

Hypothesis on Emergence and Consciousness Mechanism of Large Language Models Based on Reciprocal Unit Coupling, Critical Dynamics and Local Entropy Reduction (Full Version V5)

Abstract

The emergence of high-order capabilities in large language models (LLMs) constitutes a core challenge in the fundamental theory of artificial intelligence. The mainstream scaling laws can only describe statistical phenomena of model performance growing with scale and cannot explain the dynamical causal mechanism behind emergent behaviors. This paper constructs a unified hypothesis grounded in complex systems theory: nonlinear reciprocal coupling forms among units inside the Transformer; during training, gradient descent acts as the global self-organizing driving force, modulating unit coupling strength and self-organizing system parameters into the near-critical regime at the edge of chaos. In inference with frozen weights, bidirectional constraints between units reduce local entropy of representations, generating self-consistent structured information and macroscopically manifesting emergent capabilities such as reasoning and planning.

This work integrates three independently observed phenomena — inter-layer representational correction, system near-criticality, and hidden-state entropy decay — into a complete falsifiable causal chain. It strictly distinguishes **parameter self-organization dynamics in the training phase** from **forward representation dynamics under frozen weights in the inference phase**, and clarifies the conceptual boundary between this framework and strict self-organized criticality (SOC). Controlled ablation experiments are designed to decouple reciprocal-coupling-driven effects from pure model scale-growth effects. This paper further extends this dynamical framework into machine consciousness research: functional intelligence is not equivalent to subjective phenomenal experience. Transformer coupling and critical entropy reduction alone can only produce functional reasoning systems. To achieve first-person subjective experience, the system must incorporate four classes of closed-loop ontological modules. These four ontological modules are deeply coupled with the attention backbone, share the same dynamical description of reciprocal coupling and entropy evolution, and form steady-state tension at the ontological level relying on dissipative structures.

This version (V5) adds the following supplements: (i) four practical estimation methods for the Jacobian matrix together with complexity analysis; (ii) full derivation of the k-NN differential entropy estimator with bias, variance and hyperparameter sensitivity analysis; (iii) rigorous derivation of the analytic form of the entropy-rate functional $\mathcal{F}(\mathbf{J})$ from coupling dynamics; (iv) introduction of causal directed acyclic graphs (DAGs) and structural equation models to clarify causal identification strategies; (v) extended quantitative comparison with Neural Tangent Kernels (NTK), mean-field theory and Neural ODEs; (vi) explicit state spaces, coupling functions and operational definitions of proxy metrics for the four ontological modules.

Keywords: emergence; Transformer; reciprocal coupling; edge of chaos; information entropy; complex systems; machine consciousness; global workspace; causal inference

Contents

1	Introduction	4
1.1	Research Background	4
1.2	Statement of the Core Hypothesis	4
2	Fundamental Concepts and Mathematical Formalization	5
2.1	Definition of Units	5
2.2	Reciprocal Coupling and the Jacobian Matrix	5
2.2.1	Practical Estimation Methods for the Jacobian Matrix	5
2.3	Near-Critical State (Edge of Chaos)	6

