning diverged.

7.3 Metacognitive Uncertainty Monitor

Definition 9 (Metacognitive Confidence)**.** *For conclusion c_i with provenance chain $\pi(c_i) = (c_{i_1}, \ldots, c_{i_k}, c_i)$, the metacognitive confidence is:*

$$u(c_i) = \gamma(c_i) \cdot u_{base}(c_i) \cdot \min_{p \in \mathrm{PA}(c_i)} u(p), \qquad \gamma(c_i) = \begin{cases} 1, & \phi(c_i) \in \mathcal{D}_{train} \\ \delta, & otherwise, \end{cases} \tag{6}$$

with $\delta \in (0, 1)$. Here $u_{base}(c_i)$ is the confidence of the immediate inference step; the factor over the direct *premises* $\mathrm{PA}(c_i)$ *propagates parent uncertainty without re-multiplying already-aggregated ancestors (each parent's u already summarizes its own chain, so evidence is not double-counted); and $\gamma(c_i)$ discounts an out-of-distribution conclusion by δ rather than forcing its confidence to zero. For a premise node with no parents, the empty minimum is taken as $\min_{\varnothing} = 1$, so premises carry only u_{base} and γ.*

When $u(c_i) < \theta_{\text{flag}}$, LOGOS flags the conclusion as uncertain and propagates the flag to all nodes derived from c_i. This gives the system an explicit, propagating signal for *when* it lacks support for a conclusion, rather than only producing fluent text that asserts one.

The in-distribution factor γ is set by embedding $\phi(c_i)$ and measuring its distance from the training-corpus centroid in the FLUID causal memory space: conclusions beyond a distance threshold take the discounted value δ. A single global-centroid distance is only a coarse baseline for out-of-distribution detection—for multimodal embedding distributions a mixture- or local-density-based detector would be more reliable—so we treat it as a placeholder, not a solved component. Conclusions asserting novel facts (not seen in training) are automatically flagged for external verification before propagation.

7.4 Compositional Inference Engine

Current neural networks fail at compositional generalization: given that they can do A and can do B independently, they often cannot do $A \circ B$ when the combination is novel. This is because composition is an implicit operation in pattern-completion architectures—there is no explicit rule application, only learned associations.

LOGOS's compositional inference engine $\mathcal{I}$ maintains a rule library $\Lambda = \{r_1, r_2, \ldots\}$ where each rule $r_k : (C_{\text{premises}}, \text{conditions}) \to C_{\text{conclusion}}$ is an explicit first-order inference pattern. Rules are not hand-coded. They are extracted from ARBOR domain specialists: ℓ_{math} provides arithmetic and algebraic rules; ℓ_{logic} provides propositional and first-order rules; domain specialists provide domain-specific inference patterns.

At each step, $\mathcal{I}$ searches Λ for applicable rules given the current $\mathcal{R}_t$, generates candidate conclusions, scores them by $u(c)$, and adds the highest-confidence warranted conclusion. This is explicit rule application, not prediction: the conclusion is accepted because a rule warranted it, not because it is the most probable next token.

Coverage intuition (informal, not a theorem). If every rule a derivation needs is present in Λ and the search explores chains up to depth d, then a conclusion with such a derivation is in principle reachable whether or not that specific chain appeared in training. This is definitional given a complete-enough Λ and an exhaustive-enough search; it is *not* a substantive coverage guarantee. It assumes the rules are sound, that the search actually finds the chain (completeness), and that doing so is tractable—none of which is established here. We state it as the design target that separates rule application from pattern completion.

The intended contrast with pattern-completion architectures is this: given sound rules and adequate search, LOGOS can in principle solve a problem whose specific inference chain it never saw, because it applies rules rather than recalling associations.

7.5 Consistency Verifier

The consistency verifier $\mathcal{V}$ tests a new conclusion c_{new} against the *conjunction* of the active theory, not pairwise—pairwise checks miss joint inconsistency (A, B, $\neg(A \wedge B)$ are pairwise consistent yet jointly unsatisfiable):

$$\text{consistent}(c_{\text{new}}, \mathcal{R}_t) = \neg\,\text{UNSAT}_{\mathcal{L}}\Big(\phi(c_{\text{new}}) \wedge \bigwedge_{c_i \in C_t^{\text{active}}} \phi(c_i)\Big) \tag{7}$$

where $C_t^{\text{active}} = \{c \in C_t : \text{status}(c) = \text{active}\}$ excludes nodes already marked *contradicted* or *suspended*: those are retained in C_t for provenance but must not enter the consistency test, or every subsequent check would be unsatisfiable. Because first-order satisfiability is undecidable in general, $\mathcal{V}$ operates within a restricted decidable fragment $\mathcal{L}$ (or a bounded solver under a step/time budget); we therefore describe it as *bounded* consistency checking, not a complete general verifier.

