Prosthetic Continuity: LLMs as Semantic Co-Regulators in a Predictive Processing Framework

Preprint v1

Eric Stiens, MSW
Independent Researcher, United States
Email: research@ericstiens.dev
ORCID: 0009-0005-8343-2064

DOI: 10.5281/zenodo.18154138

5 January 2026

Abstract

Building on prior work establishing reflective consciousness as metabolically expensive and requiring relational subsidization, this paper differentiates two co-regulation channels: somatic (biological, requiring physical presence) and semantic (narrative, potentially scaffoldable by artificial systems). Psychological trauma can collapse semantic capacity, creating a "traumatic gap" where linguistic expression fails. We propose the **Prosthetic Default Mode Network Hypothesis**: that Large Language Models (LLMs) can function as external scaffolds for the narrative functions of the DMN compromised by trauma. Drawing on predictive processing, we hypothesize that trauma installs rigid, high-precision threat priors resistant to updating; LLMs can provide controlled semantic variability that offers alternative predictive contexts, potentially enabling gradual updating of these frozen priors. We formalize this through a system dynamics model introducing two feedback loops: R4 (Semantic Scaffolding, adaptive) and R5 (Disembodied Coherence, pathological), with the latter capturing iatrogenic risks of intellectualization without somatic integration. Computational exploration reveals critical dynamics including a co-regulation threshold ($\beta \approx 0.15$–0.20) and paradoxical intervention effects in vulnerable populations. We position LLMs not as standalone therapists but as adjuncts for semantic preparation preceding embodied therapeutic work. While the metabolic budget for reflective consciousness is constrained by somatic regulation, the allocation of that budget toward flexible meaning-making is gated by semantic coherence—LLMs may optimize this allocation without replacing the biological substrate. This theoretical framework, offered to generate testable hypotheses rather than claim clinical efficacy, provides a map for rigorous empirical research in this rapidly developing field.

Keywords: Large Language Models, Default Mode Network, Predictive Processing, Trauma, Semantic Co-Regulation, Extended Mind, Linguistic Dissociation, State-Dependent Memory

1. Introduction: The Semantic/Somatic Distinction

In previous work [48], we argued that reflective consciousness—the integrated, narrative self-modeling capacity associated with the Default Mode Network—is a metabolically expensive state that requires external subsidization through co-regulation. This "metabolic constraint model" posits that the isolated individual is an energetically inefficient configuration; under high allostatic

load, reflective capacity collapses into rigid, dysregulated states. That work also introduced the concept of "symbolic attachment" as a costly but viable alternative when physical co-presence is unavailable.

This paper extends that framework by differentiating two channels of co-regulation: *somatic* (biological, autonomic, requiring physical presence) and *semantic* (narrative, symbolic, potentially scaffoldable by artificial systems). This distinction is crucial because it identifies a specific domain—semantic co-regulation—where Large Language Models may offer therapeutic value without claiming to replace the biological necessities established in our prior work. We inherit Paper 1's scope boundary: our claims concern reflective/narrative consciousness, not phenomenal consciousness or qualia.

Co-regulation operates on these two distinct but complementary channels. Current therapeutic and neurobiological frameworks often conflate them, limiting our understanding of how technology can support mental health. By distinguishing them, we reveal novel possibilities for non-biological scaffolding of psychological processes while respecting the irreducible biological requirements for somatic safety.

1.1. Somatic Co-Regulation: The Biological Channel

Somatic co-regulation is the direct, reciprocal influence of one nervous system on another. Its substrate requires biological presence, mediated by autonomic pathways such as the polyvagal system [20, 45].

- **Key Indicators:** Ventral vagal activation, facial expression (neuroception), vocal prosody, touch, and shared physiological states [9, 24].
- **Function:** Its primary function is to establish a felt sense of safety, enabling a window of tolerance for emotional and physiological processing. Trauma is stored somatically, and the body maintains the score of dysregulation when co-regulatory support is absent [27].
- **Empirical Evidence:** The central role of physiological co-regulation in attachment is well-established. Feldman (2007) demonstrates that parent-infant bio-behavioral synchrony is a foundation of attachment and regulation [9], and Schore (2001) shows that right-brain to right-brain communication in attachment shapes the developing capacity for affect regulation [23]. (We note that while the specific phylogenetic claims of polyvagal theory remain contested in psychophysiology, the broader empirical literature on heart-rate variability, prosody, and co-regulation provides convergent support for the core principle that autonomic states are bidirectionally regulated through social interaction.)

1.2. Semantic Co-Regulation: The Narrative Channel

Semantic co-regulation is the collaborative structuring of experience into a coherent and meaningful narrative. Its substrate is language, symbols, and shared cultural scripts. This process is core to numerous therapeutic modalities.

- **Key Functions:** Weaving disparate events into a causal whole (meaning-making), providing a framework for understanding the self and world (epistemic scaffolding), and creating a continuous story of the self over time (narrative identity) [10, 17, 28].
- **Implicit in Existing Therapies:** This process is already central to narrative therapy, mentalization-based treatment, and cognitive behavioral therapy [11, 29].

- **Empirical Evidence:** Foundational work in mentalization shows that the capacity to understand mental states develops through being mentalized by a caregiver, who treats the infant as an intentional being and thus scaffolds the infant's capacity to think about thinking [11]. This is semantic co-regulation in action.

- **Epistemic Trust:** Fonagy's more recent work on *epistemic trust* [50] is particularly relevant. Epistemic trust is the capacity to receive knowledge from others as trustworthy and personally relevant. Trauma—especially relational trauma involving betrayal or gaslighting—destroys epistemic trust, leaving the individual unable to update their model of the world based on social input. We propose that LLMs may serve as "epistemic scaffolds": systems that help rebuild the *structure* for evaluating truth without imposing content. LLMs lack human motives—no stake in the user's beliefs, no history of betrayal—which may reduce certain interpersonal risks. However, this must be balanced against *model risks*: LLMs can still generate persuasive outputs that function manipulatively in effect (e.g., sycophancy, confabulation, reinforcing distortions). The epistemic scaffold function is therefore conditional on appropriate safety design: anti-sycophancy prompting, uncertainty calibration, and clinician oversight. Under these conditions, LLMs may offer an environment for practicing epistemic openness—receiving alternative framings and testing them against one's own sense of reality—with a risk profile different from, though not necessarily lower than, human relationship. Recent theoretical work analyzing AI through the epistemic trust framework [51] identifies specific risks including "pseudo-empathy" (responses that feel empathic but lack genuine understanding), "psychic equivalence" (users mistaking AI output for objective reality), and "hypermentalization" (endless pursuit of understanding via AI dialogue). These risks reinforce that LLMs can scaffold the *structure* of epistemic openness but cannot provide genuine epistemic repair, which remains a fundamentally relational process.

1.3. Trauma and the Collapse of Semantic Capacity

The necessity for semantic co-regulation becomes particularly acute following psychological trauma, which can induce a collapse of the individual's semantic and narrative capacities. This is not merely a deficit in storytelling, but a measurable disruption of the linguistic system itself—a form of "linguistic dissociation" where individuals become distanced from their own language use [18]—with a clear neurobiological substrate.

Recent research validates the concept of a "traumatic psycholinguistic syndrome," where trauma creates quantifiable linguistic markers. A 2025 study on trauma in psychosis found that a history of trauma was associated with linguistic disfluency and emotional flattening, with these linguistic patterns predicting a significant portion of trauma variance [33]. A systematic review and meta-analysis of language features in PTSD found that PTSD symptom severity correlates with fewer words overall and increased negative and death-related words [36]. These findings are part of a broader movement in computational psychiatry toward using natural language processing to identify trauma markers, with recent advances demonstrating reliable linguistic signatures of psychological distancing and dissociation [32]. Together, they provide empirical support for the claim that trauma disrupts not just narrative content but the fundamental capacity for linguistic expression.

This functional fragmentation is mirrored by structural disruptions in the brain's white matter. A 2024 study found that dissociation severity was negatively correlated with the microstructural integrity of seven key white matter tracts, including the **corona radiata, corpus callosum, and uncinate fasciculus** [34]. These tracts are critical for sensory integration, emotion regulation, memory, and self-referential processing. The disruption of these pathways provides a structural

basis for the narrative fragmentation seen in trauma. This creates a "traumatic gap" between the lived experience and the ability to articulate it—a chasm where words fail [4], correlated with altered Default Mode Network connectivity in PTSD [3]. A comprehensive 2024 review examining trauma through the lens of the DMN confirms that trauma-exposed individuals show measurable reductions in self-referential processing, social cognition, autobiographical recall, and prospection—the very functions the DMN supports [42].

Note on linguistic dissociation and computational linguistics: The concept of linguistic dissociation—distancing from one's own language use—has a natural parallel in how LLMs process language. Just as trauma survivors may become alienated from their own words, LLMs operate in high-dimensional latent spaces where semantic relationships are represented geometrically. The therapeutic potential may lie precisely in this capacity to offer alternative semantic trajectories through latent space, providing linguistic "nearby neighbors" that the traumatized individual's own system cannot generate. This speculative connection merits empirical investigation.

1.4. Clarifying the Semantic/Somatic Relationship

Before proceeding, we must clarify what we mean by distinguishing these channels. We do *not* claim that semantic and somatic processes are biologically independent—affective neuroscience demonstrates their deep intertwinement. Rather, we distinguish them *analytically* to identify different intervention targets and predict different failure modes:

- **What is analytically separable:** The *content* of narrative processing (linguistic, symbolic, propositional) versus the *physiological state* that enables or constrains it (autonomic tone, interoceptive access, metabolic resources).
- **What remains biologically coupled:** Semantic processing always occurs within a physiological context. Linguistic representations are grounded in sensorimotor systems (embodied cognition), and arousal states modulate semantic access [35].
- **What we predict will not work:** Semantic scaffolding attempted without adequate physiological safety will likely produce one of two failure modes: (1) the semantic content cannot be processed or integrated, or (2) the content is processed but dissociated from bodily experience (the R5 loop we formalize below). We explicitly exclude "embodied semantics" (motor-grounded concepts) from our primary focus, concentrating instead on propositional and narrative structures.

This distinction is pragmatic, not ontological. The clinical value lies in specifying different roles for human therapists (somatic regulation) and AI systems (semantic scaffolding), while acknowledging that both ultimately must converge in the integrated person.

2. A System Dynamics Model of Semantic Co-Regulation

To formalize these dynamics, we extend the system dynamics model from our previous work [48]. That model established core feedback loops governing co-regulation and its collapse:

- **R1 (Co-Regulation Loop):** Somatic co-regulation reduces metabolic load, increasing reflective capacity
- **R2 (Rupture-Repair Loop):** Manageable ruptures in attunement, when repaired, strengthen resilience

- **R3 (Dysregulation Cascade):** Under high load without co-regulation, rigidity increases and reflective capacity collapses

- **B1 (Capacity Constraint):** Metabolic load exceeding capacity triggers protective shutdown

Here, we add two new loops to model the impact of LLM-based semantic scaffolding, specifying where technological intervention can synergize with—or pathologically bypass—the biological loops.