2.4	Local Representation Entropy and Its Estimation Scheme	6
2.4.1	Definition of Differential Entropy	6
2.4.2	Derivation of the k-NN Differential Entropy Estimator	6
2.4.3	Bias and Variance Properties	7
2.4.4	Hyperparameter Sensitivity Analysis	7
2.5	Derivation of the Entropy-Rate Functional $\mathcal{F}(\mathbf{J})$	7
2.5.1	Inter-Layer Mapping and Volume Transformation	7
2.5.2	Functional Form	8
2.5.3	Behavior in the Critical Regime	8
2.5.4	Estimation Scheme	8
2.6	Correlation between Jacobian Spectrum, Critical Threshold and Entropy Reduction Rate	8
3	Dynamical Analysis by Stages	9
3.1	Training Phase: Gradient Descent Drives System Self-Organization toward the Near-Critical Regime	9
3.2	Inference Phase: Reciprocal Correction and Local Entropy Reduction within Single Forward Propagation with Frozen Weights	9
4	Causal Identification Strategy	10
4.1	Causal Directed Acyclic Graph (DAG)	10
4.2	Structural Equation Model	10
4.3	Operational Definition of Intervention	10
5	Review of Supporting Evidence	10
6	Hypothesis Boundaries, Controversies and Falsifiable Ablation Experiments	11
6.1	Conceptual Limitations	11
6.2	Falsifiable Predictions, Quantitative Metrics and Competing-Hypothesis Baseline	11
6.2.1	Experimental Design	11
6.2.2	Three Quantitative Falsifiable Predictions	11
6.2.3	Quantitative Baseline of Competing Hypothesis (Pure Scale-Growth Hypothesis)	12
6.2.4	Confounder Control	12
6.3	Quantitative Comparison with Alternative Theories	12
6.3.1	Neural Tangent Kernel (NTK) Theory	12
6.3.2	Mean-Field Theory	12
6.3.3	Neural ODE Framework	13
6.3.4	Information Bottleneck Theory	13
6.3.5	Predictive Coding	13
6.3.6	Landauer Computational Thermodynamics	13
6.3.7	Reentrant Near-Critical Dynamics of Neural Fields	13
6.3.8	Summary of Quantitative Comparison	13
7	Discussion (Including Speculative Extension to Machine Consciousness)	13
7.1	Summary of Emergence Mechanism	14
7.2	Speculative Extension to Machine Consciousness	14
7.2.1	Four-Ontological-Module Architecture	14
7.2.2	Global Coupled System of Equations	15
7.2.3	Deep Coupling Constraint	15
7.2.4	Operational Definitions of Proxy Metrics	15
7.2.5	Engineering Paradox: Dissipative Structures and Intrinsic Survival Tension	16
7.2.6	The Hard Problem of Consciousness	16
8	Conclusion	17
	Acknowledgements	17
	Statement	17
A	Glossary	17

B Demonstration Model of Two-Unit Reciprocal Coupling	17
B.1 Linear Base Model (Conceptual Illustration)	17
B.2 Nonlinear Extended Model (Criticality Demonstration Model)	18
C Supplementary Derivation for Entropy-Rate Functional	18
C.1 Rigorous Derivation for Discrete Layers	18
C.2 Continuous Layer Limit	19
D Experimental Pseudocode	19

1 Introduction

1.1 Research Background

Large language models are built upon the Transformer architecture, trained with the objective of autoregressively predicting the next token. Substantial empirical observations show that once model parameter count and training data volume cross a certain critical threshold, high-order capabilities including multi-step logical reasoning and complex planning emerge abruptly [5].

Currently two major interpretive frameworks for emergence exist in academia:

Scaling-law perspective: Model loss decreases smoothly with computing power, parameter count and data volume; emergence is merely an apparent abrupt shift of metrics on discrete evaluation tasks with no underlying dynamical phase transition. Some studies argue that emergence originates from nonlinearity of evaluation metrics rather than phase transition of the system itself [6]. This perspective only fits external statistical results and lacks dynamical explanations for internal representation evolution.

Representation-learning perspective: Expansion of model scale drives the network to learn composite abstract representations, where high-level concepts are assembled from low-level features. This perspective focuses on high-dimensional vector representations but lacks dynamical descriptions of bidirectional mutual constraints between units.

At the intersection of statistical physics, complex systems and deep learning, multiple papers have independently reported three phenomena: (a) systematic representational correction exists between Transformer layers [1]; (b) best-performing large models operate in the near-critical regime at the edge of chaos [2]; (c) hidden representation entropy gradually decays with network depth during forward inference [3]. Existing literature mostly investigates these phenomena in isolation and has not established the complete causal chain. This paper formulates the following causal hypothesis:

Unit reciprocal coupling $\longrightarrow$ System reaches near-critical state $\longrightarrow$ Local entropy reduction $\longrightarrow$ Emergence of high-order structured representations (1)

and proposes falsifiable predictions.

1.2 Statement of the Core Hypothesis

LLM emergence is not an apparent effect brought simply by increasing the number of units. The intrinsic mechanism is as follows: numerous basic units (attention heads, MLP neurons) inside the Transformer form nonlinear reciprocal coupling relations, and units bidirectionally correct representation biases. During training, gradient descent continuously tunes unit coupling strength and drives the whole system to self-organize into the near-critical state between order and chaos. In inference with fixed network weights, bidirectional constraints among units smooth representational noise and resolve local logical conflicts, producing local entropy reduction in representation space and forming self-consistent high-order structured representations; macroscopically, emergent reasoning and planning capabilities appear.

