Consciousness as Constrained Communication: Computer Study

Aman Chawla, Arun Raveendran and Salvatore Domenic Morgera
REAL Institute, Gurugram, India
University of South Florida, Tampa, FL 33620, USA

2026, May 06

Abstract

We study the application of Baar's Global Workspace Theory (GWT) to large language model (LLM)-based systems. Specifically, we are interested in the joint behavior of two metrics - system coherence and system diversity. LLM-based systems are chosen since they mimic the human mind. We construct an architecture in which multiple specialist modules compete for access to a shared workspace. By introducing a disruptor (noise) module at varying strengths, we empirically characterize the system's achievable (coherence, diversity)-pairs and find a concave hull structure analogous to multiple access channel (MAC) capacity regions studied in multi-user information theory. This characteristic structure is proposed as a candidate signature of integrative cognition.

Key Insight:

Consciousness may be amenable to treatment as a capacity constrained communication problem - we provide a proof-of-concept. The developed coherence-diversity rate region forms the basis for intelligent engineering of robust AI systems capable of functioning well in a varied and non-homogeneous set of conditions.

1 Introduction

The study of consciousness in computational systems has long been motivated by two central questions: *why* should artificial systems aspire to emulate conscious-like processing, and *so what* does such emulation achieve in practice? [1] Addressing these questions requires moving beyond descriptive analogies and toward operational frameworks that can be measured, tested, and compared across architectures. Consciousness, in biological systems, is not merely a philosophical curiosity; it is a functional mechanism that enables integration of diverse information streams, adaptive decision-making under uncertainty, and resilience to disruption. If artificial systems are to approach human-level cognition, they must grapple with these same constraints. Thus, the "so-what" of machine consciousness lies in its potential to provide robustness, adaptability, and generalization in environments where purely modular or purely statistical approaches fall short. The "why" is equally compelling: without conscious-like integration, large-scale AI systems risk remaining brittle, narrow, and incapable of sustaining coherent trajectories of thought across time.

Among theoretical frameworks, Baars' *Global Workspace Theory* (GWT) has been particularly influential in articulating how consciousness might arise from competition and cooperation among

[1] We seem to have come a long way as a society, from Alan Turing's 1951 talk titled, "Intelligent Machinery, A Heretical Theory."

specialist processes. In Baars' original formulation [1], unconscious modules generate candidate representations that compete for access to a global workspace. The winning representation is then broadcast back to all modules, creating a serial stream of conscious experience. This broadcast mechanism explains how diverse processes—perception, memory, reasoning—can be integrated into a unified trajectory. The relevance of GWT to artificial systems is immediate: it offers a blueprint for architectures that move beyond monolithic statistical models toward modular, integrative designs.

Modern large language models (LLMs) have demonstrated extraordinary capabilities in generating coherent, context-sensitive text. Yet, as powerful as they are, standard LLM architectures lack explicit mechanisms for internal competition, broadcast, and modular integration. They excel at local coherence but struggle with global integration across heterogeneous tasks. This gap motivates the construction of hybrid systems that combine LLM-based modules with a global workspace architecture. Such systems promise to inherit the generative fluency of LLMs while embedding them within a structure that enforces competition, selection, and broadcast—the hallmarks of GWT.

To make this motivation concrete, consider two emergent properties that any conscious-like system must balance: *diversity* and *coherence*. Diversity, or distinguishability, refers to the system's ability to represent and explore multiple distinct internal states. It ensures that the system does not collapse into repetitive or narrow trajectories. Coherence, or compatibility, refers to the ability to maintain a consistent and integrated trajectory of thought over time. It ensures that the system does not fragment into incoherence. The tension between these properties is not incidental; it is fundamental. Too much diversity without coherence yields chaos. Too much coherence without diversity yields rigidity. Conscious-like processing emerges precisely in the regime where both are jointly sustained.

This tradeoff has deep analogies in information theory. In multi-user communication systems, capacity regions define the achievable tradeoffs between competing informational demands. Cover's seminal work on broadcast channels [3] and Ahlswede's work on multi-way channels [4] formalized how multiple sources share a common bottleneck. The analogy to consciousness is striking: diversity and coherence act as competing informational demands on the global workspace. Just as rate pairs in a multiple access channel are bounded by a concave frontier, so too are diversity and coherence bounded by a capacity-like region in conscious-like architectures. This connection provides not only metaphorical resonance but also operational tools: entropy, mutual information, and achievable regions become metrics for consciousness.

Complementing Baars' architectural perspective, Tononi's *Integrated Information Theory* (IIT) [2] emphasizes the balance between differentiation and integration. IIT argues that consciousness corresponds to the system's ability to generate a repertoire of distinguishable states (differentiation) while simultaneously integrating them into a unified whole. This dual requirement maps directly onto our definitions of diversity and coherence. Where IIT provides a conceptual measure[2] (Φ), our framework provides operational metrics based on entropy and similarity. Together, they suggest that conscious-like processing is not an all-or-nothing phenomenon but a quantitative regime defined by tradeoffs.

[2] Integrated Information (ϕ). In Integrated Information Theory (IIT), the quantity ϕ measures the extent to which a system generates information that is irreducible to its parts. Formally, ϕ is defined as the distance between the cause–effect repertoire of the whole system and that of its minimum information partition (MIP). That is,

$$\phi = \min_{\text{partitions}} D\big(P_{\text{whole}} \,\|\, P_{\text{partitioned}}\big),$$

where $D(\cdot\|\cdot)$ denotes a suitable distance measure (e.g., Earth Mover's Distance or Kullback–Leibler divergence). A system has high ϕ if its causal structure cannot be decomposed without significant loss of information, indicating a high degree of integration.

The "so-what" of this investigation is therefore twofold. First, by characterizing the diversity-coherence region in LLM-based global workspace systems, we apparently obtain a measurable signature of integrated cognition. This signature allows us to compare architectures, tune parameters, and identify regimes of stability versus collapse. Second, by situating this region within the language of information theory, we bridge cognitive science and engineering. The capacity-like frontier is not merely a metaphor; it is a constraint that guides design. Just as communication engineers design codes to approach channel capacity, AI researchers can design architectures to approach the diversity-coherence frontier.

The "why" is equally clear. Without such a framework, discussions of machine consciousness remain speculative, philosophical, or metaphorical. By contrast, our approach provides quantitative metrics, empirical regimes, and operational signatures. It transforms consciousness from a conceptual aspiration into a measurable property. This shift is essential if machine consciousness is to be taken seriously as a scientific and engineering goal.

In summary, this work builds on three pillars of the literature: Baars' Global Workspace Theory [1], which articulates the architectural mechanism of competition and broadcast; Tononi's Integrated Information Theory [2], which formalizes the dual requirements of differentiation and integration; and Cover's information-theoretic analysis of broadcast channels [3], which provides the mathematical language of capacity regions. By synthesizing these perspectives, we construct a framework in which diversity and coherence are not abstract notions but measurable quantities. Their joint achievable region reflects fundamental limits on conscious-like processing in artificial systems. The outer boundary of this region may serve as an operational signature of integrated cognition, marking the frontier beyond which simple time-sharing or modular alternation cannot reach. It is in this frontier that conscious-like processing emerges, and it is this frontier that we seek to characterize.

Assumptions. This paper assumes: (1) The sentence-transformers model, the modules, and the GWT selection mechanism are optimal and may faithfully represent the real architecture. (2) The used metrics for coherence (C) and diversity (D) - defined later in the paper as they are used in various sections - correlate with measures of system robustness and integration. See also Section 5.1. (3) Entropy of module winners captures the semantic diversity of their proposals. (4) The disruptor is realistic.

Contribution. We demonstrate that a GWT-inspired architecture applied to LLM-based modules exhibits a nontrivial tradeoff between diversity and coherence. The achievable region of these quantities forms a capacity-like structure analogous to multi-user information systems. We argue that this region provides both a diagnostic tool for evaluating architectures and a design principle for constructing systems that sustain conscious-like processing. We also discuss tight relations with synchronization (see Section 4.3) and coding theory (see Section 4.4).

Implication. By answering the "so-what" and "why" questions, we show that machine consciousness is not merely a philosophical curiosity but a practical necessity for robust, adaptive, and generalizable AI. The diversity-coherence frontier is not only a signature of conscious-like processing but also a guide for engineering systems that can thrive in complex, disruptive environments.