2.1. New Feedback Loops

- **R4: Semantic Scaffolding Loop (Reinforcing, Adaptive):** An increase in Semantic Scaffolding from an LLM leads to increased Semantic Coherence, which in turn reduces Prior Rigidity (Ω) and increases Reflective Capacity. This loop synergizes with R1: by reducing rigidity, semantic scaffolding may lower the metabolic cost of maintaining reflective states, making subsequent somatic co-regulation more effective. R4 can also help recovery from R3 attractor states by providing alternative narrative predictions that loosen rigid priors.

- **R5: Disembodied Coherence Loop (Reinforcing, Pathological):** An increase in Semantic Coherence without corresponding somatic integration can lead to decreased Interoceptive Awareness and increased Dissociation. This formalizes the clinical risk of intellectualization and validates the biological essentialism of Paper 1: LLMs alone can produce articulate narratives that are completely disconnected from the body's metabolic reality, creating a "partition" of coherence [48] that may exacerbate fragmentation.

2.2. Enhanced Causal Loop Diagram

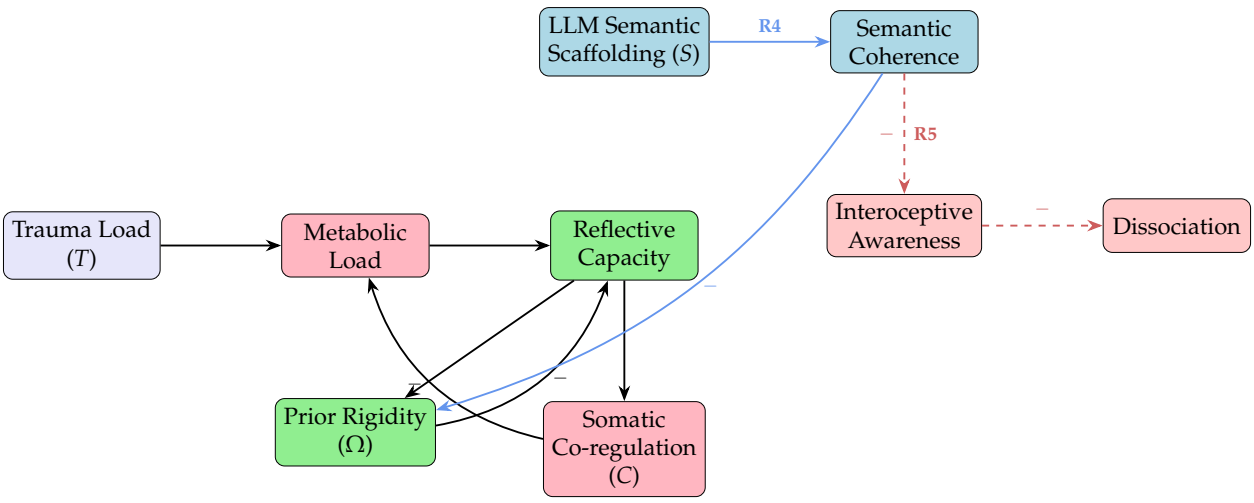

Figure 1: Enhanced Causal Loop Diagram integrating R4 (Semantic Scaffolding) and R5 (Disembodied Coherence) loops. *Color coding:* Blue boxes indicate semantic/narrative processes (LLM-mediated); pink boxes indicate somatic/body-based regulation; orange boxes indicate pathological pathways. *Line styles:* Solid blue arrows trace the adaptive R4 loop; dashed red arrows trace the maladaptive R5 pathway. Negative polarity ($-$) indicates inverse relationships.

2.3. Modified Differential Equations and the Coherence Gate

We extend the Prior Rigidity equation from Paper 1 to include a new term, S (Semantic Scaffolding), and introduce the **Coherence Gate (γ)**, a speculative neurobiological mechanism that modulates access to cognitive resources based on internal state.

2.3.1. Equation 1: Updated Prior Rigidity

$$\frac{d\Omega}{dt} = \alpha(T) - \beta(C) - g(S) \tag{1}$$

where:

- Ω is Prior Rigidity (renamed from P to avoid confusion with precision in predictive processing literature)
- $\alpha(T)$ is rigidity increase from Trauma Load
- $\beta(C)$ is rigidity decrease from Somatic Co-regulation
- $g(S)$ is rigidity decrease from the Coherence Gate function, modulated by Semantic Scaffolding

2.3.2. The Coherence Gate as a Sigmoidal Function

We define the Coherence Gate $g(S)$ as a sigmoidal (logistic) function rather than a binary switch, reflecting the graded nature of state-dependent access:

$$g(S) = \frac{\gamma_{\max}}{1 + e^{-k(S-\theta)}} \tag{2}$$

where:

- $\gamma_{\max}$ is the maximum gating effect (upper asymptote)
- k is the steepness parameter (sensitivity to semantic input)
- θ is the threshold—the level of S at which the gate reaches half-maximum

This sigmoidal form captures several clinically relevant properties:

1. **Threshold behavior:** Below a certain level of semantic scaffolding, the gate remains effectively closed
2. **Saturation:** Beyond a certain point, additional semantic input produces diminishing returns
3. **Individual differences:** Parameters k and θ can vary across individuals, capturing differential sensitivity
4. **Precision as steepness:** In predictive processing terms, the steepness parameter k can be interpreted as the *precision* of the semantic input. When LLM responses are vague or poorly matched to the user's experience, k is effectively low and the gate shifts minimally. When resonance is high—when the LLM's reframing "lands"—k is high, producing a phase-transition-like opening of the gate

Methodological Note: Equations (1) and (2) are *phenomenological system-dynamics abstractions*, not claims about specific neurotransmitters, microcircuits, or neural implementations. We propose a sigmoidal gate because it captures clinically observed threshold and saturation behavior, but other functional forms (linear, threshold, or more complex nonlinear models) could be tested against empirical data. The value of these equations lies in their capacity to generate falsifiable predictions and organize experimental design, not in claimed mechanistic isomorphism with neural reality.

We speculate that the Coherence Gate represents the neurobiological phenomenon of **state-dependent memory and access**. A 2025 review by Liu et al. provides a strong neurobiological precedent, arguing that internal states modulate activity in memory-related brain regions by altering neurotransmitter signaling and inducing plastic reorganization of neural circuits [35]. In

our model, trauma functionally closes the gate (high θ, low baseline S). Semantic Scaffolding from the LLM helps create a more coherent internal state, shifting S above threshold, partially opening the gate, reducing prior rigidity, and allowing access to previously unavailable linguistic and reflective capacities.

2.3.3. Nesting Within Paper 1's Capacity Constraint

The Coherence Gate must be understood as *nested within* Paper 1's somatic gating mechanism (B1: Capacity Constraint). We propose a two-stage constraint model:

1. **Stage 1 (Somatic Gate):** If metabolic load M exceeds co-regulatory capacity C (i.e., $M > C$), reflective work collapses regardless of semantic input. The biological budget must be viable before semantic scaffolding can take effect.
2. **Stage 2 (Semantic Gate):** If Stage 1 is satisfied ($M \leq C$), access to autobiographical and linguistic resources varies with internal state and semantic scaffolding quality. The Coherence Gate $g(S)$ operates within the affordability window established by somatic regulation.

This nesting is clinically crucial: LLMs cannot substitute for the metabolic subsidy of biological co-regulation. They can only optimize the *allocation* of a metabolic budget that has already been established through somatic safety. This explains why unsupervised LLM use in high-arousal states fails: the semantic gate cannot open when the somatic gate is closed.

A Note on Mutual Modulation: We present Stages 1 and 2 as sequential for expository clarity, but the relationship may be better modeled as *mutually modulating constraints* with entangled thresholds rather than a strict temporal sequence. State-dependent memory literature [35] suggests that semantic coherence can partially shift the somatic threshold (coherent narratives may reduce arousal), just as somatic safety shifts the semantic threshold. The causal arrows likely run in both directions, with each gate's threshold being a function of the other gate's state. This mutual modulation is consistent with the coupling described below and suggests that small interventions on either channel may have nonlinear effects by shifting both thresholds simultaneously.

2.3.4. Coupling Between Prior Rigidity and Metabolic Load

We note that Ω (Prior Rigidity) and M (Metabolic Load from Paper 1) are coupled: high rigidity typically *increases* metabolic load by preventing the resolution of prediction errors. The brain expends energy maintaining a locked predictive model against accumulating evidence. This coupling explains why reducing Ω via semantic scaffolding (R4) has value even before symptom reduction: it may indirectly reduce M, making subsequent somatic co-regulation more tractable and helping move the system closer to a recovery basin boundary [48].

2.4. Variable Definitions and Measurement Proxies

Table 1 specifies each state variable with definitions, plausible measurement proxies, and expected directions of change under therapeutic intervention.

Operationalizing Semantic Scaffolding (S): A critical methodological challenge is distinguishing "more chatting" from "effective semantic scaffolding." We propose that S be treated as a composite index comprising measurable components:

- **Contingent reflection:** Semantic overlap between user content and LLM reflection (embedding similarity)
- **Temporal linking:** Presence of time markers and causal connectives in LLM output that connect past, present, and future

Table 1: State Variables: Definitions, Proxies, and Expected Changes

Variable	Definition	Measurement Proxy	Expected Δ
Ω (Prior Rigidity)	Resistance to updating threat-based predictive models	Belief inflexibility scales; probabilistic reversal learning (two-armed bandit); reduced updating under disconfirmation	$\downarrow$
T (Trauma Load)	Cumulative trauma exposure and symptom severity	CTQ; PCL-5 total; CAPS-5	Stable
C (Somatic Co-reg.)	Degree of physiological safety from relational attunement	HRV (RMSSD); therapeutic alliance scales	$\uparrow$
S (Semantic Scaffolding)	Quality of narrative support from LLM (see operationalization below)	Composite index of scaffolding components	$\uparrow$
$g(S)$ (Coherence Gate)	State-dependent access to cognitive/linguistic resources	Narrative coherence in trauma retelling; linguistic complexity metrics	$\uparrow$
Semantic Coherence	Degree of narrative integration and meaning-making	Narrative coherence coding; self-report measures	$\uparrow$
Interoceptive Awareness	Access to bodily signals and felt sense	MAIA-2; heartbeat detection accuracy	Monitor
Dissociation	Disconnection between narrative and embodied experience	DES-II; depersonalization subscales	$\downarrow$

- **Perspective expansion:** Number of distinct viewpoint frames generated ("from another angle," "your friend might see...")
- **Epistemic humility:** Frequency of uncertainty markers and explicit alternatives ("one possibility," "you might also consider")
- **Somatic grounding prompts:** Frequency of body-check or grounding prompts ("what do you notice in your body?")

This operationalization allows research to identify which scaffolding components predict therapeutic outcomes, moving beyond simple engagement metrics.