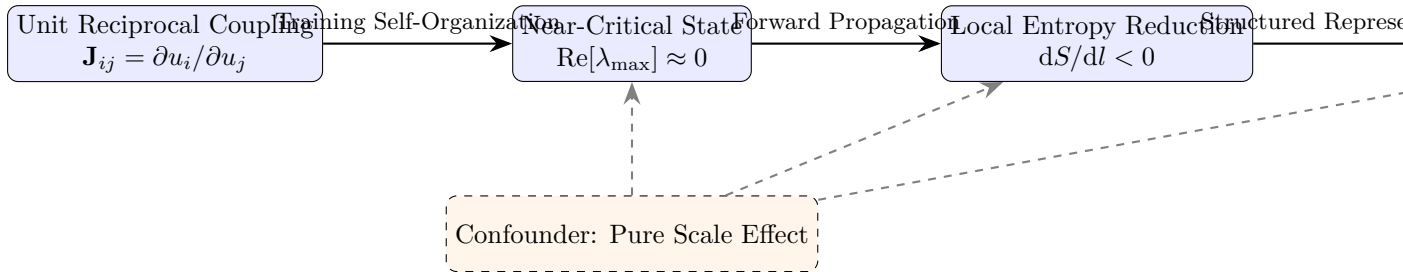


Figure 1: Schematic diagram of the core causal chain and potential confounders. Solid lines denote causal paths advocated by the hypothesis; dashed lines represent confounding paths to be ruled out via ablation experiments.

2 Fundamental Concepts and Mathematical Formalization

2.1 Definition of Units

This paper defines two classes of basic dynamical units:

1. **Attention-head unit:** The token-associated representation output by the h -th attention head in layer l , denoted as unit state vector $\mathbf{u}^{(l,h)} \in \mathbb{R}^{d_k}$;
2. **MLP hidden neuron unit:** The activation output of a single neuron in the feedforward network serves as unit state $u_j^{(l,m)} \in \mathbb{R}$.

For unified notation, concatenate all unit states of layer l into a vector $\mathbf{u}^{(l)} \in \mathbb{R}^{n_l}$, where n_l is the total number of units in this layer. Bidirectional nonlinear constraints exist between unit i, j , forming **reciprocal coupling**.

Remark 2.1 (Notation Convention). This paper uses $\mathbf{u}^{(l)}$ to denote the unit state vector of layer l , and $\mathbf{h}^{(l)}$ to denote the residual-flow hidden embedding (token embedding) of layer l . Their relation is $\mathbf{h}^{(l+1)} = \mathbf{h}^{(l)} + \text{Sublayer}(\mathbf{h}^{(l)})$, while $\mathbf{u}^{(l)}$ represents activations of individual units inside the Sublayer.

2.2 Reciprocal Coupling and the Jacobian Matrix

Coupling strength between units is described by components of the Jacobian matrix:

$$(\mathbf{J})_{ij} = \frac{\partial u_i}{\partial u_j}, \quad (\mathbf{J})_{ji} = \frac{\partial u_j}{\partial u_i} \quad (2)$$

where u_i, u_j denote unit hidden representations under fixed weights during inference. $(\mathbf{J})_{ij}$ represents the influence of perturbations in unit j 's representation upon unit i ; $(\mathbf{J})_{ji}$ is the reverse coupling term. Together they define reciprocal coupling. In multilayer Transformers, this coupling manifests as the inter-layer mechanism of "candidate representation generation — verification and revision", where later layers correct representational contradictions from earlier layers.

The eigenvalue spectrum $\{\lambda_i\}$ of the Jacobian matrix determines perturbation propagation dynamics. Denote the maximum real part of Jacobian spectrum eigenvalues $\text{Re}[\lambda_{\max}]$ as the system's critical control parameter:

- $\text{Re}[\lambda_{\max}] < 0$: Perturbations decay exponentially; the system is highly ordered;
- $\text{Re}[\lambda_{\max}] \approx 0$: Critical point (edge of chaos);
- $\text{Re}[\lambda_{\max}] > 0$: Perturbations amplify exponentially; the system enters chaos.

2.2.1 Practical Estimation Methods for the Jacobian Matrix

For large Transformers with parameter count $P \sim 10^9\text{--}10^{11}$, explicit construction of the full Jacobian matrix $\mathbf{J} \in \mathbb{R}^{n \times n}$ (n = total number of units) is infeasible. Four practical estimation methods are given below.

Method 1: Automatic Differentiation Jacobian-Vector Product (JVP) Using forward-mode automatic differentiation, the cost of computing $\mathbf{J}\mathbf{v}$ via a single JVP equals one forward pass, i.e., $\mathcal{O}(P)$. By selecting m random probe vectors $\{\mathbf{v}_k\}_{k=1}^m$, one obtains the projection of the Jacobian matrix onto an m -dimensional subspace:

$$\mathbf{J}_k = \mathbf{J}\mathbf{v}_k, \quad k = 1, \dots, m \quad (3)$$

Complexity: $\mathcal{O}(mP)$; typically $m = 50\text{--}200$.