Organization of the Work. The paper is structured as follows. Section 1 introduces the motivation for studying machine consciousness through Baars' Global Workspace Theory, situating the problem within the dual questions of "so-what" and "why," and surveying foundational contributions from Baars, Tononi, and Cover. Section 2 formalizes the system model, specifying the modular architecture, the quantitative metrics of diversity and coherence, and the disruption parameter as a probe of stability, while also illustrating snapshots of conscious-like inner dialog. Section 3 presents the empirical results, including visualizations of competition dynamics among modules under varying disruption strengths and the capacity-like tradeoff surface between diversity and coherence. Section 4 interprets these findings through an information-theoretic lens, deriving bounds on the achievable region and connecting the framework to notions of distinguishability, compatibility, and phase-transition behavior. Finally, Section 5 covers "Limitations and Discussion," highlighting the restricted scope of metrics, the limited number of disruption strengths simulated, and the interpretive boundaries of the model, while pointing toward directions for future refinement.

Table of Notation

The adjacent table, Table 1, sets the notational framework for the work, providing principal symbols used and their meanings.

Table 1: Notation used in the present work.

Symbol	**Meaning**
$M = \{M_i\}$	Set of specialist modules (memory, reasoning, analogy, language, disruptor) commonly discussed in cognitive psychology and not necessarily mutually independent
$Z_i(t)$	Proposal generated by module i at time t
$S_i(t)$	Relevance score assigned to proposal $Z_i(t)$
$z(t)$	Selected workspace content at time t (winning proposal)
P_i	Fraction of timesteps for which module i wins
D	Diversity metric: entropy of winning module distribution
C	Coherence metric: average cosine similarity (defined in Subsection 2.2) between consecutive workspace embeddings
$e(z(t))$	Embedding vector of workspace content at time t
$\alpha \in [0, 1]$	Disruption strength parameter controlling influence of disruptor module
R	Achievable region of diversity–coherence pairs under given architecture
K	Effective capacity constant bounding diversity–coherence tradeoff
$D + C \leq K$	Tradeoff constraint analogous to sum-rate bound in multi-user information theory
α_c	Critical disruption threshold beyond which coherence collapses
T_i	Fraction of timesteps allocated to module M_i (time-sharing parameter)
X	Random variable representing the active module at a given timestep
Y	Workspace state random variable
$Y_{\max}$	Maximum achievable coherence (e.g., when a single module dominates)
V_{ij}	Expected similarity $\mathbb{E}[s(Y_t, Y_{t+1}) \mid X = i, X' = j]$
π	Stationary distribution over modules
P_{ij}	Transition probability from module i to module j
$H(X_t)$	Entropy of the stationary distribution (diversity measure)
$H(X_{t+1} \mid X_t)$	Conditional entropy of transitions (used in coherence bounds)
$s(Y_t, Y_{t+1})$	Similarity kernel function measuring coherence between successive states

2 System Model

Building on the motivation outlined in the introduction, we now formalize the architecture through which large language model (LLM)-based modules can be embedded within a Global Workspace Theory (GWT) framework. The goal is to operationalize the tension between diversity and coherence in a way that is both measurable and experimentally tractable. This requires specifying (i) the modular architecture, (ii) the metrics by which diversity and coherence are quantified, and (iii) the role of disruption as a control parameter that probes the system's stability.

2.1 Global Workspace Architecture

The system is composed of multiple specialist modules, each implemented using LLM pipelines or embedding-based processes. Examples include memory retrieval, reasoning, analogy generation, language production, and an explicitly introduced disruptor module. At each timestep, all modules generate candidate outputs ("proposals") along with associated relevance scores. A central selection mechanism chooses the highest-scoring proposal, which becomes the current content of the global workspace. This content is then broadcast back to all modules, influencing their subsequent computations.

Formally, let $\mathcal{M} = \{M_i\}$ denote the set of modules. At time t, each module produces a proposal $z_i(t)$ with score $s_i(t)$. The workspace content is defined as:

$$z(t) = \arg\max_i s_i(t). \tag{1}$$

The broadcast step updates each module's internal state as a function of $z(t)$, thereby implementing parallel competition combined with serial integration, consistent with GWT.

2.2 Metrics: Diversity and Coherence

To characterize system behavior, we define two metrics over the sequence of workspace states $\{z(t)\}_{t=1}^{T}$.

Diversity. Diversity measures the distinguishability of states across time. Operationally, we compute the entropy of the distribution of winning modules or, alternatively, the spread of embeddings associated with $z(t)$. Let p_i denote the fraction of timesteps for which module i wins. Then:

$$D = -\sum_i p_i \log p_i. \tag{2}$$

This captures how many distinct processes contribute to the system's evolution, ensuring that the workspace does not collapse into repetitive trajectories.

Coherence. Coherence measures compatibility between successive states. We compute the average cosine similarity[3] between embeddings of consecutive workspace contents:

$$C = \frac{1}{T-1} \sum_{t=1}^{T-1} \frac{e(z(t)) \cdot e(z(t+1))}{\|e(z(t))\| \, \|e(z(t+1))\|}. \tag{3}$$

High coherence indicates smooth and consistent transitions, while low coherence reflects fragmentation. Together, D and C provide a quantitative lens on the balance between exploration and integration.

2.3 Disruption as a Control Parameter

To probe the robustness of the system, we introduce a disruptor module whose strength is controlled by a parameter $\alpha \in [0, 1]$. The disruptor generates scrambled or noise-like proposals, with scores proportional to α. As α increases, the system experiences greater interference, allowing us to explore transitions between ordered and chaotic regimes.

- **Low disruption ($\alpha \approx 0$):** High coherence, moderate diversity. The system is stable but somewhat repetitive.
- **Intermediate disruption ($0^+ < \alpha < 1^-$):** Balanced diversity and coherence. The system exhibits rich, adaptive behavior.
- **High disruption ($\alpha \approx 1$):** High diversity, low coherence. The system becomes fragmented and unstable.

This parameterization allows us to treat disruption as a tunable control knob, analogous to noise in communication systems or temperature in statistical physics. By varying α, we can map out the achievable region in the diversity–coherence plane and identify critical thresholds where system behavior undergoes qualitative shifts.

2.4 Summary

The system model thus integrates three essential components: a modular architecture inspired by GWT, quantitative metrics for diversity and coherence, and a disruption parameter that probes stability. Together, these elements provide a foundation for empirical analysis and theoretical interpretation. In the sections that follow, we use this model to characterize the achievable region of diversity and coherence, interpret its capacity-like structure, and situate it within broader information-theoretic and cognitive frameworks.

[3] Cosine similarity is a measure of similarity between two nonzero vectors based on the cosine of the angle between them in an inner product space. Given vectors $x, y \in \mathbb{R}^n$, it is defined as

$$\text{cos\_sim}(x, y) = \frac{x \cdot y}{\|x\| \, \|y\|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \, \sqrt{\sum_{i=1}^{n} y_i^2}}.$$

Its value lies in $[-1, 1]$, where 1 indicates identical orientation, 0 orthogonality, and -1 opposite orientation. Cosine similarity is widely used in text analysis and high-dimensional data to compare patterns independent of magnitude.

2.5 Snapshots of Conscious-Like Inner Dialog

In this subsection we present three instances of the inner-dialog of the system. A system is said to be "conscious-like" in our view if it exhibits sustained, self-referential, and context-sensitive patterns of internal dialogue that (i) maintain thematic coherence over time, (ii) reflect an awareness of its own informational limitations, (iii) generate distinctions between alternative states or possibilities, and (iv) demonstrate recursive structure in which statements about reasoning themselves become objects of further reasoning.

> The best way to learn about patterns of reasoning and reasoning is to learn about patterns of reasoning and reasoning in parallel. We are all thinking about our future. We can all make up our minds and minds. The truth is that we don't really have any information about reality. We have learned that there are many ways in which we can communicate. But in order to achieve that we have to live in this

Source: gwtlog2.json

> The difference between the two is that a single machine can do very little more than a single machine. There are some problems, and one is probably not enough. But it's worth noting that the two systems have a very different problem: they're not both: they're both very different from one another. In fact, they have the same problem. The problem is that we also have different problems. We can do pretty much any number of things, most of the time, and if we learn something, we can do it with nothing but

Source: gwtlog3.json

> That is what we should be doing in our everyday life: to understand that consciousness is not a form of consciousness. And that is what we need to be doing in our everyday life: to understand that consciousness is not a form of consciousness. There are two kinds of consciousness: the mind and the physical world. The other kind of consciousness is the mind. The mind is the physical world.