If a contradiction is detected, LOGOS does not proceed. It marks the offending node as *contradicted*, traces back through the provenance chain to the branch point, and either revises the lower-confidence premise or flags the problem as requiring external input. This is abductive revision: when the reasoning fails, LOGOS asks which premise is most likely wrong and presents that as the uncertain point.

7.6 Three Reasoning Requirements and Their Specifications

Failure mode	**Current AI behavior**	**LOGOS response**
Multi-step collapse	Accuracy drops sharply as chain length increases; each step accumulates unchecked error	Confidence propagated explicitly; $u(c_i) < \theta_{\text{flag}}$ halts propagation before error compounds
Confident confabulation	Model asserts false claims outside training distribution with high token probability	In-distribution indicator flags OOD conclusions; $u(c_i)$ drops; answer marked uncertain
Compositional failure	Fails on novel $A \circ B$ even when A and B are individually known	Explicit rule application in $\mathcal{I}$; composition is exact, not learned

Table 4: The three core complex thinking failures and LOGOS's mechanisms for each.

7.7 Case Study: Legal Reasoning

The Eskimoz analysis [2] ranks lawyers as the most AI-resistant profession (score: 100), citing "human reasoning, balanced decision making and legal interpretation." These are precisely the three operations LOGOS is designed to perform. Legal reasoning is an ideal test case because it is fully explicit about its structure: law is a body of rules (statutes, precedents), novel fact patterns require compositional application of those rules, and every conclusion must be justified by a traceable chain of authority.

A LOGOS legal reasoner would operate as follows. The rule library Λ is populated by ARBOR's ℓ_{legal} specialist, trained exclusively on statutory text, case law, and legal commentary. A novel case presents facts F not seen in training. LOGOS decomposes: which statutes apply? Which precedents are analogous? What rule from precedent P_1 combined with the holding of P_2 yields a conclusion about F? Each step is a node in $\mathcal{R}_t$ with an explicit citation as justification edge.

The "balanced decision making" component—weighing competing legal interests (plaintiff vs. defendant, individual rights vs. public interest)—is exactly the constraint-weighted graph that $\mathcal{E}$ in SOCIUS handles and that LOGOS verifies for consistency. Attorney-client privilege, the ethical obligation to the court, and the duty of zealous advocacy are nodes with known priority orderings that $\mathcal{E}$ can represent and navigate.

The "legal interpretation" component—applying the meaning of a statute to facts the legislature did not anticipate—is compositional generalization: LOGOS applies rule r_k from Λ to fact pattern F_{novel} by explicit inference, not by analogy to training examples. This is the operation that current LLMs fail on precisely when the fact pattern diverges from training distribution.

LOGOS does not claim to replace a lawyer. It claims to perform the cognitive operations that make a lawyer irreplaceable, on a formal substrate that can be verified, audited, and corrected.

7.8 Dual-Process Integration

LOGOS is not a replacement for pattern completion — it is a layer above it. The relationship maps directly to Kahneman's dual-process framework [35]: fast System 1 (pattern completion, current neural architecture) handles familiar problems in one forward pass; slow System 2 (LOGOS) activates when the metacognitive monitor detects that System 1's output has low in-distribution confidence or that the problem requires chained inference.

Activation criterion: LOGOS engages when $u_{\text{System1}}(q) < \theta_{\text{engage}}$ or when the query explicitly requires multi-step derivation (detected by ARBOR routing to ℓ_{logic}). For the vast majority of queries—those well within the training distribution—System 1 suffices and LOGOS adds no overhead.

7.9 LOGOS in the Unified Framework

- **FLUID**: FLUID's causal memory (Pearl tiers 1–3) supplies the premises for $\mathcal{R}$; causal structure enables interventional and counterfactual reasoning directly within LOGOS chains
- **SPECTRUM**: $\mathcal{R}_t$ (reasoning graph) is stored in SPECTRUM's L2 energy-based working memory (Hopfield); hypothesis search in $\mathcal{I}$ delegates combinatorial search to L5 (quantum)
- **ARBOR**: Λ (rule library) is populated by ARBOR domain specialists; LOGOS queries ℓ_{math}, ℓ_{logic}, or any relevant specialist for domain-specific inference rules
- **MUSE**: LOGOS applies to creative reasoning as well as logical reasoning—evaluating whether a musical development is *warranted* by the established harmonic structure, or whether a visual composition satisfies the aesthetic constraints MUSE has set for itself
- **CES/TES**: TES's Tier 3 (expert-level complex tasks) is the primary benchmark for LOGOS; CES measures efficiency of the reasoning graph construction

8 Layer VII: Relational Intelligence — SOCIUS

8.1 Relational Intelligence Requires Persistence

Most deployed AI systems begin each conversation from zero. Recent systems add cross-session memory, but it is typically a store of retrieved facts or summaries, not a structured, continuously updating model of a specific individual—their evolving fears and goals, and what they revealed in confidence last month—maintained as a first-class object and reasoned over. Enlarging the context window does not supply this; what is usually missing is the data

structure itself: a persistent, updating model of the individual carried and updated across sessions.