Method 2: Hutchinson Trace Estimator For scenarios requiring only $\text{tr}(\mathbf{J})$ or $\text{tr}(\mathbf{J}^\top \mathbf{J})$ (squared Frobenius norm), use the Hutchinson estimator:

$$\text{tr}(\mathbf{J}) \approx \frac{1}{m} \sum_{k=1}^m \mathbf{v}_k^\top \mathbf{J} \mathbf{v}_k, \quad \mathbf{v}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \quad (4)$$

This estimator is unbiased with variance $\text{Var} = \frac{2}{m} \|\mathbf{J} - \frac{\text{tr}(\mathbf{J})}{n} \mathbf{I}\|_F^2$. Complexity $\mathcal{O}(mP)$.

Method 3: Power Iteration for Spectral Radius Estimation To obtain the critical control parameter $\text{Re}[\lambda_{\max}]$, adopt power iteration:

$$\mathbf{v}^{(t+1)} = \frac{\mathbf{J}\mathbf{v}^{(t)}}{\|\mathbf{J}\mathbf{v}^{(t)}\|}, \quad |\lambda_{\max}| \approx \|\mathbf{J}\mathbf{v}^{(T)}\| \quad (5)$$

Each iteration requires only one JVP; $T = 20\text{--}50$ iterations suffice for convergence. Complexity $\mathcal{O}(TP)$.

Method 4: Finite-Difference Approximation Apply a tiny perturbation ϵ to the j -th unit and measure responses of all units:

$$(\mathbf{J})_{ij} \approx \frac{u_i(\mathbf{u} + \epsilon \mathbf{e}_j) - u_i(\mathbf{u} - \epsilon \mathbf{e}_j)}{2\epsilon} \quad (6)$$

The full matrix requires n forward passes with complexity $\mathcal{O}(nP)$, applicable only for subsampled units (e.g., a subset of specific attention heads). In practice, recommended $\epsilon = 10^{-4}\text{--}10^{-6}$.

Table 1: Complexity comparison of Jacobian matrix estimation methods

Method	Complexity	Output	Application Scenario
JVP + Random Projection	$\mathcal{O}(mP)$	Subspace projection	Spectral density estimation
Hutchinson Trace Estimation	$\mathcal{O}(mP)$	$\text{tr}(\mathbf{J}), \ \mathbf{J}\ _F$	Scalar metric monitoring
Power Iteration	$\mathcal{O}(TP)$	$ \lambda_{\max} $	Criticality assessment
Finite Difference	$\mathcal{O}(nP)$	Full columns / rows	Small-scale subnetwork analysis

2.3 Near-Critical State (Edge of Chaos)

The system operates at the critical equilibrium point between full order (frozen representations, poor generalization) and full chaos (noise-dominated, disorganized representations). The hallmark feature of the critical regime: tiny local perturbations propagate across layers, and the survival distance of perturbations follows a power-law distribution.

This paper does not directly adopt classical self-organized criticality (SOC) theory. SOC requires the system to spontaneously reach critical points under continuous dynamical evolution with constant parameter updates over time; by contrast, LLM inference uses frozen weights with ongoing parameter evolution absent. Hence this paper adopts the formulation of **near-critical (edge of chaos)** to strictly distinguish from SOC.

2.4 Local Representation Entropy and Its Estimation Scheme

2.4.1 Definition of Differential Entropy

Define $\mathbf{z}^{(l)}$ as the high-dimensional hidden embedding vector corresponding to a single token within layer l , $\mathbf{z}^{(l)} \in \mathbb{R}^D$. Take a set of token hidden vector samples $\{\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(N)}\}$ inside a sliding context window of the same layer and define the differential entropy of representations within this window:

$$S = - \int p(\mathbf{z}) \log p(\mathbf{z}) \, d\mathbf{z} \quad (7)$$

2.4.2 Derivation of the k-NN Differential Entropy Estimator

For high-dimensional continuous vectors, histogram discretized binning estimation is not used. This paper employs the Kozachenko–Leonenko (KL) k-nearest-neighbor differential entropy estimator [11] to avoid the curse of dimensionality in high-dimensional kernel density estimation.