Source: gwtlog55.json

3 Results

Having established the system model and given a flavor for its internal milieu, we now turn to its empirical characterization. The architecture, metrics, and disruption parameter together allow us to generate quantitative data describing how diversity and coherence interact under varying conditions. This section presents the results of those experiments, with particular attention to the visualizations in Figures 1 and 6. These figures illustrate the dynamic competition among modules and the emergent tradeoff surface between diversity and coherence.

3.1 Competition Dynamics in the Workspace

Figure 1 depicts the sequence of streams that compete for the Baars-inspired global workspace over successive timesteps. Each module—memory, reasoning, language, analogy, and disruptor—submits proposals, and the highest-scoring proposal is selected as the conscious content. The vertical axis shows the winning scores, while the labels indicate which module prevailed at each timestep.

This visualization highlights several important features of the system:

- **Serial integration of parallel competition.** Although multiple modules generate proposals simultaneously, only one is selected at each timestep, producing a serial stream of conscious content. This is consistent with Baars' original conception of GWT.
- **Module variability.** The figure shows alternating dominance among modules, with memory and language often prevailing but reasoning and analogy occasionally contributing. This variability reflects the system's ability to sustain diversity.
- **Disruptor influence.** At higher disruption strengths, the disruptor module occasionally wins, producing scrambled or noise-like content. The subsequent recovery of coherence depends on whether other modules can reassert dominance in later timesteps.

Taken together, Figure 1 provides a dynamic view of how competition, broadcast, and disruption interact. It demonstrates that conscious-like processing is not static but emerges from the continual negotiation among modules.

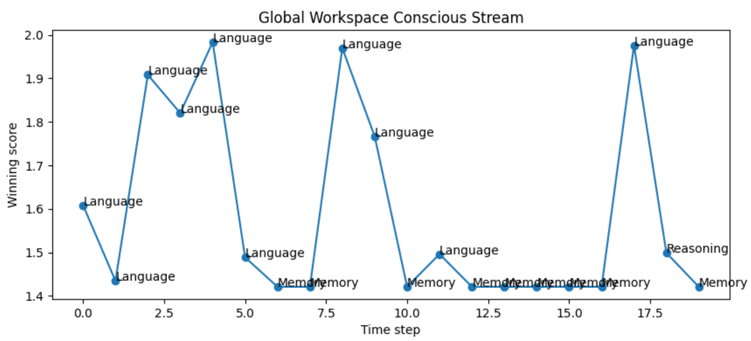

Figure 1: Sequence of streams that compete for the Baars workspace. Each timestep shows the winning module and its score, illustrating the alternation between memory, reasoning, language, and disruptor contributions. Disruption strength is 0 percent.

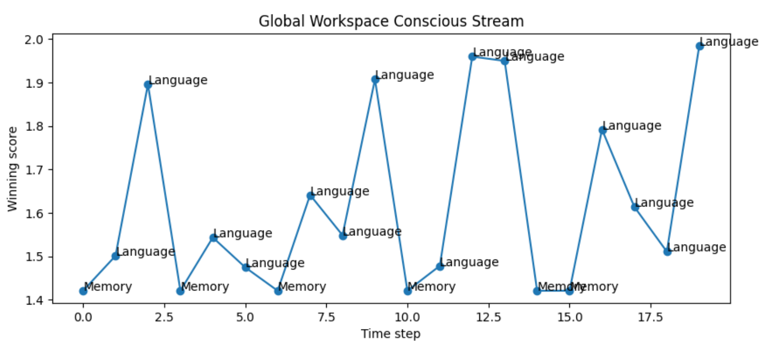

Figure 2: Sequence of streams that compete for the Baars workspace. Each timestep shows the winning module and its score, illustrating the alternation between memory, reasoning, language, and disruptor contributions. Disruption strength is 40 percent.

3.2 Capacity-Like Tradeoff Between Diversity and Coherence

Figure 6 presents the achievable region in the diversity–coherence plane. Each point corresponds to a run of the system under a particular disruption strength α. The horizontal axis represents

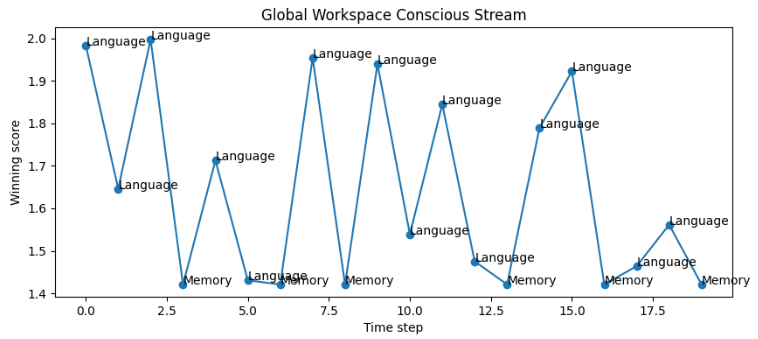

Figure 3: Sequence of streams that compete for the Baars workspace. Each timestep shows the winning module and its score, illustrating the alternation between memory, reasoning, language, and disruptor contributions. Disruption strength is 48.5 percent.

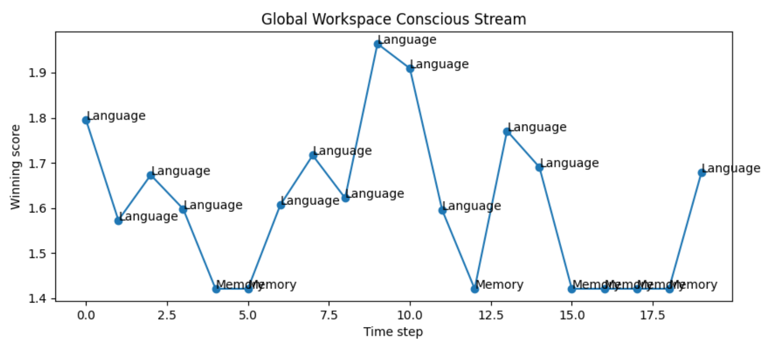

Figure 4: Sequence of streams that compete for the Baars workspace. Each timestep shows the winning module and its score, illustrating the alternation between memory, reasoning, language, and disruptor contributions. Disruption strength is 60 percent.

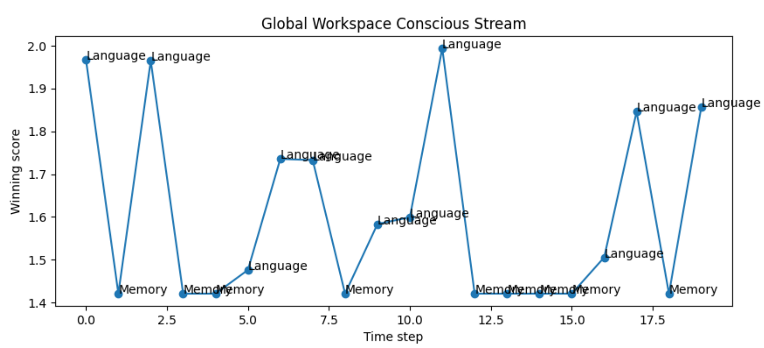

Figure 5: Sequence of streams that compete for the Baars workspace. Each timestep shows the winning module and its score, illustrating the alternation between memory, reasoning, language, and disruptor contributions. Disruption strength is 95 percent.

normalized diversity, while the vertical axis represents normalized coherence. The outer boundary of the region forms a concave frontier, reminiscent of capacity regions in multi-user information theory.

Several observations emerge from this figure:

- **Low disruption regime.** When $\alpha \approx 0$, coherence is high but diversity is moderate. The system produces stable, repetitive trajectories, dominated by a few modules.
- **Intermediate disruption regime.** At moderate values of α, the system achieves a balance: diversity increases without catastrophic loss of coherence. This regime corresponds to rich, adaptive behavior and is arguably the most "conscious-like" in its integration of multiple streams.
- **High disruption regime.** As $\alpha \to 1$, diversity rises sharply but coherence collapses. The system becomes fragmented, with frequent disruptor wins and unstable trajectories.
- **Concave frontier.** The outer boundary of the region is concave, indicating that diversity and coherence cannot be simultaneously maximized. This nontrivial tradeoff mirrors the sum-rate constraints in multiple access channels, where increasing one user's rate reduces the achievable rate of another.