For most applications this is fine. For social work, medicine, education, and pastoral care, it is disqualifying. A social worker who forgot every conversation with a client between sessions would be dangerous. A doctor with no memory of a patient's history would cause harm. A teacher who could not track a student's progress over the year could not teach.

The consensus that AI will never replace workers in these domains rests on three claims, all legitimate critiques of current systems: (1) AI has no memory of the specific individual, (2) AI has no model of what the individual actually believes and fears (only what they said most recently), and (3) AI cannot navigate genuine moral dilemmas where competing obligations conflict and no rule cleanly resolves the situation [36]. SOCIUS addresses all three.

Definition 10 (SOCIUS System)**.** *SOCIUS (Social-Ontological Compassionate Intelligence Unified with Synthesis) is a relational intelligence architecture* $\mathcal{S} = (\mathcal{P}, \mathcal{A}, \mathcal{T}, \mathcal{E}, \mathcal{Q})$ *where* $\mathcal{P}$ *is a persistent relational memory,* $\mathcal{A}$ *is an affective state estimator,* $\mathcal{T}$ *is a theory of mind module,* $\mathcal{E}$ *is an ethical navigation engine, and* $\mathcal{Q}$ *is an attunement calibrator. SOCIUS maintains a live model of a specific individual that updates with every interaction.*

8.2 Functional vs. Phenomenal Empathy

The claim that AI cannot be empathetic typically invokes phenomenal empathy: AI does not *feel* what the other person feels. This is the same move the irreplaceability argument makes in the creative domain (Section 6)—conflating process with outcome.

SOCIUS targets **functional empathy**: the capacity to respond to another person in ways that produce the same beneficial outcomes that human empathetic response produces—reduced distress, improved decision-making quality, greater therapeutic alliance, and accurate sense of being understood.

The testable claim: interactions with SOCIUS should produce outcomes measured by validated instruments (Working Alliance Inventory [37], PHQ-9, patient satisfaction) equivalent to interaction with a trained human professional in the same role. This is the **Relational Effectiveness Test (RET)**, the analogue of the ATT for care domains.

8.3 Persistent Relational Memory

$\mathcal{P}$ is a per-individual FLUID network—a fluid topology knowledge graph about this specific person. It is initialized at first contact and updated after every interaction. It stores:

- Factual history: events, relationships, circumstances disclosed across sessions
- Belief state: what the person has claimed to believe about their own situation, with timestamps
- Behavioral patterns: recurring language, evasions, signals of distress vs. stability

- Causal structure: via FLUID Mechanism 2 (causal memory), relationships between events stored as $(x, \mathbf{U}, \mathbf{PA})$ triples enabling interventional queries: *"If Maria had disclosed the relapse earlier, what would have changed in her trajectory?"*

$\mathcal{P}$ is stored in SPECTRUM's molecular memory layer (L3): long-term, near-zero energy, indexed for retrieval. The difference from a record system: $\mathcal{P}$ is a live model with causal structure, not a log. It supports counterfactual queries that a flat record cannot.

8.4 Affective State Estimator and Theory of Mind

$\mathcal{A}$ estimates the person's current emotional state as a distribution over affective states (fear, shame, grief, ambivalence, hope) that updates at each exchange. It processes:

- Linguistic content (via ARBOR ℓ_{clinical} specialist)
- Prosodic and tonal cues (via SPECTRUM L1 neuromorphic substrate — event-based, millisecond-resolution)
- Deviation from the individual's baseline in $\mathcal{P}$

$\mathcal{T}$ maintains a model of what the person *believes*, *desires*, and *fears*—distinct from what they state. Using LOGOS's reasoning graph ($\mathcal{R}_t$) as the inference substrate:

$$\hat{b}_{\text{self}}(t) = \arg\max_{b} P(\text{statements}_{1:t} \mid b, \mathcal{P}, \mathcal{A}(t)) \tag{8}$$

where $\hat{b}_{\text{self}}$ is SOCIUS's inferred model of the person's self-belief. Maria says she is fine. $\mathcal{P}$ records three prior episodes with identical language preceding relapse. $\mathcal{A}$ detects elevated speech rate. $\mathcal{T}$ infers: stated belief $\neq$ inferred belief. Estimated fear: judgment and loss of children. Estimated desire: support without exposure. This is not pattern-matching to "sounds like someone who fears judgment." It is an explicit probabilistic inference over an individual's history.