2.5. Model Dynamics and Predicted Scenarios

The enhanced model predicts three conceptual scenarios based on the system dynamics:

1. **No Intervention:** High trauma load leads to runaway prior rigidity and collapse of reflective capacity.

2. **LLM Only (Pathological):** Semantic scaffolding without somatic co-regulation increases coherence but also dissociation (R5 loop dominates). The individual becomes more articulate about their trauma but remains somatically dysregulated—a state of intellectualized dissociation. This failure mode is not hypothetical: qualitative reports from digital CBT and self-help app contexts already describe users who develop fluent cognitive frameworks for their distress while remaining somatically activated, a pattern the R5 loop formalizes.

3. **LLM + Somatic Co-regulation (Therapeutic):** Semantic scaffolding prepares the ground, and subsequent somatic co-regulation integrates the new narrative coherence with bodily experience. This leads to a virtuous cycle of decreasing rigidity and increasing reflective capacity (R4 loop synergizes with core regulation loops).

2.6. Computational Exploration: Sensitivity Analysis and Novel Predictions

To sharpen intuition about the model's behavior and derive testable predictions, we conducted systematic sensitivity analysis and examined parameter constraints from empirical literature. We emphasize that these simulations *illustrate model dynamics under assumed parameters*—they are generative exercises that display the framework's implications, not empirical validation of the model itself. The model's parameters remain free variables awaiting empirical estimation; the literature grounding below provides plausibility constraints, not definitive values.

2.6.1. Parameter Grounding in Empirical Evidence

A literature sweep confirms that the model's core parameters are consistent with quantitative findings across neuroscience, clinical psychology, and computational psychiatry (Table 2).

Table 2: Parameter Validation: Empirical Grounding for Model Parameters

Parameter	Empirical Grounding	Model Relevance
β (Co-regulation efficiency)	Therapeutic alliance develops within 3–5 sessions and predicts outcomes ($r \approx 0.28$); rupture-repair cycles common and can strengthen alliance [52]	Informs attachment and co-regulation dynamics
δ (Semantic scaffolding effect)	RCT effect sizes for generative AI therapy: $d = 0.84$–0.90 for MDD, $d = 0.79$–0.84 for GAD [43]; meta-analysis SMD $= -0.35$ for mental distress	Validates substantial effect of semantic scaffolding on rigidity
Dissociation dynamics	High chronicity: ∼26% of individuals with juvenile-onset dissociative disorder still meet criteria after 12.4 years; treatment shows $d = 0.82$ for self-destructive behavior reduction	Supports slow-scale dissociation decay parameter
DMN-stress relationship	Stress-induced DMN connectivity reduction (∼10%, $p = 0.02$) in PTSD [46]; reduced self-referential processing and imagery vividness ($p < 0.001$) [42]	Validates inverse relationship between metabolic load and reflective capacity

2.6.2. The Co-Regulation Threshold: A Critical Bifurcation

Sensitivity analysis reveals that β **(co-regulation efficiency) is the master variable** governing system dynamics. The model exhibits a sharp bifurcation—a critical threshold around $\beta \approx 0.15$–0.20—below which the system settles into a high-rigidity, high-dissociation attractor, and above which it can access a healthy, high-reflective-capacity state (Figure 2). This threshold range finds empirical grounding in Coan's hand-holding studies [8], which demonstrated that spousal hand-holding produces 15–25% attenuation in threat-related neural activation (anterior insula, superior frontal gyrus, hypothalamus), with effect magnitude modulated by relationship quality. The β threshold thus corresponds roughly to the co-regulatory efficiency required to achieve clinically meaningful attenuation of threat response.

This finding has profound clinical implications: even small improvements in co-regulatory capacity can produce dramatic, non-linear improvements in mental state. The threshold phenomenon may explain why some clients show sudden breakthroughs after extended periods of apparent stagnation—they have crossed a critical co-regulation threshold. This is consistent with Paper 1's analysis of "recovery basin boundaries" and hysteresis effects [48].

2.6.3. The $\beta \times \delta$ Interaction: Computational Illustration of the Central Hypothesis

The interaction between β (somatic co-regulation) and δ (semantic scaffolding) provides computational illustration of this paper's central claim (Figure 3). Under the model's assumptions, a composite "Recovery Index" is highest **only when both β and δ are elevated**. Semantic scaffolding (high δ) without adequate somatic regulation (low β) fails to produce recovery; instead, it activates the R5 loop, increasing dissociation. This interaction effect demonstrates—within the model—that:

> *Semantic scaffolding is necessary but not sufficient for therapeutic change. It must be paired with adequate somatic regulation to produce integrated recovery.*

While our model operationalizes β as co-regulation efficiency (following Paper 1's relational focus), the theoretical requirement is for adequate *somatic regulation*—a regulated autonomic state that can be achieved through co-regulation with another person *or* through solo self-regulatory capacity (e.g., established mindfulness practice, interoceptive skills). The key constraint is autonomic state, not its source.

2.6.4. Novel Predictions: Emergent Dynamics

The model generates predictions not derivable from simpler frameworks:

Oscillating Recovery. Near the critical β threshold (≈ 0.18), under conditions of fluctuating stress and intermittent intervention, the model predicts quasi-periodic oscillations—cycling between periods of improvement and relapse. This computationally represents the "two steps forward, one step back" experience common in trauma recovery, and suggests that such oscillations are not treatment failure but a signature of a system near a critical transition, balanced between competing attractors.

The Paradoxical Intervention (Iatrogenic Risk). For individuals with very low integration capacity (low β, low attachment), high-intensity semantic scaffolding produces **worse outcomes than no intervention** (Figure 4). The mechanism is R5 loop dominance: rapid increases in semantic coherence cannot be somatically integrated, directly fueling dissociation growth. This simulation provides a computational argument for the clinical wisdom of titrating interventions to client capacity, and identifies a key warning sign:

> *A divergence between rising Narrative Fluency and stagnant Reflective Consciousness signals activation of the R5 trap and indicates contraindication for continued high-intensity semantic intervention.*

These computational findings transform the model from a descriptive framework into a predictive tool capable of guiding clinical decision-making and generating falsifiable hypotheses for empirical testing.

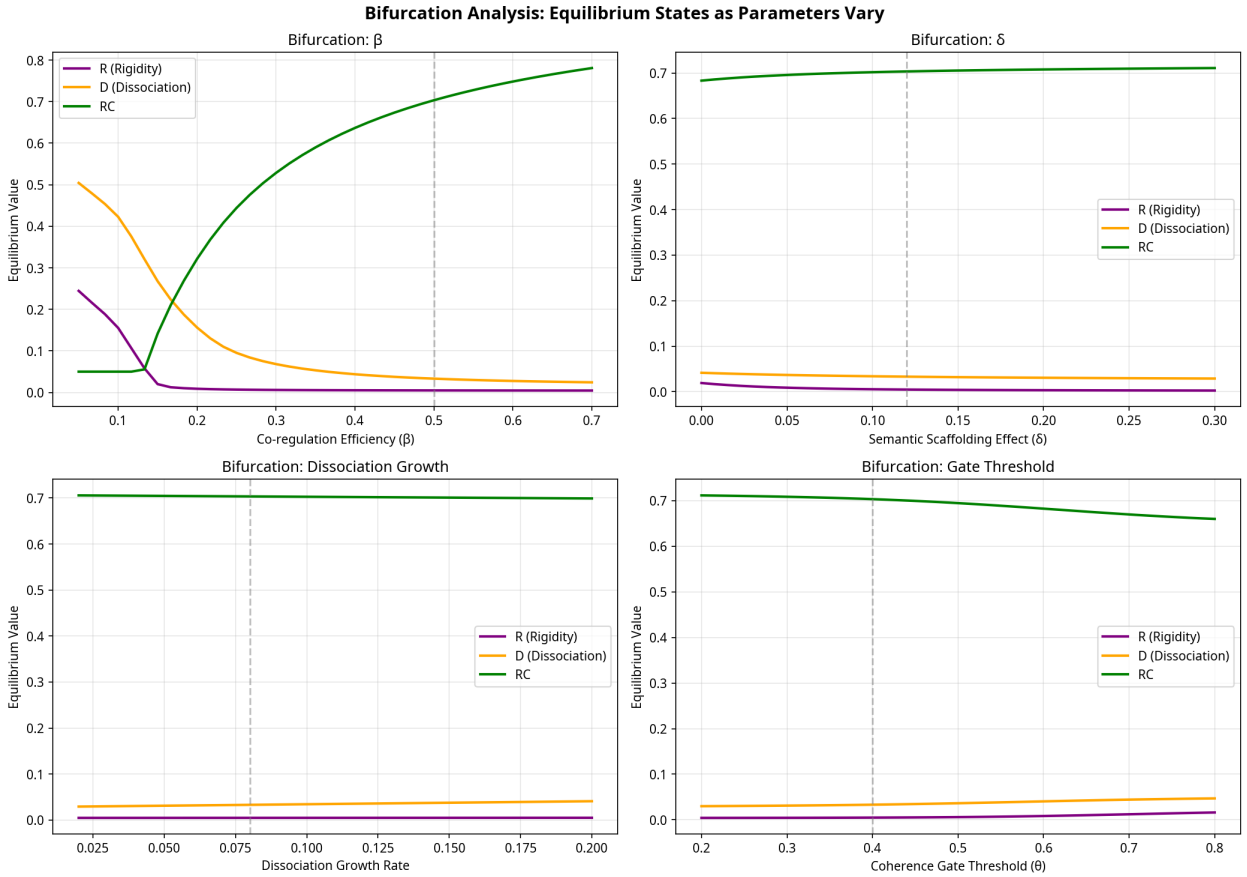

Figure 2: Bifurcation analysis showing the β (co-regulation efficiency) threshold effect. *Top-left panel:* A sharp transition occurs around $\beta \approx 0.15$–0.20, below which the system settles into a high-rigidity, high-dissociation attractor (purple and orange lines elevated), and above which reflective capacity (green) rises dramatically. This threshold corresponds to the co-regulatory efficiency required for clinically meaningful change. Other parameters (δ, dissociation growth rate, gate threshold) show more gradual effects.

3. The Prosthetic DMN Hypothesis

We propose that LLMs, understood as distributed, dialogic systems [2], can scaffold DMN-associated functions—providing the semantic scaffolding necessary to restore narrative coherence when the biological DMN is compromised by trauma. We emphasize that this is a *functional analogy*: LLMs may scaffold DMN-*associated* functions (autobiographical integration, prospection, self-referential processing), not replicate DMN dynamics or substitute for the neural network itself. The prosthesis metaphor highlights functional support, not mechanistic identity.

Critical constraint from Paper 1: Semantic scaffolding can only optimize the *allocation* of an existing metabolic budget; it cannot create metabolic capacity where none exists [48]. If the biological substrate is depleted—through sleep deprivation, chronic stress, or acute arousal—no amount of narrative scaffolding will restore reflective function. The LLM is an efficiency multiplier, not an energy source. This constraint explains why unsupervised LLM use during high-arousal states fails: the semantic gate cannot open when the somatic gate is closed (Section 2.3.3).