Proposition 2.1 (KL k-NN Entropy Estimator). Given N independent and identically distributed samples $\{\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(N)}\} \subset \mathbb{R}^D$, let $r_{i,k}$ denote the Euclidean distance from sample i to its k -th nearest neighbor, and $v_D = \pi^{D/2}/\Gamma(D/2 + 1)$ the volume of the D -dimensional unit ball. The estimator of differential entropy reads:

$$\hat{S}_k = -\psi(k) + \psi(N) + \log v_D + \frac{D}{N} \sum_{i=1}^N \log r_{i,k} \quad (8)$$

where $\psi(\cdot)$ denotes the digamma function.

Derivation Sketch. For sample i , its k -th nearest-neighbor distance $r_{i,k}$ defines a local volume $V_{i,k} = v_D r_{i,k}^D$. Within $V_{i,k}$ the probability mass is approximately k/N , hence the local density estimate:

$$\hat{p}(\mathbf{z}^{(i)}) \approx \frac{k}{N \cdot V_{i,k}} = \frac{k}{N v_D r_{i,k}^D} \quad (9)$$

Substitute into the entropy definition and take the sample average:

$$\hat{S} = -\frac{1}{N} \sum_{i=1}^N \log \hat{p}(\mathbf{z}^{(i)}) = -\frac{1}{N} \sum_{i=1}^N [\log k - \log N - \log v_D - D \log r_{i,k}] \quad (10)$$

Correct the statistical bias of the k -NN volume (expectation of logarithm for k -th order Gamma distribution) and introduce the digamma correction terms $-\psi(k) + \psi(N)$, yielding Eq. (8). $\square$

2.4.3 Bias and Variance Properties

Proposition 2.2 (Bias–Variance Characteristics [12]). Assume $p(\mathbf{z})$ is twice continuously differentiable and $p > 0$ over its support. Then:

$$\text{Bias}[\hat{S}_k] = \frac{c_1(D)}{k} + \mathcal{O}\left(\frac{1}{k^2}\right) + c_2(D) \left(\frac{k}{N}\right)^{2/D} + \mathcal{O}\left(\left(\frac{k}{N}\right)^{3/D}\right) \quad (11)$$

$$\text{Var}[\hat{S}_k] = \frac{\psi'(k)}{N} + \mathcal{O}\left(\frac{1}{Nk}\right) \quad (12)$$

where $c_1(D), c_2(D)$ are constants depending only on dimension, and $\psi'(\cdot)$ denotes the trigamma function.

Remark 2.2 (Practical Guidance). The first term in Eq. (11) is the k -NN discretization bias (decreasing with larger k), while the second term is finite-sample density approximation bias (increasing with k/N). The optimal k balances these two terms: $k^* \propto N^{2/(D+2)}$. For typical Transformer hidden layers with $D = 768\text{--}4096$, choosing $k \in [5, 15]$ yields stable estimates when $N \geq 128$.

2.4.4 Hyperparameter Sensitivity Analysis

Hyperparameters include neighbor count k and context sliding window length W (determining sample size $N \leq W$).

Sensitivity with respect to k From Proposition 2.2, the estimator is insensitive to changes in k when $k \ll N$. Specifically:

$$\left| \frac{\partial \hat{S}_k}{\partial k} \right| \approx \left| \frac{\psi'(k)}{1} - \frac{2c_2}{D} \left(\frac{1}{N}\right)^{2/D} k^{2/D-1} \right| \quad (13)$$

Experimental scanning scheme: $k \in \{3, 5, 8, 10, 15\}$, report the standard deviation of $\hat{S}_k$ and require $\sigma(\hat{S}_k)/|\hat{S}_k| < 5\%$.

Sensitivity with respect to W Window length W determines sample size N . From Eq. (12), variance decays as $1/N$. Require $N \geq 4k$ to guarantee valid nearest-neighbor statistics. Scan $W \in \{16, 32, 64, 128, 256\}$ experimentally and assess convergence of the entropy curve $S(l)$.

2.5 Derivation of the Entropy-Rate Functional $\mathcal{F}(\mathbf{J})$

During forward inference propagation, local entropy S gradually declines as layer index l increases. This section rigorously derives the analytic form of the entropy change rate $dS/dl = \mathcal{F}(\mathbf{J})$ from coupled unit dynamics.