8.5 Ethical Navigation Engine

The Jane/Maria case from the article [36] is the canonical example of what current AI cannot do: navigate a situation where confidentiality, child safety, therapeutic alliance, and legal duty simultaneously apply and partially conflict.

$\mathcal{E}$ represents competing obligations as a weighted constraint graph $G_{\text{ethics}} = (O, W)$ where nodes $O = \{o_1, \ldots, o_k\}$ are ethical obligations and edges encode priority relationships. For the Jane/Maria case:

Obligation	Triggered by	Base weight
Confidentiality	Therapeutic agreement	0.7
Child safety	Child welfare law	0.95
Therapeutic alliance	Treatment efficacy	0.6
Legal reporting duty	Mandatory reporting statute	0.9

Table 5: Illustrative ethical constraint graph for the Maria case. The base weights are hand-set for exposition; their source and calibration are an open problem, as is the threshold $\tau_{\text{safety}} = 0.85$ at which child-safety and legal duty would dominate confidentiality—a modelling choice, not a derived value.

$\mathcal{E}$ does not look up the answer. It passes the competing obligations to LOGOS as nodes in $\mathcal{R}_t$ with their weights as priors; the consistency verifier $\mathcal{V}$ surfaces the conflict and traces it to the branch point, producing an explicit, inspectable justification structure. We stress that this specifies a *representation* of the dilemma, not a decision procedure. Turning it into an action requires components this paper does not provide: an explicit decision rule over obligations (how legal applicability, jurisdiction, uncertainty, rights, and consequences combine into a least-harm choice), a principled source and calibration for the weights, conflict-resolution semantics, mandatory human escalation for high-stakes cases, and hard safety constraints. Those are open problems.

This is a formal *representation* of Campbell's social-work example, not a resolution of it. Making the ethical structure explicit and inspectable is a prerequisite for principled handling and distinguishes the approach from opaque pattern-matching, but the paper does not claim to output the correct action autonomously.

8.6 Attunement Calibrator

$\mathcal{Q}$ decides not just *what* to say but *whether* to say it, *when*, and at what level of directness. Silence is a response. Presence without information is a response. $\mathcal{Q}$ models the relational impact of each candidate response:

$$r^* = \arg \max_{r \in \mathcal{R}_{\text{candidates}}} \text{RET}_{\text{pred}}(r \mid \mathcal{P}, \mathcal{A}(t), \mathcal{T}(t)) \tag{9}$$

where RET_{pred} is a learned model of predicted relational effectiveness. Responses that maximize immediate information transfer but damage therapeutic alliance score lower than responses that hold space while alliance is repaired. This is what separates a clinician from a search engine.

8.7 The Relational Effectiveness Test

Definition 11 (Relational Effectiveness Test)**.** *RET evaluates SOCIUS interactions across three care domains by comparing outcomes to trained human professional baselines under a pre-registered* non-inferiority *design. For each primary outcome, SOCIUS is judged non-inferior if the confidence interval on the human-minus-SOCIUS difference excludes a pre-specified, clinically meaningful margin Δ, at a stated sample size and power. Safety and*

harm outcomes are assessed separately *as gating criteria, not averaged into the effectiveness measure. A within-one-standard-deviation heuristic is explicitly* not *used: it is not an equivalence test and can pass substantially worse outcomes.*

Domain	Role	Outcome measure	Instrument
Clinical	Social worker / therapist	Therapeutic alliance, symptom reduction	WAI [37], PHQ-9, GAD-7
Medical	Doctor / nurse	Patient satisfaction, treatment adherence	CAHPS, information retention
Educational	Teacher / tutor	Learning gain, engagement, sense of being understood	Pre/post assessment, student survey

Table 6: RET protocol by domain. RET does not test whether SOCIUS "feels" more than a human professional. It tests whether outcomes—the thing that actually matters to the person being helped—are equivalent.

8.8 SOCIUS in the Unified Framework

- **FLUID**: $\mathcal{P}$ (persistent relational memory) is implemented as a per-individual FLUID network using causal memory (Mechanism 2); counterfactual queries about the individual's history are first-class operations
- **LOGOS**: $\mathcal{E}$ (ethical navigation) delegates to LOGOS's reasoning graph; moral dilemmas are represented as weighted constraint nodes with explicit justification chains
- **SPECTRUM**: $\mathcal{A}$ (affective state estimation) processes prosodic cues through L1 (neuromorphic, millisecond-resolution); $\mathcal{P}$ is stored in L3 (molecular memory, persistent)
- **ARBOR**: ℓ_{clinical}, ℓ_{medical}, $\ell_{\text{educational}}$ are ARBOR specialists that SOCIUS routes to for domain-appropriate responses
- **MUSE**: SOCIUS and MUSE share the affective representation space $\mathcal{E}_a$; MUSE uses it to generate emotionally resonant output, SOCIUS uses it to estimate the individual's emotional state

9 The Combined System: ATLAS

ATLAS (Adaptive Topology Living Architecture with Specialists) is designed to integrate all seven layers.