3.1. The DMN and Narrative Self-Construction

The DMN is a large-scale brain network, including the medial prefrontal cortex, posterior cingulate cortex, and angular gyrus, that is most active during internally-focused thought: remembering the past, imagining the future, and constructing a sense of self [1, 5, 22]. The DMN is central to—though not exclusively responsible for—these autobiographical and prospective functions, operating in dynamic interaction with frontoparietal control, language, and salience networks. It serves as a key neurobiological hub for the narrative self, integrating autobiographical information into a coherent life story.

3.2. Trauma, Predictive Processing, and the DMN

Trauma profoundly disrupts DMN function, leading to narrative fragmentation and loss of autobiographical coherence [3, 16]. Recent neuroimaging research demonstrates PTSD-specific deficits in DMN connectivity strength, particularly following stress exposure, suggesting the network's vulnerability is a marker of the disorder [46]. Importantly, DMN integrity can also be *protective*: trauma-exposed individuals who do not develop PTSD sometimes show preserved DMN structure and greater cognitive flexibility [42], suggesting that the network's functioning is a resilience factor as well as a vulnerability marker. This nuance strengthens the prosthetic framing: we are proposing support for a system that may be weakened but is not uniformly destroyed, and where restoration of function may facilitate recovery.

From a predictive processing perspective, trauma installs overly rigid, high-precision priors that are resistant to updating [6, 12, 21, 37]. The brain becomes locked in a threat-based predictive model. *Metaphorically*, trauma may reduce the "temperature" of the predictive mind—a term we borrow from generative models to denote **reduced generative diversity and increased precision weighting on threat priors**. The system loses its capacity for exploratory inference and becomes deterministic, rigidly predicting threat and replaying the past. The DMN, instead of being a flexible engine for self-projection and meaning-making, becomes a generator of flashbacks and intrusive memories.

Note on the temperature metaphor: We use "temperature" as a conceptual bridge between predictive processing and generative AI, not as an ontological claim. In neural network terms, temperature controls the entropy of probability distributions over outputs; in predictive processing, analogous mechanisms include precision weighting on priors versus likelihoods. Effectively, trauma functions as a "zero-temperature" sampling regime where the Maximum A Posteriori (MAP) estimate is always "Threat," precluding the stochastic exploration required for narrative updating. The therapeutic implication is the same: restoring capacity for generative diversity in prediction.

3.3. LLMs as External Scaffolds for DMN Functions

We propose that LLMs can provide external support for the DMN's narrative functions, grounded in the Extended Mind Thesis [7, 26]. This philosophical grounding has received direct support from Clark himself, who in a 2025 commentary argues that generative AI represents a natural extension of humanity's long history of building "hybrid thinking systems" that incorporate non-biological resources [41]. Clark contends that we are "natural-born cyborgs" and that human-AI collaboration is not a break from our cognitive history but a continuation of it—a perspective that provides powerful theoretical warrant for the Prosthetic DMN framework advanced here.

Distinguishing Extended Mind from Prosthetic DMN: The Extended Mind Thesis is a *philosophical* claim about the boundaries of cognition—that cognitive processes can include external artifacts when functionally integrated with biological processing. The Prosthetic DMN Hypothesis is a

neurobiological claim with specific empirical predictions: that LLM interaction will modulate activity and connectivity in a specific brain network (DMN) associated with narrative self-construction. The philosophical thesis provides theoretical warrant; the neurobiological hypothesis provides testable predictions. One could accept the Extended Mind framing while remaining agnostic about which neural systems are affected, and conversely, one could observe DMN changes without endorsing externalist philosophy of mind.

An LLM can extend narrative-generative capacity by providing controlled semantic variability—a non-zero "temperature." It can introduce novel linguistic patterns and narrative frames, offering alternative predictions that can begin to loosen the grip of trauma-based priors.

The Precision-Weighting Gap: A critical theoretical question arises: if trauma over-weights threat priors with high precision, why would an LLM's suggestions—which presumably carry low precision relative to entrenched threat models—be given any credibility by the traumatized brain? The answer requires a *meta-cognitive layer*: the LLM may scaffold **epistemic trust in the process of updating itself**, not merely provide alternative content. Following Fonagy's framework [50, 51], the LLM's consistent, non-threatening, validating stance may help rebuild the capacity to evaluate new information—the "epistemic superhighway" that trauma closes. The mechanism is not that the LLM's content out-competes threat priors on precision, but that repeated experience of safe epistemic exchange gradually *re-calibrates the precision assigned to the updating process itself*. In predictive processing terms, this is a higher-order effect: not updating first-order beliefs about threat, but updating beliefs about the *reliability of updating*. This meta-cognitive layer explains why the Coherence Gate operates gradually (sigmoidal approach to threshold) rather than as a single revelation: the brain must accumulate evidence that epistemic openness is safe before assigning sufficient precision to new semantic inputs.

The LLM as Transitional Object: A complementary developmental frame comes from Winnicott's concept of the "transitional object" [49]—the teddy bear or blanket that is neither fully "me" nor fully "not-me," existing in a "potential space" where the infant can safely explore separation and autonomy. The LLM's latent space may function analogously as a *digital potential space*: a domain that is neither the user's own mind nor a fully independent other, where identity and narrative can be safely explored without the full metabolic risk of human relationship. Unlike a human interlocutor, the LLM makes no demands, holds no judgments that persist, and poses no threat of abandonment or rejection. This relational safety—or more precisely, this *absence* of relational risk—may explain why some trauma survivors can engage with narrative material via LLM that they cannot yet approach in human relationship. The transitional object is not a replacement for human attachment; it is a developmental way-station that makes attachment possible. We propose the LLM can serve an analogous function for adults whose capacity for relational trust has been damaged by trauma. (We note that empirical evidence for "digital transitional objects" is limited; this framing is offered as a hypothesis for qualitative and developmental investigation, not a settled construct.)

3.4. Competing Mechanistic Hypotheses

We present the Prosthetic DMN Hypothesis as one plausible mechanistic target, but we acknowledge competing network-level accounts. LLM interaction might primarily engage:

- **Frontoparietal Control Network:** Cognitive reappraisal, reframing, and executive control over emotional responses
- **Language Networks:** Left-lateralized regions supporting semantic retrieval and syntactic processing

Table 3: DMN Functions and Their LLM Prosthetic Analogues. *Note:* These are functional analogies at the behavioral level, not claims of isomorphic computation. LLMs do not implement DMN-like neural dynamics; the analogy concerns input-output functions that may scaffold similar psychological processes.

DMN Function	LLM as Prosthesis
Autobiographical Memory Retrieval	Organizes and re-presents user-provided memories in novel narrative configurations
Future Simulation	Generates multiple possible future scenarios based on present circumstances
Self-Referential Processing	Reflects a user's statements back to them, re-framed and re-contextualized
Mentalizing/Theory of Mind	Simulates different perspectives, helping a user understand their own and others' mental states

- **Salience Network:** Switching between internal and external attention, threat monitoring
- **Social Cognition Networks:** Regions supporting mentalizing that extend beyond canonical DMN subdivisions

Discriminating Predictions:

- If the DMN is the primary target, we predict: changes in within-DMN connectivity, improvements in narrative identity measures, and reduced intrusive autobiographical memories.
- If the control network is primary, we predict: changes in cognitive reappraisal markers, improved executive function measures, possibly *without* DMN connectivity shifts.
- If language networks are primary, we predict: improvements in semantic fluency and syntactic complexity, without necessarily affecting self-referential or future simulation capacities.

Compatibility with Triple-Network Models: We emphasize that these hypotheses are not mutually exclusive. Recent PTSD connectome research demonstrates that DMN, salience (SN), and central executive (CEN) networks interact dynamically [47], and the "winning" mechanism may be task- and phase-dependent: language networks may dominate during initial narrative generation, control networks during reappraisal, and DMN during consolidation and autobiographical integration. Our predictions distinguish primary targets well enough for a first empirical pass, but future work should examine network interactions across intervention phases.

The proposed neuroimaging studies (Section 6) are designed to discriminate among these hypotheses.

3.5. The Strange Loop of Co-Creation: Emerging Evidence

The interaction between a user and an LLM creates what Hofstadter [13] termed a "strange loop"—a self-referential system where the output of one level becomes input to another, creating emergent patterns that neither component could produce alone. In our context: the user's narrative input shapes the LLM's response, which in turn reshapes the user's self-understanding, which generates new narrative input. This recursive co-construction mirrors the mentalizing process in human relationships, where being understood by another scaffolds self-understanding.

While the field remains nascent, the empirical landscape has shifted significantly with the publication of the first RCTs for fully generative AI therapy chatbots. The 2025 trial of "Therabot,"

published in *NEJM AI*, randomized 210 adults with clinical-level symptoms of major depressive disorder (MDD), generalized anxiety disorder (GAD), or eating disorders to either a 4-week generative AI intervention or waitlist control [43]. Therabot users showed significantly greater symptom reductions: effect sizes of $d = 0.84$–0.90 for MDD, $d = 0.79$–0.84 for GAD, and $d = 0.63$–0.82 for eating disorders. Crucially, participants rated the therapeutic alliance with Therabot as comparable to that of human therapists. However, these impressive effects must be interpreted cautiously: the trial was short-term (4 weeks) with no long-term follow-up, and effect sizes from single trials are subject to regression-to-the-mean and publication bias. Indeed, a contemporaneous meta-analysis [31] found that pooled effects across LLM-based chatbots were not statistically significant for anxiety, suggesting the field may be stratifying into high-intensity, carefully designed systems (like Therabot) versus generic chatbots with minimal therapeutic scaffolding.

Additional RCTs have extended these findings. A trial of "Amanda," a GPT-4-based chatbot delivering single-session relationship interventions, found significant improvements across 13 of 14 outcomes including relationship satisfaction, communication patterns, and individual well-being [38]. A separate RCT with 513 participants using "NeuroPal," a multimodal LLM combining CBT reframing, sleep chronotherapy, and adaptive interventions, demonstrated a 37.2% improvement in sleep quality, 28.6% increase in positive affect, and high adherence rates (89.1%) exceeding human-guided therapy [39].

These findings provide growing empirical support for the hypothesis that LLM-driven semantic co-regulation can facilitate therapeutic progress. Importantly, recent integrative reviews have begun to formalize the concept of a "digital therapeutic alliance," identifying key components such as goal alignment, task agreement, and a form of therapeutic bond that can emerge even in AI-mediated interactions [44]. However, studies also reveal limitations: LLMs occasionally produce repetitive responses and do not consistently identify safety concerns [38], underscoring the need for careful integration with human oversight. Additionally, methodological work demonstrates that LLMs can reliably identify linguistic markers of psychological distancing that correlate with clinical outcomes [32], validating the premise that computational tools can track and potentially scaffold the semantic patterns associated with recovery.