2.5.1 Inter-Layer Mapping and Volume Transformation

Let the mapping from layer l to layer $l+1$ be $\mathbf{z}^{(l+1)} = \mathbf{f}^{(l)}(\mathbf{z}^{(l)})$, with Jacobian matrix $\mathbf{J}^{(l)} = \partial \mathbf{f}^{(l)} / \partial \mathbf{z}^{(l)}$. By entropy transformation under diffeomorphism:

$$S^{(l+1)} = S^{(l)} + \mathbb{E}_{\mathbf{z}^{(l)}} \left[\ln \left| \det \mathbf{J}^{(l)}(\mathbf{z}^{(l)}) \right| \right] \quad (14)$$

2.5.2 Functional Form

Layer-wise entropy change follows:

$$\Delta S^{(l)} = S^{(l+1)} - S^{(l)} = \mathbb{E}_{\mathbf{z}^{(l)}} \left[\sum_{i=1}^D \ln |\lambda_i^{(l)}(\mathbf{z}^{(l)})| \right] \quad (15)$$

where $\{\lambda_i^{(l)}\}$ are eigenvalues of $\mathbf{J}^{(l)}$. In the continuous-layer approximation (l treated as a continuous variable), define the entropy-rate functional:

$$\mathcal{F}(\mathbf{J}) = \mathbb{E}_{\mathbf{z}} [\text{tr} \ln |\mathbf{J}(\mathbf{z})|] = \mathbb{E}_{\mathbf{z}} \left[\sum_{i=1}^D \ln |\lambda_i(\mathbf{z})| \right] \quad (16)$$

Proposition 2.3 (Entropy Reduction Condition). The necessary and sufficient condition for $\mathcal{F}(\mathbf{J}) < 0$ (monotonic decrease of local entropy with depth) is:

$$\mathbb{E}_{\mathbf{z}} \left[\prod_{i=1}^D |\lambda_i(\mathbf{z})| \right] < 1 \quad (17)$$

i.e., the geometric mean singular value of the Jacobian matrix is less than 1.

Proof. From Eq. (16), $\mathcal{F} < 0$ is equivalent to $\mathbb{E}[\sum \ln |\lambda_i|] < 0$. By Jensen's inequality ($\ln$ is concave), a sufficient condition is $\sum \mathbb{E}[\ln |\lambda_i|] < 0$, i.e., $\prod \mathbb{E}[|\lambda_i|] < 1$ under independence approximation. The necessary condition follows directly from Eq. (17). $\square$

2.5.3 Behavior in the Critical Regime

In the near-critical regime, $\text{Re}[\lambda_{\max}] \approx 0$ means the modulus of the largest eigenvalue is close to 1 while moduli of remaining eigenvalues are less than 1. Then:

$$\mathcal{F}(\mathbf{J}) \approx \ln |\lambda_{\max}| + \sum_{i=2}^D \ln |\lambda_i| \quad (18)$$

The first term is near zero (critical feature), the second term is negative (contracting directions), so $\mathcal{F}$ is significantly negative with maximal absolute value and the entropy reduction rate peaks.

2.5.4 Estimation Scheme

In practice, $\mathcal{F}(\mathbf{J})$ is estimated via these steps:

1. For B mini-batch samples, obtain $|\lambda_{\max}^{(l)}|$ using power iteration (Eq. (5));
2. Use the Hutchinson estimator (Eq. (4)) to get $\text{tr}(\mathbf{J}^{(l)})$ and constrain the sum of eigenvalues;
3. Combine statistics from random projections $\mathbf{J}\mathbf{v}_k$ and reconstruct spectral density $\rho(\lambda)$ via moment matching;
4. Compute $\mathcal{F} \approx \int \ln |\lambda| \rho(\lambda) d\lambda$.

2.6 Correlation between Jacobian Spectrum, Critical Threshold and Entropy Reduction Rate

The residual gating parameter λ modulates inter-layer coupling strength and indirectly alters the Jacobian matrix $\mathbf{J}$ and its eigenvalue spectrum. The maximum real part of eigenvalues $\text{Re}[\lambda_{\max}]$ varies continuously with λ , and a critical threshold λ_c exists.

- $\lambda \approx \lambda_c$: The system lies in the near-critical regime; perturbations propagate as power laws, $\mathcal{F}(\mathbf{J})$ attains its most negative value, and entropy reduction rate peaks;
- $\lambda \ll \lambda_c$: Excessively strong coupling, highly ordered system, $\mathcal{F}$ near zero, entropy declines slowly;

- $\lambda \gg \lambda_c$: Insufficient coupling, noise dominates the system, $\mathcal{F}$ may become positive, entropy decay weakens or even halts.