The analogy to information theory is particularly illuminating. Just as achievable rate pairs (R_1, R_2) in a multiple access channel are bounded by mutual information constraints, achievable diversity–coherence pairs (D, C) are bounded by the system's architecture. States along the frontier cannot be reproduced by simple time-sharing among modules. They may require genuine integration.

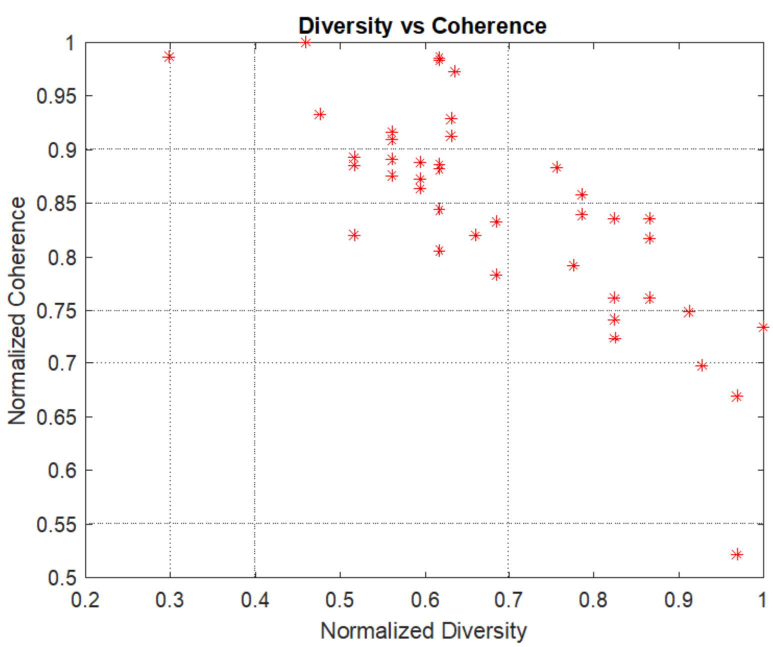

Figure 6: MAC-like outer bound region for diversity and coherence as two "competing users" of the Baars-inspired global workspace. The concave frontier illustrates the tradeoff constraint: increasing diversity reduces coherence, and vice versa.

3.3 Transition Summary

Together, Figures 1 and 6 provide complementary perspectives. The first shows the micro-dynamics of module competition over time, while the second shows the macro-structure of achievable diversity–coherence pairs. The smooth transition from ordered to chaotic regimes, controlled by the disruption parameter α, reveals threshold behavior akin to phase transitions. Below a critical α_c, coherence is sustained; above it, coherence collapses abruptly. Thus it appears that conscious-like processing operates near criticality, balancing diversity and coherence at the edge of instability.

4 Discussion

4.1 Capacity Region Interpretation

We interpret the observed region as an achievable set:

$$\mathcal{R} = (D, C) : \text{achievable under given architecture.} \tag{4}$$

This region exhibits two key properties:

1. **Non-convexity relative to time-sharing:** Points on the boundary cannot be reproduced by alternating between simpler regimes.
2. **Tradeoff constraint:** Increasing diversity beyond a point reduces coherence, and vice versa.

This structure is reminiscent of capacity regions in multi-user communication systems, particularly the multiple access channel (MAC), where multiple sources share a common decoding bottleneck.

4.2 Information-Theoretic Lens on the Diversity–Coherence Tradeoff

Going beyond the simulation framework and its D and C definitions, here we formalize the global workspace as a stochastic process and derive inner and outer bounds on the achievable diversity–coherence region. The resulting structure yields a coupled, capacity-like tradeoff analogous to a multiple access channel. In essence, the cocave hull like bound represents timesharing between coherence and diversity, with both coherence and diversity experiencing a peak cut-off when the other is below a threshold.

4.2.1 Information Theoretic Model

Let $\mathcal{M} = \{1, \ldots, K\}$ denote the set of modules. At each time t, a random variable

$$X_t \in \mathcal{M}$$

represents the index of the winning module.

The workspace emits a state $Y_t \in \mathcal{Y}$ according to a conditional distribution

$$Y_t \sim P_{Y|X}(\cdot \mid X_t).$$

We impose the following assumptions:

(A1) (**Memoryless emissions**) Y_t depends only on X_t.

(A2) (**Stationarity**) $\{X_t\}$ is a stationary stochastic process.

(A3) (**Ergodicity**) Time averages converge to ensemble expectations.

4.2.2 Metrics

Diversity. We define diversity as the entropy rate:

$$D \triangleq H(X_t).$$

Coherence. We define coherence as the mutual information between successive workspace states:

$$C \triangleq I(Y_t; Y_{t+1}).$$

This definition captures temporal predictability and enforces a direct connection to information-theoretic structure.

4.2.3 Achievable Region

Definition 1. A pair (D, C) is achievable if there exists a stationary process $\{X_t\}$ and channel $P_{Y|X}$ such that

$$D = H(X_t), \quad C = I(Y_t; Y_{t+1}).$$

Let $\mathcal{R}$ denote the set of all achievable pairs.

4.2.4 Main Result

Theorem 1 (Diversity–Coherence Tradeoff)**.** The achievable region $\mathcal{R}$ satisfies

$$\mathcal{R}_{\text{in}} \subseteq \mathcal{R} \subseteq \mathcal{R}_{\text{out}},$$

where

$$\mathcal{R}_{\text{in}} = \text{conv} \left\{ \begin{array}{l} (D, C) : D = H(\pi), \\ C \leq I(X_t; X_{t+1}) \end{array} \right\}, \tag{5}$$

$$\mathcal{R}_{\text{out}} = \{(D, C) : 0 \leq D \leq \log K, \quad 0 \leq C \leq D\}. \tag{6}$$

4.2.5 Achievability

Theorem 2 (Achievability)**.** For any stationary Markov chain $\{X_t\}$ with distribution π and transition matrix P_{ij}, the pair

$$D = H(\pi), \quad C \leq I(X_t; X_{t+1})$$

is achievable.

Proof. Construct a stationary Markov chain $\{X_t\}$ with marginal π and transition matrix P_{ij}. Let $Y_t \sim P_{Y|X}(\cdot \mid X_t)$.

By stationarity,

$$D = H(X_t) = H(\pi).$$

For coherence,

$$C = I(Y_t; Y_{t+1}).$$

Since $Y_t - X_t - X_{t+1} - Y_{t+1}$ forms a Markov chain, the data processing inequality yields

$$I(Y_t; Y_{t+1}) \leq I(X_t; X_{t+1}).$$

Thus any pair satisfying

$$C \leq I(X_t; X_{t+1})$$

is achievable.

Time-sharing between two stationary processes preserves achievability and yields convex combinations of (D, C) pairs. Hence the convex hull is achievable. □

4.2.6 Converse

Theorem 3 (Converse)**.** Any achievable pair (D, C) must satisfy

$$0 \leq D \leq \log K, \quad 0 \leq C \leq D.$$

Proof. The diversity bound follows from entropy:

$$D = H(X_t) \leq \log K.$$

For coherence,

$$C = I(Y_t; Y_{t+1}).$$

Using the Markov structure $Y_t - X_t - X_{t+1} - Y_{t+1}$ and data processing,

$$I(Y_t; Y_{t+1}) \leq I(X_t; X_{t+1}).$$

By the chain rule,

$$I(X_t; X_{t+1}) = H(X_t) - H(X_t|X_{t+1}) \leq H(X_t).$$

Thus,

$$C \leq D.$$

Non-negativity of entropy and mutual information gives $C \geq 0$ and $D \geq 0$. □

4.2.7 Interpretation

The outer bound

$$C \leq D$$

establishes a fundamental coupling between diversity and coherence: temporal coherence cannot exceed the information content of the module distribution.

At one extreme:

- $D = 0$: a single module dominates, yielding maximal coherence.

At the other:

- $D = \log K$: modules are uniformly active, but coherence is limited by reduced temporal predictability.

The achievable region is convex and exhibits a concave frontier under suitable transition structures, reflecting a capacity-like tradeoff between diversity and temporal integration.