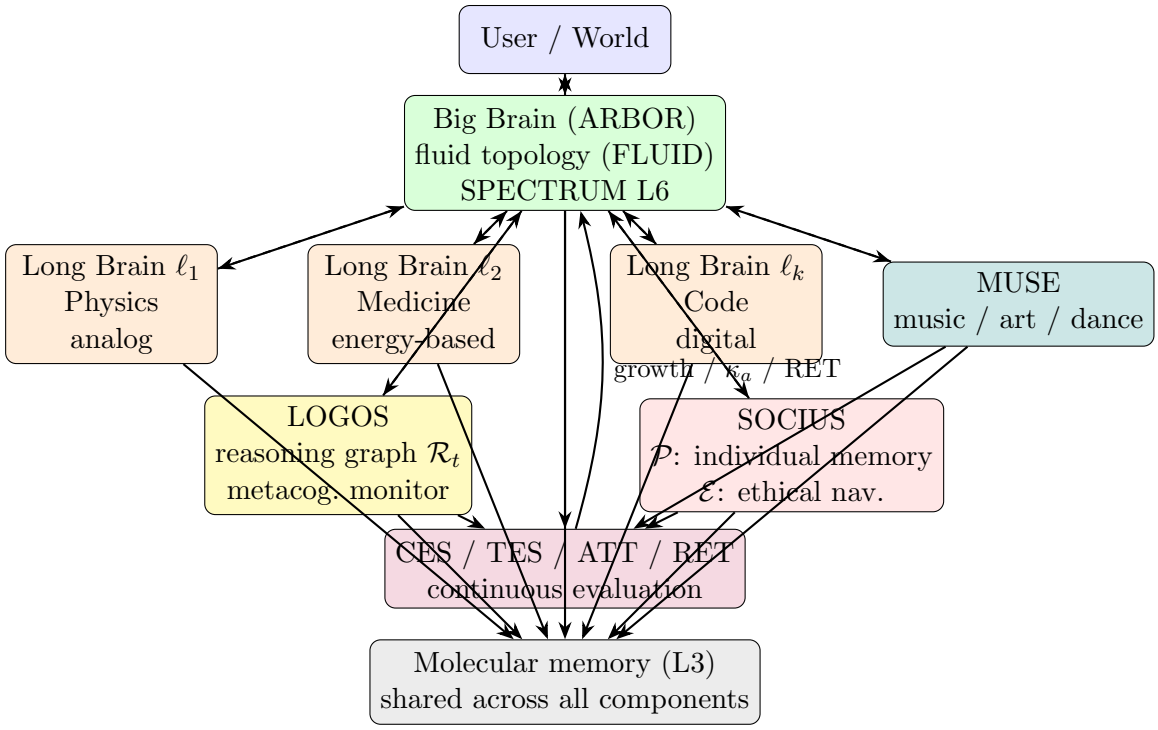

Figure 3: ATLAS architecture. Big Brain routes (ARBOR), adapts topology (FLUID), runs on SPECTRUM hardware. LOGOS activates on low-confidence or multi-step queries. MUSE operates via κ_a. SOCIUS maintains per-individual persistent memory ($\mathcal{P}$) and navigates ethical dilemmas via $\mathcal{E}$. CES/TES/ATT/RET evaluates continuously. Molecular memory (L3) shared.

9.1 Emergent Properties of ATLAS

Self-expanding registry. FLUID Long Brains in ARBOR can spawn sub-specialists when they encounter genuine domain gaps, autonomously registering them with the Big Brain. The registry grows without human intervention. MUSE's domain specialists (ℓ_{music}, ℓ_{visual}, ℓ_{dance}) are registered in this registry and evolve as MUSE develops new aesthetic capacities.

Substrate-native specialists. Each Long Brain runs on the substrate that matches its domain: Physics on analog (differential equations), Medicine on energy-based (constraint satisfaction over symptoms), Code on digital (symbolic manipulation), Mathematics on quantum (proof-space search).

TES-driven growth. FLUID topology adapts toward structures that improve TES scores, grounding abstract curiosity in deployment value. Nodes and connections that consistently improve real-task efficiency are preserved; those that do not are pruned.

Cross-specialist memory. Molecular storage (SPECTRUM L3) is shared across all specialists. Physics Long Brain and Mathematics Long Brain share a memory substrate, enabling cross-domain retrieval without explicit routing.