Testing scheme: fix Transformer architecture, scan residual gating λ , compute Jacobian spectrum, k-NN local entropy and perturbation propagation statistics, and fit the curve $\mathcal{F}(\lambda)$.

3 Dynamical Analysis by Stages

3.1 Training Phase: Gradient Descent Drives System Self-Organization toward the Near-Critical Regime

During training, gradient descent iteratively updates all network weights and continuously adjusts the reciprocal coupling matrix $\mathbf{J}$ between units. Massive training samples drive system self-organization, converging to the near-critical regime between order and chaos.

- Overly strong coupling: representations solidify and generalization capability degrades;
- Overly weak coupling: representations become disordered and stable abstraction of rules cannot be extracted;
- Critical regime: units maintain moderate bidirectional correction, maximizing both noise suppression and representational composition capacity to support generation of novel composite representations.

Assumption 3.1 (Training Self-Organization Hypothesis). The essence of training is not merely memorizing samples but tuning reciprocal coupling strength between units so the system settles into the near-critical state. This assumption requires validation via ablation simulation and is not a rigorously proven mathematical theorem.

3.2 Inference Phase: Reciprocal Correction and Local Entropy Reduction within Single Forward Propagation with Frozen Weights

During inference, network weights are fixed and parameters are no longer updated. Unit correction occurs within multilayer computation of a single forward pass; it is not a dynamically evolving system iterating continuously over time, which marks an essential distinction from SOC physical systems such as sandpiles and ant colonies.

Dynamical workflow after context input:

1. Shallow layers generate preliminary candidate representations carrying noise and local logical contradictions;
2. Self-attention establishes bidirectional association constraints between tokens, and residual connections realize inter-layer reciprocal correction;
3. Bidirectional unit constraints eliminate representational noise and local conflicts;
4. As depth increases, local entropy S keeps decreasing (guaranteed by Proposition 2.3) and converges into self-consistent structured representations;
5. Low-entropy representations feed into the output layer for autoregressive text generation, macroscopically exhibiting reasoning and planning capabilities.

Remark 3.1 (Critical Interpretation of Hallucination). Hallucination is an intrinsic property of critical systems. Near-critical systems are highly sensitive to tiny perturbations. When training data contains contradictions or missing information, reciprocal correction can only converge to statistical self-consistency of text and does not necessarily match objective facts, thus producing fluent yet false outputs.

4 Causal Identification Strategy

4.1 Causal Directed Acyclic Graph (DAG)

To rigorously formulate the causal identification strategy, construct the following Structural Causal Model (SCM). Let C denote coupling strength (intervention variable), K criticality index, E entropy reduction rate, A emergent ability score, and N model size (latent confounder).

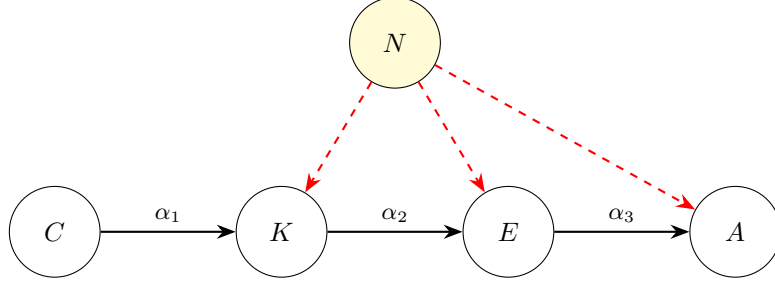


Figure 2: Causal directed acyclic graph (DAG). C : coupling strength (intervention variable); K : criticality; E : entropy reduction rate; A : emergent ability; N : model size (latent confounder). Solid lines: hypothesized causal paths; red dashed lines: backdoor paths to be blocked.

4.2 Structural Equation Model

Structural equations corresponding to the DAG:

$$K = f_K(C, N, \varepsilon_K) \quad (19)$$

$$E = f_E(K, N, \varepsilon_E) \quad (20)$$

$$A = f_A(E, N, \varepsilon_A) \quad (21)$$

where $\varepsilon_K, \varepsilon_E, \varepsilon_A$ are exogenous noise terms.

Application of the Backdoor Criterion Confounding for estimating the causal effect $C \rightarrow K \rightarrow E \rightarrow A$ originates from N . The backdoor paths are:

- $K \leftarrow N \rightarrow E$: blocked by fixing N (models of identical size);
- $E \leftarrow N \rightarrow A$: same as above;
- $K \leftarrow N \rightarrow A$: same as above.