Theorem 4 (Concavity of the Frontier)**.** The function

$$C_{\max}(D) = \sup\{I(Y_t; Y_{t+1}) : H(X_t) = D\}$$

is concave in D.

Proof. Consider the Lagrangian

$$\mathcal{L}(\pi, P; \beta) = I(X_t; X_{t+1}) - \beta H(\pi).$$

Define

$$F(\beta) = \sup_{(\pi,P)} \mathcal{L}(\pi, P; \beta).$$

Then $F(\beta)$ is convex in β, and

$$C_{\max}(D) = \inf_{\beta \geq 0} \{F(\beta) + \beta D\},$$

which is a pointwise infimum of affine functions in D. Hence $C_{\max}(D)$ is concave. □

4.3 Synchronization, Distinguishability, and Phase Behavior

We now reinterpret the diversity–coherence tradeoff through the lens of *synchronization* in coupled stochastic systems along the lines of [6]. This perspective unifies the notions of distinguishability and compatibility with well-studied phenomena in dynamical systems, neuroscience, and distributed information processing.

From Coherence to Synchronization. In the preceding development, coherence was defined either via embedding similarity or via mutual information between successive workspace states. We now refine this notion by interpreting coherence as a form of *synchronization* among modules.

Let $\{Z_i(t)\}_{i=1}^K$ denote the proposals generated by the K modules at time t. We define a synchronization-based coherence measure as

$$C_{\text{sync}} \triangleq \frac{1}{K(K-1)} \sum_{i \neq j} I\big(Z_i(t); Z_j(t)\big), \tag{7}$$

which captures the degree of statistical dependence among module outputs. High values of C_{sync} indicate that modules are aligned in a shared informational state, while low values indicate independent or competing representations.

An alternative, dynamical interpretation may be obtained by associating each module with a latent phase $\theta_i(t)$ and defining the order parameter

$$R(t) \triangleq \left| \frac{1}{K} \sum_{i=1}^{K} e^{j\theta_i(t)} \right|, \qquad C_{\text{sync}} \approx \mathbb{E}[R(t)], \tag{8}$$

which measures the degree of phase alignment across modules. This formulation connects the present framework to classical models of synchronization such as the Kuramoto system[4].

Diversity as Desynchronization. Under this reinterpretation, diversity corresponds to the degree of *desynchronization* or dispersion across modules. The entropy-based definition

$$D = H(X_t) \tag{9}$$

may be viewed as measuring the uncertainty in the identity of the dominant module, and thus the spread of activity across competing processes.

More generally, diversity captures the extent to which modules occupy distinct informational or dynamical states. In the phase-based picture, this corresponds to dispersion in $\{\theta_i(t)\}$, while in the information-theoretic picture it corresponds to reduced dependence among $\{Z_i(t)\}$.

---

[4]**Kuramoto System:**

A *Kuramoto system* is a model of a population of N coupled phase oscillators. Each oscillator is characterized by a phase $\theta_i(t) \in \mathbb{S}^1$ and a natural frequency $\omega_i \in \mathbb{R}$. The dynamics are governed by the system of differential equations

$$\frac{d\theta_i}{dt} = \omega_i + \frac{K}{N} \sum_{j=1}^{N} \sin\big(\theta_j - \theta_i\big), \qquad i = 1, \ldots, N,$$

where $K \geq 0$ is the global coupling strength.

The Kuramoto system exhibits a transition from incoherence to partial or full synchronization as the coupling K increases. A common measure of synchronization is given by the complex *order parameter*

$$r(t)e^{i\psi(t)} = \frac{1}{N} \sum_{j=1}^{N} e^{i\theta_j(t)},$$

where $r(t) \in [0, 1]$ quantifies the degree of phase coherence and $\psi(t)$ represents the average phase.

Distinguishability and Compatibility. We may therefore reinterpret:

- **Distinguishability** as desynchronization, i.e., the ability of the system to sustain multiple, weakly coupled or independent states.
- **Compatibility** as synchronization, i.e., the ability of the system to align its components into a unified, mutually consistent state.

This mapping provides a concrete operationalization of longstanding notions in consciousness research, where differentiation and integration are viewed as dual requirements for conscious processing.

Tradeoff as a Synchronization Constraint. Within this framework, the diversity–coherence tradeoff becomes a constraint between dispersion and alignment in a coupled system. The outer bound

$$C \leq D \tag{10}$$

may be reinterpreted as stating that the degree of achievable synchronization is limited by the informational diversity of the system.

More refined characterizations may be obtained by introducing an effective coupling parameter κ (implicit in the scoring, selection, and broadcast mechanisms) and the disruption parameter α. The achievable region may then be viewed as

$$\mathcal{R}(\kappa, \alpha) = \{(D, C_{\text{sync}})\}, \tag{11}$$

with the boundary determined by the interplay between coupling and disruption.

Phase Behavior and Criticality. The introduction of synchronization reveals a natural phase structure in the system:

- **Desynchronized regime** (high α, low effective coupling): modules behave largely independently, yielding high diversity and low coherence.
- **Partially synchronized regime** (intermediate α): modules exhibit intermittent alignment, producing a balance of diversity and coherence. This regime corresponds to the empirically observed "conscious-like" operating region.
- **Fully synchronized regime** (low α, strong coupling): one or a few modules dominate, leading to high coherence and reduced diversity.

Empirically, these regimes are separated by a critical disruption threshold α_c, beyond which coherence approaches zero rapidly. This behavior is consistent with phase transitions observed in coupled dynamical systems, where synchronization emerges only when coupling exceeds a critical value.

Interpretation. This synchronization-based view provides a unifying interpretation of the global workspace dynamics. Rather than treating diversity and coherence as independent metrics, we view them as dual aspects of a single underlying phenomenon: the degree of coordination among interacting modules.

In this interpretation, conscious-like processing corresponds to operation near the boundary between synchronization and desynchronization, where the system simultaneously maintains a rich repertoire of states and the ability to integrate them into a coherent trajectory. This boundary coincides with the concave frontier of the achievable region and represents a regime that cannot be reproduced by simple time-sharing, but instead requires genuine interaction among modules.

4.4 Coherence as Coding

The synchronization-based interpretation developed in the previous subsection admits a complementary, structural perspective rooted in coding theory [7]. While synchronization characterizes how modules dynamically align over time, coding theory characterizes which joint configurations of module outputs are *admissible* in the first place. This distinction allows us to reinterpret coherence not merely as alignment, but as membership in a structured set of valid joint states.

Codebook Structure. Let $\{Z_i(t)\}_{i=1}^{K}$ denote the outputs of the K modules at time t. We posit the existence of a subset

$$\mathcal{C} \subseteq \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_K$$

of admissible joint configurations, which we interpret as a *codebook*. An element of $\mathcal{C}$ represents a mutually compatible combination of module outputs. In contrast, configurations outside $\mathcal{C}$ correspond to incoherent or semantically inconsistent states.

This viewpoint is analogous to linear coding theory, where only specific linear combinations of codewords yield valid codewords, while arbitrary combinations produce invalid or undecodable sequences. In the present setting, each module output plays the role of a component of a distributed codeword, and coherence corresponds to whether the joint configuration lies within the admissible set $\mathcal{C}$.

Coherence as Code Membership. Under this interpretation, coherence can be viewed as the probability that the system occupies a valid codeword:

$$C \;\approx\; \mathbb{P}\big((Z_1(t), \ldots, Z_K(t)) \in \mathcal{C}\big).$$

High coherence corresponds to trajectories that remain within the codebook, while low coherence corresponds to excursions into invalid configurations. This reframes the earlier similarity- and information-based measures of coherence as proxies for code membership.

Diversity as Codebook Exploration. Diversity, defined as $D = H(X_t)$, measures how broadly the system explores different modules and, by extension, different regions of the codebook. From the coding perspective, diversity reflects the extent to which multiple distinct codewords are visited over time. Thus, the diversity–coherence tradeoff may be interpreted as a tension between exploring a large number of codewords and remaining within the admissible set.

Noise, Disruption, and Decoding. The disruption parameter α may be interpreted as a noise level that perturbs the system away from valid codewords. At low α, the system remains close to $\mathcal{C}$, yielding high coherence. As α increases, the system is more likely to produce configurations outside the codebook, resulting in incoherent or “gibberish” states. In this sense, the global workspace may be viewed as performing an implicit decoding operation, selecting and broadcasting the most plausible codeword given noisy proposals.