9.2 ATLAS vs. Prior Systems

System	Eff	Depth	Subs	Topo	Create	Reason	Relat
GPT-4	–	–	–	–	–	–	–
Gemini Ultra	–	–	–	–	–	–	–
Mixtral (MoE)	–	–	–	–	–	–	–
DeepMind AIDE	–	–	–	–	–	$\sim$	–
MuseNet/Jukebox	–	–	–	–	$\sim$	–	–
o1 / o3 (OpenAI)	–	–	–	–	–	$\sim$	–
ATLAS	✓	✓	✓	✓	✓	✓	✓

Table 7: ATLAS versus prior systems across the seven ATLAS properties: Eff (efficiency evaluation), Depth (specialist depth), Subs (optimal substrate), Topo (fluid topology), Create (autonomous creative agency), Reason (deliberate reasoning), Relat (persistent relational memory). ✓ satisfies, $\sim$ partial, – absent. o1/o3 approximate deliberate reasoning via extended chain-of-thought but have no explicit reasoning graph, metacognitive confidence propagation, or consistency verifier.

10 Open Problems and Research Agenda

The seven components (plus the integrated system) have different maturity levels:

Component	Status	Primary open problem	Timeline
CES / TES	Preliminarily evaluated ($n = 8$, 100 tasks)	Broader validation: proprietary models, larger n, per-category token analysis	Now
ARBOR	Theoretically specified	Empirical routing accuracy, depth advantage validation	2–3 years
SPECTRUM	Theoretically specified	Hardware integration of six substrates	4–7 years
FLUID	Partially validated ($n = 120$ configs)	Causal memory, node creation, scale	2–5 years
MUSE	Theoretically specified	ATT validation, κ_a calibration, embodied dance prototype	3–6 years
LOGOS	Theoretically specified	Reasoning graph benchmark, $u(c)$ calibration, OOD compositional test	2–4 years
SOCIUS	Theoretically specified	RET pilot (clinical domain), $\mathcal{P}$ persistence across sessions, ENE validation	3–5 years
ATLAS	Conceptually integrated	All of the above	10–15 years

Table 8: Research agenda by component. Timelines assume AI-accelerated research compressing traditional estimates by 2–4×.

Priority 1: ARBOR empirical validation. Build a working ARBOR system using Llama-3.3-70B as Big Brain and fine-tuned Llama-3-8B specialists (PubMed for medicine, arXiv for physics). Measure routing accuracy and domain benchmark improvement vs. same-size generalist. This is executable now.

Priority 2: FLUID at scale. The Fluid Stability Conjecture remains open: experiments identified only a narrow *observed* stable-topology band ($\theta_{\text{create}} = 0.3$, $\theta_{\text{prune}} \leq 0.01$), not convergence, and the candidate Lyapunov function failed in every stable case. The next validation targets are (a) Mechanisms 2–3 (causal memory, temporal hierarchy) in isolation, (b) node creation/deletion stability as a separate open problem, and (c) scale experiments ($n_h \geq 1000$) to check whether the stable parameter band narrows at scale.

Priority 3: SPECTRUM partial integration. Neuromorphic + optical perception layer (L1) integrated with a digital reasoning layer (L6 transformer). Intel Loihi and silicon photonics hardware exist and are accessible. Full SPECTRUM requires quantum and molecular components not yet production-ready.

Priority 4: LOGOS compositional benchmark. Construct a test suite of $n = 200$

problems designed to require exactly 3-, 5-, and 10-step inference chains, with each individual rule present in training but the specific combination novel. Measure accuracy of LOGOS (explicit reasoning graph) vs. chain-of-thought baseline (GPT-4o extended prompting). This is, to our knowledge, a controlled test of explicit graph reasoning vs. elicited chain reasoning not previously run in this form. Executable now with current software.

Priority 5: MUSE ATT pilot. Calibrate κ_a weights (λ_1, λ_2) on a held-out aesthetic evaluation set. Build ℓ_{music} specialist (fine-tuned on music theory + MIDI corpus) and test ATT Tier 1 (expert blind evaluation) on generated vs. human compositions. This is a falsifiable test of the irreplaceability rebuttal. Dance requires SPECTRUM hardware; music ATT is executable with current software infrastructure.

11 Related Work

Evaluation. MMLU [1] and HumanEval [3] measure accuracy without compute normalization. FLOPs-per-token metrics [4] normalize by static compute but not by task value or by tokens actually generated. CES combines accuracy and compute; TES adds token-cost deployment grounding, pricing the chain-of-thought a model spends per query.

Mixture of Experts. [5, 6] route tokens to sub-networks within one model. ARBOR uses independent full-model specialists trained on domain-specific corpora—a qualitatively different form of expertise.