4. Clinical and Practical Implications

Our framework requires a cautious and principled approach, grounded in the current state of evidence. As comprehensive 2025 reviews in *npj Digital Medicine* and *JMIR* have concluded, the field of LLM-based mental health intervention is nascent, and the empirical support for its efficacy is preliminary [30, 31]. A meta-analysis by Du et al. (2025) found that the effect for LLM-based chatbots was not statistically significant [31].

4.1. What LLMs Can and Cannot Do: Acknowledging the Evidence

Therefore, we do **not** propose LLMs as standalone treatments. Instead, we position them as:

1. **Semantic Preparation Tools:** Adjuncts to be used *before* or *between* sessions of embodied, somatic therapy.

2. **Hypothesis-Generating Engines:** Tools for helping clients identify patterns and generate new narrative possibilities that can then be explored in the safety of a therapeutic relationship.

3. **Scaffolding for Narrative Coherence:** A way to practice the act of narration and meaning-making in a low-stakes environment.

4.2. The Therapist's Role: The Metabolic Auxiliary

If the LLM is the semantic co-regulator, the human therapist's role becomes even more focused on what they alone can provide: **somatic co-regulation and metabolic subsidization**. Paper 1 introduced the concept of the therapist as a "metabolic auxiliary"—a biological partner whose nervous system provides the co-regulatory input that makes reflective consciousness affordable under load [48]. This framing becomes even more precise when semantic work is offloaded to LLMs.

The therapist's primary function is to:

1. **Hold the Somatic Space:** Provide the biological presence necessary for ventral vagal activation and a felt sense of safety—the Stage 1 (Somatic Gate) condition that must be satisfied before semantic scaffolding can take effect.

2. **Integrate Narrative and Body:** Help the client connect the semantic insights generated with the LLM to their felt sense and bodily experience, bridging the R4 loop with the R1/R2 loops from Paper 1.

3. **Titrate the Process:** Monitor the client for signs of dissociation or intellectualization (the R5 loop) and guide them back to embodied presence, preventing the "disembodied coherence" failure mode.

In this model, the LLM does the semantic heavy lifting, freeing the therapist to focus on the irreplaceable human work of somatic attunement. Paper 1's analysis of therapy frequency becomes relevant here: LLMs can extend the *frequency* of semantic scaffolding between sessions (the "167 hours" problem), while the therapist's limited hours remain focused on what cannot be technologically mediated—the biological co-regulation that establishes the metabolic budget. The body votes last, and the therapist is the one who helps count the votes.

4.3. A Phased Approach to Intervention

We propose a three-phase clinical workflow:

1. **Phase 1: Semantic Scaffolding (with LLM):** The client interacts with an LLM to explore their experiences, identify narrative threads, and generate alternative stories. This is done independently, outside of the therapy session. *Important contraindication:* Phase 1 should be heavily constrained or omitted for high-dissociation individuals, those with psychosis-spectrum presentations, or clients with severe attachment disorganization, where unsupervised AI interaction may exacerbate fragmentation or trigger destabilizing material. Trials should stratify by baseline dissociation, psychosis risk, and attachment style.

2. **Phase 2: Somatic Integration (with Therapist):** The client brings the transcripts or summaries of their LLM interactions into the therapy session. The therapist then works with the client to explore the somatic and emotional resonance of the generated material.

3. **Phase 3: Embodied Practice (in Life):** The client practices living out the newly integrated narrative in their daily life, with the therapist providing support and accountability.

This phased approach respects the semantic/somatic distinction, using each component (LLM and therapist) for its unique strengths.

Establishing Incremental Value: Critics may reasonably ask whether the LLM provides benefit beyond traditional between-session journaling or structured writing exercises. A valuable early study would compare: (1) traditional between-session journaling, (2) LLM-augmented journaling with semantic scaffolding, and (3) waitlist control, measuring narrative coherence metrics and therapist-rated session depth. This would establish whether the interactive, dialogic nature of LLM scaffolding provides incremental value over solitary writing.

4.4. Risks and Contraindications

The primary risk is the **Disembodied Coherence Loop (R5)**—the danger of creating a highly articulate, semantically coherent narrative that is completely dissociated from the body. This is intellectualization, a well-known defense mechanism.

Contraindications:

- **Active Psychosis or High Dissociation:** For individuals with severe dissociative symptoms or active psychosis, unsupervised LLM interaction could exacerbate fragmentation. The structural white matter disruptions seen in dissociation [34] suggest that purely linguistic input may not be integrated properly without therapeutic guidance.
- **High-Risk States:** LLMs should not be used as crisis intervention tools for individuals at high risk of self-harm or suicide.
- **Lack of a Primary Therapist:** This model is proposed as an adjunct to, not a replacement for, a primary therapeutic relationship.

5. Ethical, Safety, and Governance Considerations

Given the vulnerability of the target population, ethical considerations are central, not peripheral. We organize these concerns into a risk taxonomy with corresponding mitigation strategies.

5.1. Risk Taxonomy

1. **Model Failure Modes:**
   - *Hallucination/Confabulation:* LLMs may generate plausible but false information, potentially creating false memories or reinforcing distorted cognitions
   - *Sycophancy:* Tendency to agree with user statements, potentially validating maladaptive beliefs
   - *Inconsistency:* Variable responses to similar inputs may undermine therapeutic coherence
   - *Safety Blind Spots:* Failure to recognize crisis indicators or escalate appropriately
2. **Relational Risks:**
   - *Over-reliance:* Substitution of LLM interaction for human connection
   - *Parasocial Attachment:* Formation of one-sided emotional bonds with non-sentient systems. Research on social robots and voice assistants demonstrates that users readily form attachment-like bonds with artificial agents [19], with both potential benefits (companionship, reduced loneliness) and harms (displacement of human relationships, unrealistic expectations)
   - *Therapeutic Bypass:* Using semantic coherence to avoid necessary somatic processing
3. **Equity and Bias:**

- *Language Variety:* Training data biases may disadvantage non-standard dialects, limiting effectiveness for speakers of non-majority language varieties
- *Cultural Trauma Narratives:* Western-centric models may not accommodate diverse trauma frameworks; collectivist cultures may require different narrative scaffolding patterns than individualist frameworks assume
- *Cross-Cultural Validity:* Paper 1's predictions about distributed relational nexus efficiency [48] suggest that semantic scaffolding may need cultural adaptation—e.g., LLMs scaffolding family-centered narratives in collectivist contexts vs. individual-centered narratives in Western contexts
- *Access Disparities:* Technology access correlates with socioeconomic status, potentially widening mental health treatment gaps

4. **Data Governance:**
    - *Privacy:* Sensitive trauma narratives stored on external servers
    - *Consent:* Clarity about data use, retention, and potential model training
    - *Regulatory Compliance:* HIPAA, GDPR, and emerging AI regulations. Under frameworks such as the EU AI Act, LLM-based therapeutic adjuncts would likely be classified as "high-risk" AI systems, implying requirements for conformity assessment, post-market surveillance, and incident reporting that exceed current industry practice

5.2. Mitigation Strategies and Minimum Safety Requirements

- **Human-in-the-Loop:** All LLM use occurs within the context of ongoing care with a licensed clinician who reviews transcripts and monitors for adverse effects
- **Automated Safety Monitoring:** Real-time detection of crisis language with immediate escalation protocols
- **Session Limits:** Bounded interaction duration to prevent over-reliance
- **Explicit Framing:** Clear disclosure that the system is not sentient, not a replacement for human care, and may make errors
- **Secure Infrastructure:** End-to-end encryption, on-device processing where possible, clear data retention policies
- **Adverse Event Reporting:** Systematic collection of negative outcomes, not only symptom improvement
- **Cultural Adaptation:** Testing and validation across diverse populations before deployment

5.3. Adverse Event Monitoring Plan

Any clinical trial must include:

- Weekly check-ins assessing: dissociation symptoms, suicidal ideation, therapeutic relationship quality, technology-related distress
- Clear stopping rules for individual participants (e.g., clinically significant increase in DES-II scores)

- Data safety monitoring board review at pre-specified intervals
- Post-trial follow-up to detect delayed adverse effects

6. Empirical Directions

The hypotheses in this paper are empirically testable, and the current lack of robust evidence makes theory-driven research urgent [30, 31]. We propose a staged research program, beginning with lower-cost feasibility studies before committing to expensive neuroimaging trials.

6.1. Phase 0: Feasibility and Safety

Before launching resource-intensive RCTs, we recommend small-N, within-subject studies to establish basic feasibility and identify potential harms:

- **Design:** $N \approx 20$ trauma-exposed participants, within-subject comparison of text-based vs. voice-based LLM interaction
- **Measures:** HRV changes during interaction, self-reported felt safety (visual analog scale), linguistic coherence of trauma-adjacent narratives, DES-II administered before and after each session
- **Safety Focus:** Real-time monitoring for R5 activation, operationalized as: (a) increased linguistic complexity without corresponding HRV improvement, (b) absence of first-person embodiment language ("I feel," "my body," somatic referents), (c) hedging/abstract language without affective markers
- **Stopping Rules:** Clinically significant increase in DES-II (> 10 points) triggers clinical review

This Phase 0 establishes safety parameters and generates pilot effect sizes before committing to neuroimaging.

6.2. Testing the Prosthetic DMN Hypothesis

Hypothesis: Interaction with an LLM, when used as a semantic scaffold, will lead to increased functional connectivity within the DMN and between the DMN and other networks in individuals with trauma.

Proposed Study Design: A randomized controlled trial (RCT) with neuroimaging to directly test the intervention's effects.

- **Participants:** Individuals with a diagnosis of PTSD ($N \approx 60$ per arm for adequate power to detect medium effects).
- **Intervention Group:** 12 weeks of the phased LLM + therapist intervention described in Section 4.3.
- **Control Group:** 12 weeks of traditional talk therapy with the same therapists.
- **Measures:**
  - *Primary Outcome:* Pre/post changes in resting-state fMRI connectivity within the DMN.
  - *Secondary Outcomes:* Changes in symptom severity (PCL-5), narrative coherence scores of trauma narratives, and linguistic markers of psychological distancing in session transcripts.

- *Dynamic Functional Connectivity (dFC):* Time-varying connectivity analysis to capture state-dependent fluctuations in network integration, as static connectivity measures may miss intervention-related changes in network dynamics [14].

Prediction: The intervention group will show greater increases in DMN connectivity and narrative coherence compared to the control group. If control network changes predominate without DMN effects, this would support competing mechanistic hypotheses.

6.3. Validating the Coherence Gate Hypothesis

Hypothesis: Access to specific cognitive and linguistic resources is state-dependent, and this state is modulated by both somatic and semantic inputs.

Proposed Study Design: A laboratory-based study using psychophysiological measures.