Relation to Synchronization. The coding and synchronization perspectives are complementary. Synchronization describes the dynamical process by which modules align, while the codebook structure defines the manifold onto which such alignment must project in order to be meaningful. High synchronization without adherence to $\mathcal{C}$ would yield aligned but semantically invalid states, whereas valid codewords require both alignment and structural compatibility.

Interpretation. This perspective suggests that conscious-like processing corresponds to trajectories that remain within a structured manifold of admissible states under noisy, coupled dynamics. The diversity–coherence frontier can thus be reinterpreted as the boundary between reliable decoding (remaining within the codebook) and failure modes where the system explores invalid configurations. In this sense, coherence is not merely smoothness or predictability, but membership in a constrained, information-bearing structure.

5 Limitations and Further Observations

While the results presented in the preceding section provide a compelling picture of the diversity–coherence tradeoff in a Global Workspace–inspired LLM system, several limitations must be acknowledged. These limitations highlight the provisional nature of the findings and point toward directions for future refinement.

5.1 Metric Limitations

The analysis relies primarily on two metrics: entropy-based diversity and cosine-similarity based coherence. Although these measures are intuitive and computationally tractable, they represent only a narrow slice of the broader conceptual landscape of consciousness. Entropy captures distinguishability among module contributions, but it does not fully account for the semantic richness or functional relevance of those contributions. Similarly, cosine similarity between embeddings provides a proxy for coherence, yet it reduces integration to a geometric measure of proximity in latent space. This simplification ignores higher-order temporal dependencies, causal relationships, and task-specific compatibility that may be crucial for genuine conscious-like processing.

Moreover, the justification for these metrics is minimal. They were chosen for their operational convenience and alignment with existing information-theoretic analogies, rather than as the result of a comprehensive evaluation of alternatives. Other measures, such as mutual information, predictive stability, or integrated information (Φ), could provide complementary or even superior insights. The reliance on only two metrics thus constrains the interpretive scope of the results.

5.2 Interpretive Boundaries

Finally, the framework itself is limited by its reliance on a small set of modules (memory, reasoning, analogy, language, disruptor) and a simplified competition mechanism. Real-world cognitive architectures involve far richer interactions, hierarchical organization, and multi-level integration. The present model captures only the first-order dynamics of competition and broadcast. As such, the results should be interpreted as a proof-of-concept rather than a definitive characterization of machine consciousness.

5.3 Information Theoretic Alternatives

Though the work suggests a converse-achievability like inner and outer bounded "capacity" region for the coherence-diversity pair, another line of analysis could have been considered, but was not. We are referring to [5] where the system is the Kalman filter. Restructuring that work to apply to the Baars setting would have captured nuances of entropy movement whereas the inner-outer bound style of the present work has an asymptotic origin.

6 Summary, Conclusion and Future Work

Despite these limitations, the study demonstrates that even a simplified system exhibits nontrivial tradeoffs between diversity and coherence. The concave frontier observed in the diversity–coherence plane provides a useful operational signature, suggesting that conscious-like processing may indeed be bounded by capacity-like constraints. The limitations identified here underscore the need for richer metrics, broader experimental sweeps, and more complex architectures - tasks for a future work. Addressing these gaps will be essential for advancing from preliminary demonstrations to robust, general frameworks for machine consciousness.

In summary, the results should be viewed as an initial step: they provide evidence that diversity and coherence interact in measurable ways, but they do so under restricted conditions. Future work must expand the metric set, justify choices more rigorously, and explore larger parameter spaces to fully capture the dynamics of conscious-like processing in artificial systems. Neurocognitive applications such as to diseased states of the brain and psyche must also be explored under the present framework.

Acknowledgments

This work was produced with the assistance of language models. A.C. thanks Dr. Lav Varshney for numerous formative discussions on the overlapping zone between information theory and the math of consciousness.

Appendix: Further Experimentation Confirms Hull

In this appendix we document additional experiments conducted to test the robustness of the diversity–coherence tradeoff reported in the main text. The central question is whether the observed concave achievable region is an artefact of the specific similarity-based metrics used in earlier versions, or a structural property of the underlying Global Workspace dynamics.

.1 Experimental Extension: Version 3

Version 3 (v3) extends the original simulator along three key dimensions. First, the number of specialist modules is increased to ten, including Memory, Reasoning, Analogy, Language, Disruptor, Emotion, Attention, Metacognition, Prediction, and Curiosity. This expansion introduces richer competition dynamics and reduces the likelihood that previously observed behavior was due to a small or biased module set.

Second, the simulation engine is refactored into a parallel-competition framework with controlled stochasticity. In particular, a small Gaussian perturbation is added to proposal scores at each timestep to avoid degeneracies arising from discretized similarity values. Third, a systematic sweep over disruption strength $\alpha \in [0, 1]$ is performed across 200 trials, with identical architectural configuration and varying noise levels.

Under these conditions, diversity and coherence are measured using embedding-based quantities: diversity as one minus mean pairwise cosine similarity across broadcast states, and coherence as the average cosine similarity between consecutive states. The resulting scatter cloud exhibits a clear concave upper boundary in the diversity–coherence plane, consistent with the capacity-like region described in the main text.

.2 Metric Replacement: Version 4

To test whether this concave structure depends on the embedding-based metric, Version 4 (v4) replaces both diversity and coherence with information-theoretic quantities while leaving the simulation dynamics unchanged.

Let $W = (w_0, \ldots, w_{T-1})$ denote the sequence of winning module identities. Diversity is defined as the normalized Shannon entropy:

$$D = \frac{H(W)}{\log N}, \tag{12}$$

where N is the number of modules. Coherence is defined as normalized next-step predictability:

$$C = 1 - \frac{H(W_{t+1} \mid W_t)}{\log N}. \tag{13}$$

These quantities depend only on the discrete winner process and are independent of any embedding geometry.

Importantly, both v3 and v4 metrics are computed on the *same underlying runs*. That is, for each trial in the disruption sweep, the workspace trajectory is generated once, and both metric families are evaluated post hoc. This ensures that any observed agreement cannot be attributed to differences in simulation conditions.

.3 Hull Comparison

Figure 7 presents a side-by-side comparison of the achievable regions under the two metric regimes. Each panel corresponds to 200 trials with disruption strength swept linearly from 0 to 1, and points are color-coded by α.

Two key observations emerge:

(1) Persistence of the Concave Frontier. Both panels exhibit a clear concave upper boundary: as diversity increases, coherence decreases in a nonlinear fashion. No trial achieves simultaneously high diversity and high coherence. This confirms that the tradeoff is not specific to cosine similarity, but persists when measured using entropy and conditional entropy.

(2) Consistent Negative Dependence. The relationship between diversity and coherence remains strongly negative across both metrics. In v3, the Pearson correlation is approximately -0.86, while in v4 it is approximately -0.77. Although the magnitude decreases slightly under the information-theoretic metric, the sign and qualitative structure are preserved.

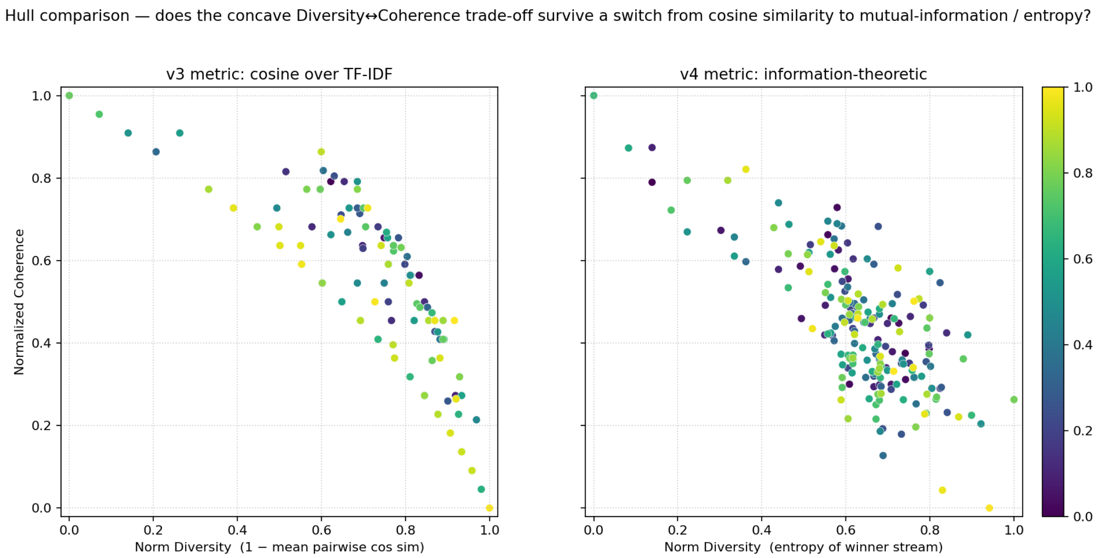

Figure 7: Comparison of diversity–coherence tradeoffs under two metric regimes. Left: v3 metric based on cosine similarity over TF-IDF embeddings. Right: v4 metric based on entropy and next-step predictability of the winner stream. Each point corresponds to a single trial; color encodes disruption strength.