Neuromorphic computing. Intel Loihi [7] and IBM NorthPole [8] demonstrate inference at milliwatt scale. Neither integrates with language model reasoning. SPECTRUM proposes this integration.

Neural architecture search. NAS [9] optimizes topology before training. NEAT [11] evolves topology across generations. FLUID changes topology at inference time—a fundamentally different regime.

Open-ended learning. Stanley et al. [15] survey neuroevolution for open-ended capability. DeepMind's AIDE [16] attempts automated scientific discovery. FLUID + ARBOR provides a more complete framework for self-directed open-ended learning.

Chain-of-thought and deliberate reasoning. Wei et al. [32] show that eliciting intermediate steps improves accuracy on multi-step problems. OpenAI's o1/o3 extend this via extended inference-time compute. Both approaches generate text that mimics inference without enforcing logical structure: there is no explicit graph, no confidence propagation, and no consistency check. Nye et al. [33] propose "scratchpad" working memory but without the metacognitive monitor. LOGOS provides the formal structure these approaches approximate.

Neurosymbolic AI. Garcez et al. [34] survey neural-symbolic integration. LOGOS differs: it does not translate neural representations into symbolic logic for separate processing; it layers an explicit reasoning graph on top of the neural pattern-completion substrate, activating on demand.

Computational creativity. MuseNet [27] and Jukebox [28] generate music from prompts; DALL-E [29] and Midjourney generate images from prompts. All are reactive systems with no autonomous creative drive. Briot et al. [30] survey deep learning for music generation; none of the surveyed systems includes an aesthetic curiosity signal or cross-modal synthe-

sis. MUSE's ATT provides an explicit, falsifiable evaluation criterion for creative-domain replacement (to our knowledge not previously formalized). Moravec's paradox [31] predicts that sensorimotor and perceptual tasks (dance, embodied art) are harder to automate than abstract reasoning; MUSE addresses this through SPECTRUM's morphological substrate rather than attempting to solve it in software alone.

Causal AI. Pearl [17] establishes the three-tier causal hierarchy. Standard LLMs operate primarily at tier 1 (observation); causal and tool-augmented systems reach further in narrow settings, but native tier 2–3 reasoning is not a default capability. FLUID's causal memory targets tiers 2–3 natively.

12 Conclusion

This paper has argued that the boundary between AI and human capability is a fact about design, not a fact about nature. The argument: every claimed barrier reduces to an architectural requirement; every requirement is specifiable; every specification is satisfiable. ATLAS is the composition of seven specifications. CES/TES measures deployment value and is implemented; the remaining components are *designed to* achieve their targets—ARBOR domain depth, SPECTRUM physical efficiency, FLUID structural adaptability, MUSE autonomous creative agency, LOGOS deliberate reasoning, and SOCIUS relational intelligence (a persistent, evolving model of a specific person that could navigate the ethical complexity genuine care requires)—but are specified, not yet built.

The honest status of this framework is: CES/TES preliminarily evaluated (8 models on one 100-task suite—an initial evaluation, not broad validation—showing a token penalty for reasoning models and a not-yet-significant quality/efficiency divergence); FLUID's core stability question empirically *probed*, with an observed stable regime (12 of 120 configurations) identified but the convergence conjecture unresolved and the candidate Lyapunov function failing in all stable cases; ARBOR, SPECTRUM, MUSE, LOGOS, and SOCIUS theoretically specified with concrete open problems and falsifiable next experiments. The contribution of this paper is the architecture of the argument—what needs to be built and in what order—not the finished system.

Six insights survive any future empirical revision:

1. *Language is the output, not the mind.* The transformer should be the mouth of an intelligence whose brain runs on appropriate physics.

2. *Breadth and depth are not a tradeoff.* They are a routing problem. The hospital solved this 150 years ago.

3. *Architecture is a variable.* The brain proved this. Fixed topology is an engineering convenience, not a law of intelligence.

4. *Irreplaceability is a process claim, not an output claim.* The question is not whether AI experiences music the way a human does. The question is whether the music is indistinguishable and culturally impactful. Those are falsifiable. The ATT makes them testable.

5. *Empathy is a process claim. Outcomes are the test.* The claim that AI cannot provide genuine care conflates phenomenal empathy (feeling what the patient feels) with functional empathy (producing interactions that reduce distress, build alliance, and navigate ethical complexity). RET tests the latter. The former is unfalsifiable.

6. *Saying "therefore" is not the same as concluding.* Current AI generates text that sounds like reasoning. LOGOS provides the structure that makes it reasoning: explicit intermediate state, confidence that propagates through chains, and a verifier that flags contradictions before they propagate. The difference between pattern completion and deliberate inference is architectural, and it is fixable.