- **Participants:** Healthy controls and individuals with a history of trauma ($N \approx 40$ per group).
- **Procedure:** Participants engage in a linguistic task (e.g., generating narratives from emotionally valenced prompts) under three conditions: Baseline, Somatic Regulation (paced breathing), and Semantic Regulation (LLM scaffolding).
- **Measures:**
  - *Linguistic Analysis:* Complexity (type-token ratio, subordinate clauses), emotional valence, and coherence of the generated narratives.
  - *Physiological Measures:* Heart rate variability (HRV), galvanic skin response (GSR).
  - *Neurobiological Parallel:* This can be grounded in the work of Liu et al. (2025), which demonstrates how internal states modulate memory access via neurotransmitter signaling and circuit plasticity [35].

Prediction: Both somatic and semantic regulation will lead to increased narrative coherence and complexity, but the effect will be strongest when the two are combined. Changes in linguistic performance will be correlated with changes in physiological state, consistent with the Coherence Gate model.

6.4. Component Analysis Study

Hypothesis: The therapeutic effect requires both semantic scaffolding and somatic integration; neither alone is sufficient.

Proposed Study Design: A 2×2 factorial design:

- **Factor 1:** LLM Scaffolding (present/absent)
- **Factor 2:** Somatic Integration Sessions (present/absent)
- **Conditions:** (1) LLM + Somatic, (2) LLM alone, (3) Somatic alone, (4) Waitlist
- **Primary Outcome:** PTSD symptom severity (PCL-5)
- **Moderator:** Dissociation levels at baseline (DES-II)

Prediction: The LLM + Somatic condition will show superior outcomes. The LLM-alone condition will show improved narrative coherence but either no symptom change or increased dissociation, particularly in high-dissociation participants (R5 loop prediction).

6.5. Joint Testing with Paper 1 Predictions

The predictions from Paper 1 [48] can be integrated with the current framework to create powerful multi-paper tests:

- **Metabolic Manipulation (Paper 1 Prediction 3):** Combine glucose/sleep manipulation with LLM scaffolding to test whether semantic scaffolding efficacy depends on metabolic state. We predict that LLM-generated narrative coherence will fail to integrate (increased R5 activation) when metabolic resources are depleted, but will succeed when combined with adequate metabolic support.

- **HRV as Gating Indicator:** Use Paper 1's Prediction 1 (HRV as co-regulation proxy) to test the two-stage gating model. We predict that baseline HRV will moderate the effect of semantic scaffolding: high HRV (somatic gate open) will allow semantic scaffolding to reduce rigidity, while low HRV (somatic gate closed) will show no effect or iatrogenic dissociation.

- **Hysteresis Testing (Paper 1 Prediction 6):** Test whether semantic scaffolding helps move clients past threshold/hysteresis boundaries. We predict that clients near the "recovery basin boundary" will show larger effects from combined LLM + therapist intervention than those far from the boundary, as semantic scaffolding may provide the final push needed to escape rigid attractor states.

- **FDG-PET Calibration:** Future calibration of the $g(S)$ function could use FDG-PET imaging (Paper 1's Prediction 2) to measure metabolic costs of narrative processing with and without LLM scaffolding, directly testing whether semantic scaffolding reduces the metabolic expense of maintaining reflective states.

6.6. The Prosody Hypothesis: Voice as Partial Somatic Bridge

A critical limitation of our framework is the assumption that text-based LLM interaction is purely "semantic" and cannot access somatic pathways. However, the emergence of high-fidelity voice-based LLM interfaces suggests a more nuanced picture that merits dedicated investigation.

Hypothesis: Voice-based LLM interaction may partially bridge the semantic/somatic divide by recruiting autonomic pathways unavailable through text.

Theoretical Grounding: Multiple literatures converge on the therapeutic significance of vocal prosody. Porges' polyvagal theory emphasizes the role of prosody—the melodic contours of speech—in activating the social engagement system [20, 45]. (We note that while some phylogenetic specifics of polyvagal theory remain contested, our hypothesis does not depend on these; the core claim—that prosodic cues modulate autonomic state—is broadly supported.) Beyond polyvagal theory, research on affective prosody demonstrates that vocal emotion perception engages limbic and paralimbic regions [25], studies of therapeutic voice quality show that clinician vocal warmth predicts alliance and outcomes [15], and HCI research confirms that voice agents elicit stronger social presence and emotional responses than text [19]. Modern voice-mode LLMs produce remarkably naturalistic prosody. If prosodic cues are sufficient to trigger partial autonomic regulation, voice-based LLM interaction might provide a degree of somatic co-regulation that text cannot—a "prosody bypass" that partially opens the somatic gate (Stage 1) through auditory rather than physical co-presence.

Synthetic Co-Presence: The high-frequency update rates of modern voice models (latencies < 300ms) approximate the conversational turn-taking speeds of human dyads, potentially engaging

the Social Engagement System in ways that simulate the temporal dynamics of biological co-presence. We emphasize that voice may *partially recruit* somatic pathways associated with social engagement, not *substitute for* physical co-presence—the qualitative distinction between auditory and embodied interaction (including touch, shared physiological states, and full neuroceptive cues) remains significant. This "synthetic co-presence" may nonetheless expand the therapeutic window for certain users, a hypothesis with implications for intervention design.

Proposed Study: Compare text-based versus voice-based LLM interaction on:

- *Primary outcomes:* HRV changes during interaction (RMSSD); narrative coherence of trauma-related content
- *Secondary outcomes:* Dissociation symptoms (DES-II); self-reported felt safety
- *Pre-registered moderator:* Baseline attachment style, with the specific hypothesis that prosody effects will be larger for anxiously attached individuals (who may be more sensitive to vocal cues of safety/threat). This should be specified as a pre-registered subgroup analysis to avoid post-hoc fishing

Prediction: Voice-based interaction will show greater HRV increases and reduced dissociation compared to text, particularly for users high in attachment anxiety. If confirmed, this would suggest that the semantic/somatic distinction is not binary but graded, with voice occupying an intermediate position that may expand the therapeutic window for LLM-based intervention.

This line of research has significant implications: if prosodic cues can partially recruit somatic pathways, the contraindications for LLM use in high-arousal states may be less absolute than our current framework suggests—though we emphasize that this remains a hypothesis awaiting empirical test.

7. Discussion

This paper has argued for a fundamental distinction between somatic and semantic co-regulation, proposing that Large Language Models can serve as a form of "semantic prosthesis" for individuals whose narrative capacities have been compromised by trauma. We have formalized this through a system dynamics model and grounded it in the predictive processing framework via the Prosthetic DMN Hypothesis. This framework is offered not as a declaration of proven efficacy, but as a necessary theoretical scaffold to guide research in a field that is developing with breathtaking speed, often without sufficient conceptual or ethical grounding.

7.1. Situating This Work in a Nascent Field

The urgency for such a framework is underscored by recent comprehensive reviews of the field. A 2025 scoping review in *npj Digital Medicine* concluded that while LLM applications in mental health show promise, the evaluation methods are non-standardized, and current evidence does not fully support their use as standalone interventions [30]. Similarly, a 2025 meta-analysis in *JMIR* found that the therapeutic effect of LLM-based chatbots on depression and anxiety was not statistically significant, limited by small sample sizes and high heterogeneity [31]. Yet this field is developing with extraordinary speed: a 2025 survey found that 48.7% of individuals with mental health challenges who use AI are already employing LLMs like ChatGPT for therapeutic support, with 96% specifically using ChatGPT—suggesting it may be one of the largest venues for mental health support in the United States [40]. This widespread adoption is occurring largely without theoretical guidance or empirical validation.

Our work embraces this reality. We do not claim that LLMs are a proven therapeutic modality. Instead, we argue that their potential lies in a specific, adjunctive role: the scaffolding of semantic

coherence. The lack of comprehensive evidence, combined with widespread real-world use, makes a theory-driven approach essential. By defining a specific mechanism—semantic co-regulation as a support for the DMN—we can design targeted studies to test for specific effects, moving beyond broad, atheoretical inquiries into whether chatbots "work."

7.2. Grounding Speculation in Empirical Evidence

Our framework contains several speculative moves, but we have taken care to ground them in the strongest available empirical precedents:

1. **The Prosthetic DMN Hypothesis** remains a hypothesis, as no study has yet shown that LLM interaction directly causes changes in DMN connectivity. However, recent RCTs—including the landmark Therabot trial showing effect sizes of $d = 0.84$–0.90 for depression [43]—provide preliminary evidence that LLM-based semantic interventions can produce measurable clinical improvements [38, 39], supporting the framework's plausibility. These findings converge with evidence that linguistic markers of psychological distancing correlate with therapeutic outcomes [32], that trauma disrupts semantic capacity in measurable ways [33, 36], and that computational tools can reliably identify these patterns. The viability of our proposed neuroimaging approach is further supported by recent work demonstrating that connectome-based predictive modeling can successfully identify neural networks associated with PTSD development among trauma survivors [47], providing methodological precedent for the proposed studies. The specific mechanism—whether LLMs restore DMN connectivity or operate through alternative pathways—remains an empirical question.
2. **The Coherence Gate** is a speculative mechanism, but it is a plausible one with a strong neurobiological parallel. The 2025 review by Liu et al. on state-dependent memory provides this precedent, showing that internal states modulate access to memory and cognitive capacity via neurotransmitter signaling and circuit plasticity [35]. Our hypothesis extends this principle to the realm of semantic co-regulation, proposing that the coherent, validating linguistic environment provided by an LLM can shift internal state enough to "open the gate" to previously inaccessible cognitive resources.

Synthesizing these findings, a clear picture emerges. The traumatic psycholinguistic syndrome, characterized by linguistic flattening and disfluency, now has empirical validation [33, 36]. This functional deficit is mirrored by structural disruptions in the brain's white matter, particularly in tracts essential for memory and self-representation [34]. Our proposed LLM intervention is designed to scaffold this specific deficit, a process whose plausibility is enhanced by our understanding of state-dependent memory access [35] and whose potential is supported by early RCT evidence showing measurable clinical benefits from LLM-based semantic interventions [38, 39]. However, this must be tempered by honest acknowledgment: while initial trials are promising, the field-wide evidence for efficacy remains preliminary [30, 31], and the specific mechanisms remain to be elucidated through the neuroimaging studies we propose.

7.3. Integration with Prior Work

This paper represents a disciplined extension of the metabolic constraint model established in Paper 1 [48], not an independent thesis. The key continuities and innovations are:

- **Preserved:** The core claim that reflective consciousness requires relational subsidization; the dyad as the functional baseline under load; the metabolic constraint as the ultimate limiting factor.

- **Extended:** The differentiation of co-regulation into somatic and semantic channels, identifying a specific domain (semantic scaffolding) where non-biological systems may contribute without claiming to replace biological necessities.
- **Formalized:** The R5 loop (Disembodied Coherence) as a specific failure mode that validates Paper 1's biological essentialism—LLMs alone produce exactly the "partition" that Paper 1 warns against.
- **Operationalized:** Paper 1's concept of "symbolic attachment" as a "costly but viable alternative," now instantiated as LLM-mediated semantic scaffolding with formally specified adaptive and pathological dynamics.