(3) Increased Resolution Under Information-Theoretic Metrics. The v4 scatter exhibits significantly less quantization than v3. Whereas cosine similarity produces clustered values due to the discrete TF-IDF vocabulary, entropy-based measurements vary continuously with the empirical distribution of transitions. As a result, the achievable region in v4 is more densely populated, revealing a smoother approximation to the underlying frontier.

(4) Metric-Specific Saturation Effects. Differences between the panels are most visible in the high-coherence regime. Under v4, coherence saturates rapidly when a single module dominates the winner stream, reflecting the probabilistic nature of predictability. In contrast, cosine similarity continues to increase gradually as lexical overlap grows. Despite this difference, the global shape of the achievable region remains consistent.

.4 Interpretation

The persistence of the concave hull under a complete change of measurement framework provides strong evidence that the diversity–coherence tradeoff is a property of the *workspace dynamics* rather than of the metric used to evaluate it. In particular, both geometric (embedding-based) and information-theoretic (distribution-based) measurements recover the same qualitative constraint: increasing diversity reduces achievable coherence.

This observation supports the interpretation advanced in the main text, where the diversity–coherence region is viewed as analogous to a capacity region in multi-user information theory. The concave frontier appears to reflect a structural limitation imposed by competition and broadcast in the global workspace, rather than an artefact of representation.

.5 Conclusion of Appendix

Taken together, the v3 and v4 experiments demonstrate that the concave diversity–coherence hull is robust to (i) increased architectural complexity, (ii) large-scale stochastic sweeps, and (iii) replacement of similarity-based metrics with information-theoretic ones. While further work is required to formalize this phenomenon analytically, these results substantially strengthen the claim that the observed tradeoff is intrinsic to the dynamics of modular, broadcast-based cognitive systems.

.6 Addendum Control Experiment: Removing the Disruptor (Version 5)

To isolate the role of exogenous noise in shaping the diversity–coherence tradeoff, we perform a control experiment in which the Disruptor module is removed from the system. The Disruptor is the only specialist whose behavior explicitly scales with the disruption parameter α, injecting stochastic or adversarial content into the workspace. Its removal therefore tests whether the concave achievable region is intrinsic to the global workspace dynamics or induced by explicit noise injection.

Experimental Design. Version 5 (v5) repeats the same 200-trial experimental protocol used in v3 and v4, with identical timestep horizon ($T = 30$), module implementations, and seeds, but with the Disruptor excluded from the enabled module set. This reduces the system from $N = 10$ to $N = 9$ specialists.

Each trial corresponds to a full workspace trajectory generated under a fixed random seed. For trial index i, the system is initialized with seed $s_i = s_0 + i$, which deterministically specifies all stochastic elements of the run, including proposal generation, score perturbations, and selection outcomes. Thus, across trials, variation arises from distinct realizations of the same stochastic dynamical process.

In the absence of the Disruptor, the disruption parameter α no longer acts as an exogenous control variable. Although it is retained in the configuration for consistency, it does not directly influence any module behavior. Consequently, the sweep over α in v5 should be interpreted as indexing stochastic trials rather than as a causal perturbation.

For each trajectory, both metric families are evaluated:

- v3 (embedding-based): cosine similarity–based coherence and diversity.
- v4 (information-theoretic): entropy-based diversity and next-step predictability.

As in prior experiments, both metrics are computed on identical underlying trajectories.

Results. Figure 8 presents a 2×2 comparison: rows correspond to module roster (full vs. no Disruptor), and columns correspond to metric regime (v3 vs. v4). Each point represents one trajectory; the horizontal axis is normalized diversity and the vertical axis is normalized coherence.

Three observations follow.

(1) Persistence of the concave hull. The diversity–coherence tradeoff remains clearly visible after removing the Disruptor. In both metric regimes, the achievable region exhibits a concave upper boundary, and no trajectories occupy the high-diversity/high-coherence region. This demonstrates that the hull is not generated by explicit noise injection, but is intrinsic to the competition-and-broadcast dynamics of the workspace.

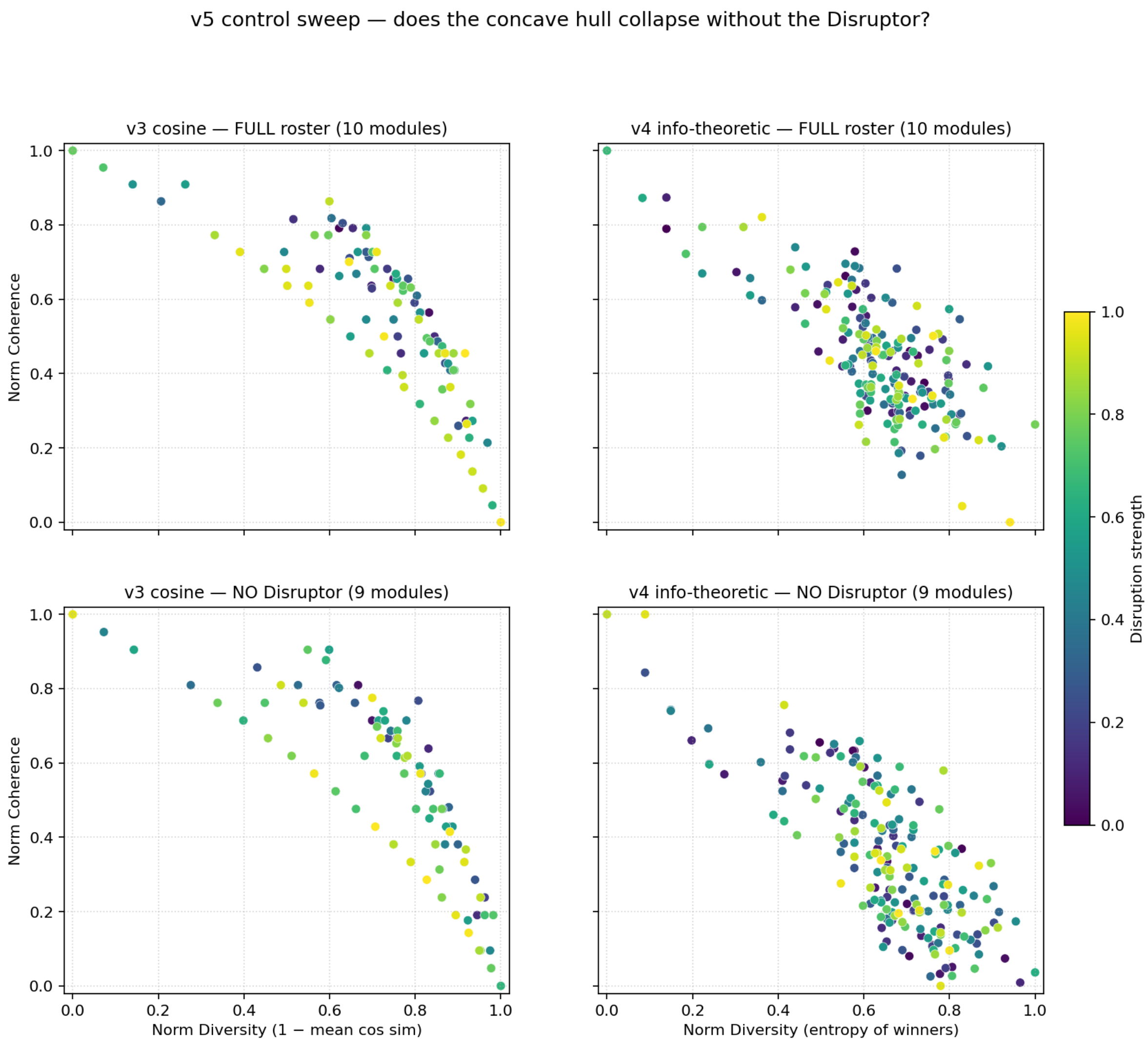

Figure 8: Control sweep comparing the full system (with Disruptor) and the ablated system (without Disruptor), under both cosine-based (v3) and information-theoretic (v4) metrics. Each point corresponds to a single stochastic trajectory. Colour encodes the nominal disruption parameter α, which is inactive in the no-Disruptor condition and serves only as a trial index.