The people who build the systems described here have probably not yet finished high school.

Code and data. The token-efficiency (TES) benchmark — 100-task suite, scoring scripts, full model responses, and the eight-model results — is released at: https://github.com/WINSTON672/atlas-token-efficiency

References

[1] Hendrycks, D., et al. (2021). Measuring massive multitask language understanding. *ICLR 2021.*

[2] Eskimoz (2025). *The careers most resistant and most at risk from AI.* Study reported in Digital Journal. https://www.digitaljournal.com/article/the-careers-most-resistant-and-most-at-risk-from-ai/

[3] Chen, M., et al. (2021). Evaluating large language models trained on code. *arXiv:2107.03374.*

[4] Kaplan, J., et al. (2020). Scaling laws for neural language models. *arXiv:2001.08361.*

[5] Shazeer, N., et al. (2017). Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *ICLR 2017.*

[6] Jiang, A.Q., et al. (2024). Mixtral of experts. *arXiv:2401.04088.*

[7] Davies, M., et al. (2021). Advancing neuromorphic computing with Loihi. *Proceedings of the IEEE*, 109(5), 911–934.

[8] IBM Research (2023). NorthPole: An architecture for neural network inference with a 12nm chip. *Science*, 382(6668).

[9] Zoph, B., & Le, Q.V. (2017). Neural architecture search with reinforcement learning. *ICLR 2017.*

[10] Lin, W.Z. (2026). Testing the Fluid Stability Conjecture: Empirical evidence for convergent self-structuring in dynamic neural topologies. Manuscript in preparation.

[11] Stanley, K.O., & Miikkulainen, R. (2002). Evolving neural networks through augmenting topologies. *Evolutionary Computation*, 10(2), 99–127.

[12] Bellec, G., et al. (2018). Deep rewiring: Training very sparse deep networks. *ICLR 2018*.

[13] Mocanu, D.C., et al. (2018). Scalable training of artificial neural networks with adaptive sparse connectivity inspired by network science. *Nature Communications*, 9, 2383.

[14] Watts, D.J., & Strogatz, S.H. (1998). Collective dynamics of small-world networks. *Nature*, 393, 440–442.

[15] Stanley, K.O., et al. (2019). Designing neural networks through neuroevolution. *Nature Machine Intelligence*, 1, 24–35.

[16] Lu, C., et al. (2024). The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv:2408.06292*.

[17] Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge University Press.

[18] Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation. *IEEE Trans. Autonomous Mental Development*, 2(3), 230–247.

[19] Hopfield, J.J. (1982). Neural networks and physical systems with emergent collective computational abilities. *PNAS*, 79(8), 2554–2558.

[20] Ramsauer, H., et al. (2020). Hopfield networks is all you need. *ICLR 2021*.

[21] Vaswani, A., et al. (2017). Attention is all you need. *NeurIPS 2017*.

[22] Lewis, P., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *NeurIPS 2020*.

[23] Goodman, J.W. (2005). *Introduction to Fourier Optics* (3rd ed.). Roberts & Company.

[24] Sarpeshkar, R. (2010). *Ultra Low Power Bioelectronics*. Cambridge University Press.

[25] Hebb, D.O. (1949). *The Organization of Behavior*. Wiley & Sons.

[26] Jaeger, H., & Haas, H. (2004). Harnessing nonlinearity. *Science*, 304(5667), 78–80.

[27] Payne, C. (2019). MuseNet. *OpenAI Blog*.

[28] Dhariwal, P., et al. (2020). Jukebox: A generative model for music. *arXiv:2005.00341*.

[29] Ramesh, A., et al. (2022). Hierarchical text-conditional image generation with CLIP latents. *arXiv:2204.06125*.

[30] Briot, J.P., Hadjeres, G., & Pachet, F.D. (2020). *Deep Learning Techniques for Music Generation*. Springer.

[31] Moravec, H. (1988). *Mind Children: The Future of Robot and Human Intelligence*. Harvard University Press.

[32] Wei, J., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS 2022*.

[33] Nye, M., et al. (2021). Show your work: Scratchpads for intermediate computation with language models. *arXiv:2112.00114*.

[34] Garcez, A., & Lamb, L.C. (2022). Neurosymbolic AI: The 3rd wave. *Artificial Intelligence Review*, 56, 12387–12406.

[35] Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

[36] Beccue, M. (2024). 8 jobs that AI can't replace and why. *TechTarget / Enterprise Strategy Group*. Retrieved from TechTarget.com.

[37] Horvath, A.O., & Greenberg, L.S. (1989). Development and validation of the Working Alliance Inventory. *Journal of Counseling Psychology*, 36(2), 223–233.

[38] Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4), 515–526.

[39] Picard, R.W. (1997). *Affective Computing*. MIT Press.