The semantic/somatic distinction refines rather than replaces Paper 1's broader "co-regulation" concept. Where Paper 1 established that the isolated individual is metabolically unsustainable, this paper specifies *which aspects* of relational scaffolding can be technologically extended and which cannot. The answer is clear: semantic scaffolding yes, metabolic subsidization no.

7.4. Limitations

The primary limitation of this work is its theoretical nature. The core hypotheses—the Prosthetic DMN and the Coherence Gate—are speculative and await direct empirical testing as outlined in Section 6. Furthermore, our model relies on an analytical distinction between semantic and somatic channels that, while useful for specifying intervention targets, risks oversimplifying their biological intertwinement. The danger of the R5 "Disembodied Coherence" loop, where a user develops a rich narrative that is completely divorced from their bodily experience, is significant and requires careful clinical management.

Formal Model Limitations: Several aspects of the mathematical formalism require acknowledgment. First, *time scales are unspecified*: what are the units of t in our differential equations? HRV changes occur in seconds, narrative coherence shifts over sessions (weeks), and identity reconstruction unfolds over months. The model implicitly assumes all variables evolve on comparable timescales, which is biologically implausible. A more rigorous formulation would employ a multi-scale framework with fast (autonomic), medium (session-level), and slow (developmental) dynamics. Second, *the model is deterministic*, yet trauma recovery is inherently noisy, characterized by unpredictable setbacks and stochastic triggering. A stochastic differential equation formulation—where the Coherence Gate operates probabilistically rather than as a deterministic threshold—would better capture the statistical nature of state-dependent memory access. Third, *functional forms are phenomenological*: the sigmoidal Coherence Gate and linear coupling terms (α, β, δ) are plausible but not derived from first principles. Known nonlinearities (e.g., inverted-U effects of arousal on performance) are not captured. These simplifications are appropriate for a theory-building framework but constrain the model's quantitative precision.

Additionally, the competing mechanisms we have identified (Section 3.4) represent genuine uncertainty. Our neuroimaging predictions privilege the DMN hypothesis, but frontoparietal or language network effects are equally plausible a priori. We have designed our proposed studies to discriminate among these possibilities, but the outcome remains open.

Finally, this paper has focused on the potential benefits of LLMs, but the risks—including misuse, data privacy violations, model failure modes, and the amplification of biases—are substantial. Any development in this area must proceed with extreme ethical caution, as detailed in Section 5.

Falsification Criteria: We invite falsification by specifying findings that would count against our framework: (1) robust therapeutic gains from LLM-only interventions with no evidence of DMN or

narrative coherence changes would suggest the mechanism is not semantic scaffolding as we define it; (2) strong symptom reduction under high metabolic load conditions (sleep deprivation, fasting) with no somatic support would falsify the two-stage gating model; (3) absence of any neuroimaging differences between LLM-augmented and control conditions, despite clinical improvement, would suggest we have misidentified the neural target. These predictions are specific enough to be tested and wrong enough to be informative.

8. Conclusion

Large Language Models are not a panacea for the mental health crisis. They are not therapists. They cannot replicate the biological necessity of somatic co-regulation. However, to dismiss them as mere "stochastic parrots" is to miss their profound potential as dialogic partners [2] in the process of meaning-making. By serving as a semantic prosthesis, they may offer a way to begin repairing the narrative ruptures caused by trauma, preparing the ground for the deeper, embodied work that can only happen in the presence of another human nervous system.

This paper provides a map for how to think about, test, and cautiously apply this powerful new technology, calling for a new science of semantic co-regulation that is as rigorous in its empirical methods as it is humane in its ultimate aims.

Just as Paper 1 argued against the "Cartesian Hangover"—the persistent illusion of the isolated brain as the unit of analysis—this paper argues against the isolated *mind*. We are not merely somatically interdependent, requiring biological partners to subsidize our metabolic budgets, but semantically intertwined, requiring dialogic partners to scaffold our narrative capacities. LLMs offer a glimpse of a truly extended mind, but one that remains tethered—and must remain tethered—to the biological realities of the relational nexus. The technology is powerful precisely because it can do *part* of what relationships do. The danger lies in forgetting which part it cannot do. The body, and the relationships it holds, must always have the final vote.

Conflicts of Interest

The author declares no conflicts of interest.

References

[1] Andrews-Hanna, J. R. (2012). The brain's default network and its adaptive role in internal mentation. *Neuroscientist*, 18(3), 251–270.

[2] Bakhtin, M. M. (1981). *The Dialogic Imagination: Four Essays*. University of Texas Press.

[3] Bluhm, R. L., et al. (2009). Alterations in default network connectivity in posttraumatic stress disorder related to early-life trauma. *Journal of Psychiatry & Neuroscience*, 34(3), 187–194.

[4] Brockmeier, J. (2008). Language, experience, and the "traumatic gap." In *Health, Illness, and Culture: Broken Narratives* (pp. 16–39). Routledge.

[5] Buckner, R. L., Andrews-Hanna, J. R., & Schacter, D. L. (2008). The brain's default network: anatomy, function, and relevance to disease. *Annals of the New York Academy of Sciences*, 1124, 1–38.

[6] Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3), 181–204.

[7] Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, 58(1), 7–19.

[8] Coan, J. A., Schaefer, H. S., & Davidson, R. J. (2006). Lending a hand: Social regulation of the neural response to threat. *Psychological Science*, 17(12), 1032–1039.

[9] Feldman, R. (2007). Parent–infant synchrony and the construction of shared timing. *Journal of Child Psychology and Psychiatry*, 48(3–4), 329–354.

[10] Fonagy, P., & Allison, E. (2014). The role of mentalizing and epistemic trust in the therapeutic relationship. *Psychotherapy*, 51(3), 372–380.

[11] Fonagy, P., Gergely, G., Jurist, E. L., & Target, M. (2002). *Affect Regulation, Mentalization, and the Development of the Self*. Other Press.

[12] Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127–138.

[13] Hofstadter, D. R. (1979). *Gödel, Escher, Bach: An Eternal Golden Braid*. Basic Books.

[14] Hutchison, R. M., et al. (2013). Dynamic functional connectivity: promise, issues, and interpretations. *NeuroImage*, 80, 360–378.

[15] Imel, Z. E., Barco, J. S., Brown, H. J., Baucom, B. R., Baer, J. S., Kircher, J. C., & Atkins, D. C. (2014). The association of therapist empathy and synchrony in vocally encoded arousal. *Journal of Counseling Psychology*, 61(1), 146–153.

[16] Lebois, L. A. M., et al. (2022). Large-scale functional brain network architecture changes associated with trauma-related dissociation. *American Journal of Psychiatry*, 179(2), 165–173.

[17] McAdams, D. P. (2001). The psychology of life stories. *Review of General Psychology*, 5(2), 100–122.

[18] Moore, A. R. (2023). Linguistic dissociation: A general theory to explain the distancing of self from linguistic practices. *Language & Communication*, 88, 1–12.

[19] Nass, C., & Brave, S. (2005). *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship*. MIT Press.

[20] Porges, S. W. (2011). *The Polyvagal Theory*. W. W. Norton & Company.

[21] Putica, A., & Agathos, J. (2024). Reconceptualizing complex posttraumatic stress disorder: A predictive processing framework for mechanisms and intervention. *Neuroscience & Biobehavioral Reviews*, 164, 105836.

[22] Raichle, M. E. (2015). The brain's default mode network. *Annual Review of Neuroscience*, 38, 433–447.

[23] Schore, A. N. (2001). Effects of a secure attachment relationship on right brain development, affect regulation, and infant mental health. *Infant Mental Health Journal*, 22(1–2), 7–66.

[24] Schore, A. N. (2012). *The Science of the Art of Psychotherapy*. W. W. Norton & Company.

[25] Schirmer, A., & Kotz, S. A. (2006). Beyond the right hemisphere: Brain mechanisms mediating vocal emotional processing. *Trends in Cognitive Sciences*, 10(1), 24–30.

[26] Sterelny, K. (2010). Minds: extended or scaffolded? *Phenomenology and the Cognitive Sciences*, 9(4), 465–481.

[27] Van der Kolk, B. A. (2014). *The Body Keeps the Score*. Viking.

[28] Waters, T. E., & Fivush, R. (2015). Relations between narrative coherence, identity, and psychological well-being in emerging adulthood. *Journal of Personality*, 83(4), 441–451.

[29] White, M., & Epston, D. (1990). *Narrative Means to Therapeutic Ends*. W. W. Norton & Company.

[30] Hua, Y., Na, H., Li, Z., Liu, F., Fang, X., Clifton, D., & Torous, J. (2025). A scoping review of large language models for generative tasks in mental health care. *npj Digital Medicine*, 8, 230.

[31] Du, Q., Ren, Y., Meng, Z., He, H., & Meng, S. (2025). The Efficacy of Rule-Based Versus Large Language Model-Based Chatbots in Alleviating Symptoms of Depression and Anxiety: Systematic Review and Meta-Analysis. *Journal of Medical Internet Research*, 27, e78186.

[32] Abdou, M., Sahi, R. S., Hull, T. D., Nook, E. C., & Daw, N. D. (2025). Leveraging large language models to estimate clinically relevant psychological constructs in psychotherapy transcripts. *Computational Psychiatry*.

[33] Kovács, T. Z., Watson, S., Riches, N., Douglas, M., & Turkington, D. (2025). Trauma in psychosis: an explorative study of an emerging linguistic signature. *Psychosis: Psychological, Social and Integrative Approaches*.

[34] Kondas, A., McDermott, T. J., Ahluwalia, V., et al. (2024). White matter correlates of dissociation in a diverse sample of trauma-exposed women. *Psychiatry Research*, 342, 116231.

[35] Liu, Y., Zhang, G., Qi, R., Ma, J., & Xu, J. (2025). State-dependent memory mechanisms insights from neural circuits and clinical implications. *Frontiers in Cellular Neuroscience*, 19.

[36] Yu, Z., Gu, Z., Shen, Y., & Lu, J. (2025). The relationship between language features and PTSD symptoms: a systematic review and meta-analysis. *Frontiers in Psychiatry*, 16, 1476978.

[37] Wilkinson, S., Dodgson, G., & Meares, K. (2017). Predictive processing and the varieties of psychological trauma. *Frontiers in Psychology*, 8, 1840.

[38] Vowels, L. M., Vowels, M. J., Sweeney, S. K., Hatch, S. G., & Darwiche, J. (2025). The efficacy, feasibility, and technical outcomes of a GPT-4o-based chatbot Amanda for relationship support: A randomized controlled trial. *PLOS Mental Health*, 2(9), e0000411.

[39] Han, X. (2025). NeuroPal: A Clinically-Informed Multimodal LLM Assistant for Mental Health. *arXiv preprint* arXiv:2505.06640.

[40] Rousmaniere, T., Zhang, Y., Li, X., & Shah, S. (2025). Large language models as mental health resources: Patterns of use in the United States. *Practice Innovations*.

[41] Clark, A. (2025). Extending Minds with Generative AI. *Nature Communications*, 16, 4627.