(2) Asymmetric contraction of the achievable region. While the qualitative shape of the region is preserved, its extent is reduced. The primary change is the loss of trajectories in the high-diversity/low-coherence regime (right–bottom portion of the plane). As a result, the cloud no longer spans the full range observed in the presence of the Disruptor, and instead occupies a narrower subset of intermediate diversity and coherence values. Importantly, this contraction does not correspond to a collapse toward a single corner, but rather to a truncation of one side of the feasible region.

(3) Consistency across metrics. The same qualitative behavior is observed under both metric families. The information-theoretic representation produces a smoother and more finely resolved cloud, while the cosine-based representation exhibits mild quantization, but the geometry of the tradeoff and the effect of removing the Disruptor remain consistent.

Interpretation. These results suggest a decomposition of the diversity–coherence tradeoff into two components:

1. **Endogenous stochastic dynamics:** Competition among heterogeneous modules, combined with stochastic proposal generation and selection, is sufficient to produce a concave diversity–coherence tradeoff.

2. **Exogenous disruption:** The Disruptor expands the reachable region by enabling trajectories in the high-diversity/low-coherence regime.

Thus, the Disruptor is not necessary for the existence of the tradeoff, but it increases the extent of the achievable region. In an information-theoretic interpretation, it enlarges the set of attainable (D, C) pairs without altering the qualitative structure of the boundary.

Conclusion of Control Study. The v5 ablation demonstrates that the concave diversity–coherence hull is:

- *intrinsic* to modular competition under broadcast,
- *robust* to removal of explicit noise sources, and
- *expanded*—but not created—by exogenous disruption.

Taken together[5] with the v3 and v4 results, this provides converging evidence that the diversity–coherence tradeoff reflects a structural property of the global workspace dynamics rather than a metric artefact or a consequence of externally injected noise.

.7 Postscript: Related Studies in the Literature

While the specific framing of consciousness-like processing as a capacity-constrained communication problem—with an empirical characterization of the achievable (diversity, coherence) region as a concave hull analogous to multiple-access channel (MAC) capacity regions—is novel, the broader idea of embedding large language models (LLMs) within a Global Workspace Theory (GWT)-inspired modular architecture is not entirely new. Prior computational implementations and agent frameworks have explored specialist modules that compete for access to a shared global workspace, followed by selection and broadcast mechanisms, to achieve more integrated, robust, and human-like cognition in LLM-based systems.

The following works represent key examples from the recent literature (primarily 2024–2026):

- **Unified Mind Model (UMM) and MindOS** [8]: Proposes a cognitive architecture explicitly built upon Baars' Global Workspace Theory as its macro-level structure. Specialist modules handle perception, planning, reasoning, tool use, memory, reflection, and motivation, with LLMs serving as the core processing engine. A central global workspace coordinates competition and integration, enabling human-like cognitive abilities. The authors further introduce MindOS, an agent-building engine that allows creation of domain-specific autonomous agents via natural language descriptions.

[5] As a further check, the experiments were repeated on a second personal computer as well and the diversity-coherence concave hull was confirmed.

- **Global Workspace Agents (GWA) / "Theater of Mind" for LLMs** [9]: Introduces Global Workspace Agents, a cognitive architecture directly inspired by GWT. It features a central broadcast hub coupled with a diverse set of functionally specialized LLM-based agents. The system implements a dynamic cognitive cycle with explicit competition for the workspace, aiming to overcome limitations of static memory pools and enable sustained, self-directed reasoning in LLM systems.

- **CogniPair / GNWT-Agents** [10]: Presents what is described as a computational implementation of Global Neuronal Workspace Theory (GNWT) for LLM-based generative agents. Multiple specialized sub-agents (covering emotion, memory, social norms, planning, and goal-tracking) compete and coordinate through a global workspace broadcast mechanism. The framework was validated in large-scale social simulations (e.g., 551 agents) and demonstrated improved psychological authenticity in tasks such as dating and hiring scenarios.

- **Language Agents and Global Workspace Theory** [11]: Analyzes existing LLM-based language agents (including memory-stream architectures such as those in Park et al.'s Generative Agents) through the lens of GWT. The authors discuss partial compatibility with GWT principles and propose architectural modifications to better incorporate competition, selection, and global broadcast, potentially enhancing conscious-like properties in agentic systems.

- **Brain-Inspired Graph Multi-Agent Systems (BIGMAS)**: Employs a graph-structured multi-agent setup for LLM reasoning, with specialized agents coordinating exclusively through a centralized shared workspace inspired by GWT principles.

- **Pre-LLM Foundations and Extensions**: Earlier symbolic cognitive architectures such as LIDA (Learning Intelligent Distribution Agent) [12, 13] implemented GWT with competing coalitions, attention codelets, and workspace broadcast. Modern discussions have explored extensions or mappings of these ideas onto LLM-based systems. Related theoretical analyses, including those by Chalmers (2023) and others, have examined the potential (and limitations) of incorporating global workspace mechanisms into LLMs or LLM-augmented agents.

These works demonstrate an active research direction in adapting GWT to overcome limitations of monolithic LLMs (e.g., lack of explicit internal competition, brittle cross-task integration, and reduced robustness under noise or heterogeneity). Most focus on architectural design, functional integration, or application-specific validation. In contrast, the present study provides an operational, information-theoretic characterization of the emergent diversity–coherence tradeoff in such systems, treating disruption as a tunable probe and identifying a candidate quantitative signature of integrative cognition.

Essential Reading

[1] Baars, B. J. (1988). *A cognitive theory of consciousness.* Cambridge University Press.

[2] Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience, 5*(1), 42. https://doi.org/10.1186/1471-2202-5-42

[3] Cover, T. M. (1972). Broadcast channels. *IEEE Transactions on Information Theory, 18*(1), 2–14. https://doi.org/10.1109/TIT.1972.1054746

[4] Ahlswede, R. (1971). Multi-way communication channels. In *Proceedings of the 2nd International Symposium on Information Theory* (pp. 23–52). Tsahkadsor, Armenian SSR.

[5] Mitter, S. K., & Newton, N. J. (2005). Information and entropy flow in the Kalman–Bucy filter. *Journal of Statistical Physics, 118*(1), 145–176. https://doi.org/10.1007/s10955-004-8821-3

[6] Chawla, A. (2024). Some thoughts on cosmic-conscious devices. *Zenodo.* https://doi.org/10.5281/zenodo.10606490

[7] Chawla, A. (2023). Of codes and consciousness. *Zenodo.* https://doi.org/10.5281/zenodo.10425949

[8] Hu, P., & Ying, X. (2025). Unified Mind Model: Reimagining Autonomous Agents in the LLM Era. arXiv preprint arXiv:2503.03459.

[9] "Theater of Mind" for LLMs: A Cognitive Architecture Based on Global Workspace Theory. arXiv:2604.08206v1 [cs.MA], April 2026.

[10] Li, A., et al. (2025). CogniPair: From LLM Chatbots to Conscious AI Agents – GNWT-Based Multi-Agent Digital Twins for Social Pairing – Dating & Hiring Applications. arXiv preprint (related to 2506.03543).

[11] Goldstein, S., & Kirk-Giannini, C. D. (2024). Language Agents and Global Workspace Theory. arXiv preprint arXiv:2410.11407.

[12] Franklin, S., & Graesser, A. (1999). A software agent model of consciousness. Consciousness and Cognition, 8(3), 285–301.

[13] Franklin, S. (2003). A conscious artifact? Journal of Consciousness Studies, 10, 47–66.

[14] Chalmers, D. J. (2023). Could a Large Language Model be Conscious? (Manuscript).