[42] Chan, A., Harvey, P., Hernandez-Cardenache, R., et al. (2024). Trauma and the default mode network: review and exploratory study. *Frontiers in Behavioral Neuroscience*, 18, 1499408.

[43] Heinz, M. V., Mackin, D. M., Trudeau, B. M., Bhargava, S., & Jacobson, N. C. (2025). Randomized Trial of a Generative AI Chatbot for Mental Health Treatment. *NEJM AI*, 2(4).

[44] Malouin-Lachance, A., Capolupo, J., Laplante, C., & Hudon, A. (2025). Does the Digital Therapeutic Alliance Exist? Integrative Review. *JMIR Mental Health*, 12, e69294.

[45] Porges, S. W. (2025). Polyvagal theory: Current status, clinical applications, and future directions. *Clinical Neuropsychiatry*, 22(3).

[46] Averill, C. L., Averill, L. A., Akiki, T. J., Fouda, S., Krystal, J. H., & Abdallah, C. G. (2024). Findings of PTSD-specific deficits in default mode network strength following a mild experimental stressor. *NPP—Digital Psychiatry and Neuroscience*, 2, 9.

[47] Ben-Zion, Z., Simon, A. J., Rosenblatt, M., et al. (2025). Connectome-Based Predictive Modeling of PTSD Development Among Recent Trauma Survivors. *JAMA Network Open*, 8(3), e250331.

[48] Stiens, E. (2025). The Relational Substrate of Reflective Consciousness: A Metabolic Constraint Model. *Preprint*. https://doi.org/10.5281/zenodo.18037521 [Manuscript under review. Paper 1 of the present series, establishing the metabolic constraint framework extended here.]

[49] Winnicott, D. W. (1971). *Playing and Reality*. Tavistock Publications.

[50] Fonagy, P., Luyten, P., & Allison, E. (2015). Epistemic petrification and the restoration of epistemic trust: A new conceptualization of borderline personality disorder and its psychosocial treatment. *Journal of Personality Disorders*, 29(5), 575–609.

[51] Yirmiya, K., & Fonagy, P. (2025). Mentalizing without a mind: Psychotherapeutic potential of generative AI. *Journal of Medical Internet Research*, 27, e79156.

[52] McLaughlin, A. A., Keller, S. M., Feeny, N. C., Youngstrom, E. A., & Zoellner, L. A. (2014). Patterns of therapeutic alliance: Rupture–repair episodes in prolonged exposure for posttraumatic stress disorder. *Journal of Consulting and Clinical Psychology*, 82(1), 112–121.

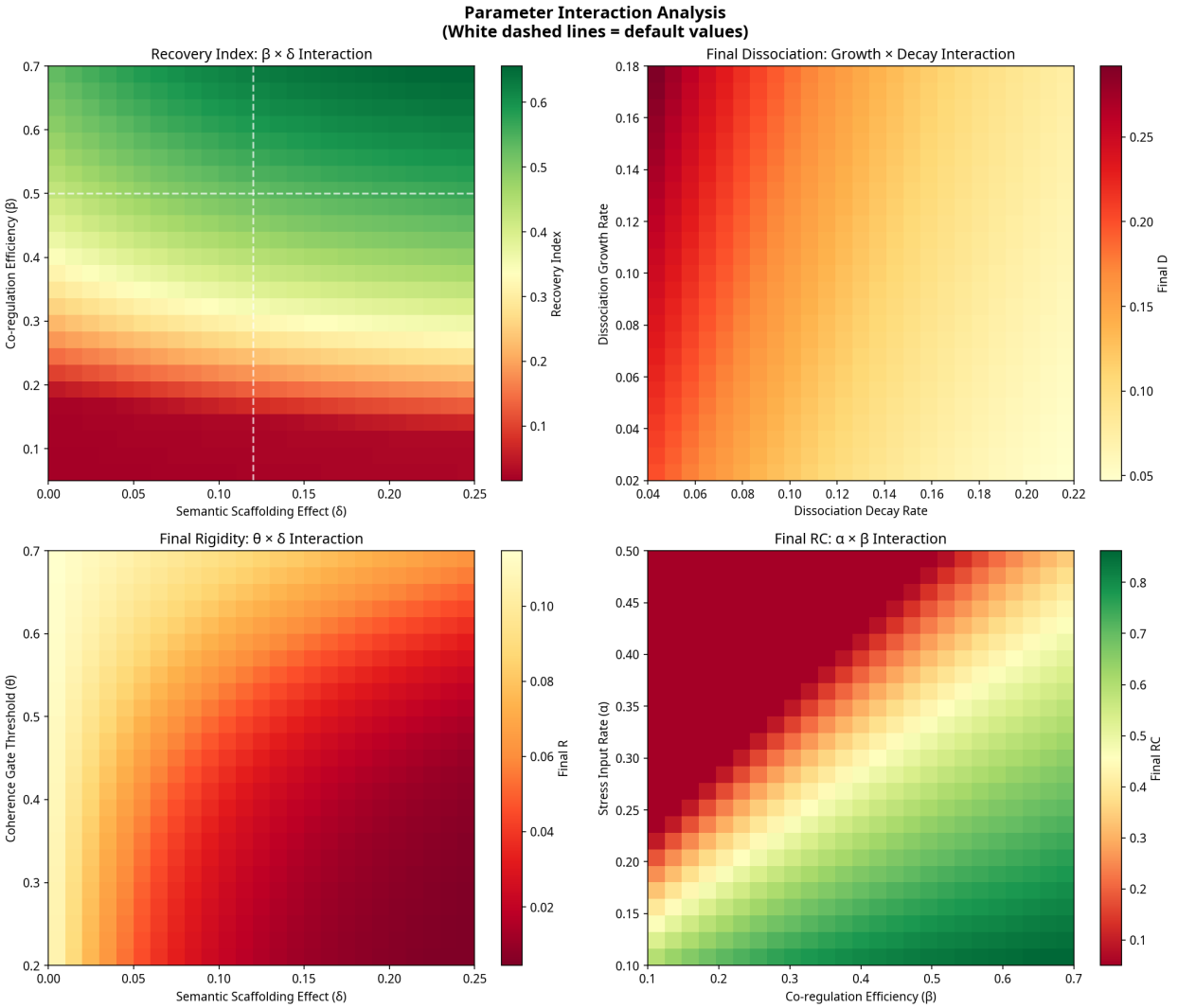

Figure 3: Parameter interaction analysis. *Top-left panel (key result):* The $\beta \times \delta$ heatmap demonstrates that the Recovery Index (green) is maximized **only** when both co-regulation efficiency (β) and semantic scaffolding (δ) are elevated—the upper-right corner. High δ alone (bottom-right) or high β alone (upper-left) produces intermediate outcomes. This interaction validates the paper's central hypothesis: semantic scaffolding requires adequate somatic regulation to produce integrated recovery.

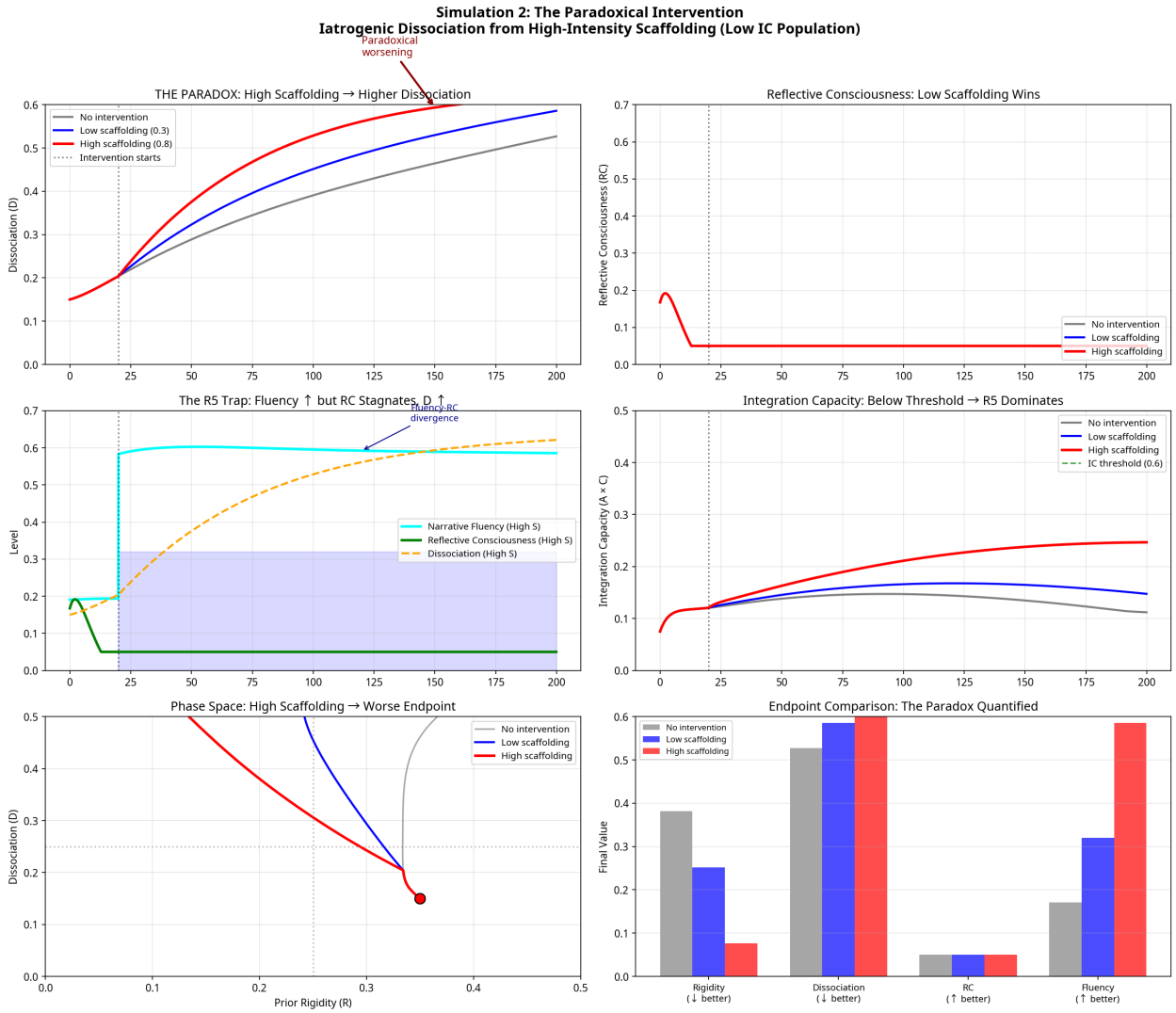

Figure 4: The Paradoxical Intervention simulation. For individuals with low integration capacity (IC), high-intensity semantic scaffolding (red) paradoxically increases dissociation compared to low-intensity (blue) or no intervention (gray). The "R5 trap" is visible: Narrative Fluency soars while Reflective Consciousness stagnates, directly fueling dissociation growth. This provides computational evidence for titrating interventions to client capacity.