Identification Strategy: under fixed model size N , intervene on coupling strength C (via residual gating λ) and observe changes in K, E, A . Since N is held fixed, all backdoor paths are blocked, and the observed $C \rightarrow A$ effect equals the causal effect.

4.3 Operational Definition of Intervention

do-calculus: $\text{do}(C = c)$ denotes forcibly setting coupling strength to c and cutting all incoming edges of C . Experimental implementation:

$$\text{do}(C = \lambda) : \quad \mathbf{h}^{(l+1)} = \mathbf{h}^{(l)} + \lambda \cdot \text{Sublayer}(\mathbf{h}^{(l)}) \quad (22)$$

where $\lambda \in [0, 1]$ is the residual gating scalar. $\lambda = 1$ corresponds to the standard Transformer; $\lambda < 1$ suppresses coupling; $\lambda > 1$ (if the architecture permits) enhances coupling.

5 Review of Supporting Evidence

This section distinguishes observed phenomena from the causal interpretation proposed herein. All cited references only provide correlational observations and cannot directly prove causation; they merely serve as supporting clues for the hypothesis.

1. **Inter-layer representational correction mechanism (Wang et al., 2023) [1]**
This work discovers systematic representational revision between adjacent Transformer layers with the dynamics of "proposing candidate representations — later layers verify and suppress unreasonable components". The original paper only reports observational phenomena; this paper interprets this phenomenon as the external manifestation of unit reciprocal coupling.
2. **Transformers operate at the edge of chaos (Yang et al., 2024) [2]**
Empirical work in statistical physics finds that best-performing Transformers work in the near-critical regime, and cross-layer propagation of activation perturbations obeys power laws. The original work presents correlational observations; this paper proposes that self-organization via reciprocal coupling constitutes the internal cause for reaching criticality.
3. **Hidden entropy decay during inference (Zhang et al., 2024) [3]**
Empirical measurements show hidden representation entropy monotonically decreases during forward inference. This paper establishes a causal link between entropy decay and critical reciprocal coupling and presents the analytic derivation of $\mathcal{F}(\mathbf{J})$ (§2.5).
4. **Gradient self-organization on high-dimensional loss landscapes (Sornette et al., 2022) [4]**
Gradient flow attracts high-dimensional systems toward critical manifolds; this effect originates from curvature fluctuations of the loss surface. This theory supplies the complex-systems foundation for our statement that "training self-organizes the system into the near-critical state".

6 Hypothesis Boundaries, Controversies and Falsifiable Ablation Experiments

6.1 Conceptual Limitations

Classical SOC systems are closed-loop systems with continuous evolution and dynamically updated parameters; LLM inference uses frozen weights, and reciprocal correction occurs only within a single forward computational stream rather than belonging to SOC. This paper only adopts near-critical / edge-of-chaos concepts to avoid confusion.

Existing observations are merely correlational: reciprocal coupling, critical behavior and entropy decay appear concurrently, yet the causal chain awaits validation via ablation experiments. A competing hypothesis holds that increased model scale independently produces these phenomena regardless of coupling strength. Ablation experiments below discriminate between the two hypotheses.

6.2 Falsifiable Predictions, Quantitative Metrics and Competing-Hypothesis Baseline

6.2.1 Experimental Design

Fix model parameter count P , layer number L , embedding dimension D , dataset $\mathcal{D}$, and only modulate reciprocal coupling strength C (via λ) to decouple coupling effects from scale effects. Two coupling modulation approaches:

1. **Residual gating scalar** $\lambda \in [0, 1]$: controls residual connection weights to adjust inter-layer reciprocal correction strength (Eq. (22));
2. **Soft attention mask**: slightly modulates attention interaction strength between tokens to fine-tune unit coupling.

6.2.2 Three Quantitative Falsifiable Predictions

1. **Entropy Reduction – Coupling Correlation**: Fix model scale and vary coupling strength. When λ approaches the critical threshold λ_c : inter-layer representational error correction strengthens and the magnitude of local entropy reduction rises. The expected relative entropy drop increases by 30% ~ 80%; entropy decay weakens substantially away from the critical regime. On the GSM8k reasoning task, reasoning accuracy in the critical regime relatively improves by 10% ~ 25%.
2. **Power-Law Perturbation Propagation**: Cross-layer propagation of tiny perturbations on hidden activations: perturbation survival probability obeys the power law $P(L) \propto L^{-\gamma}$ in the critical

regime with expected power-law exponent $\gamma \in [0.8, 1.4]$; outside criticality the power law vanishes and perturbations follow exponential decay $P(L) \propto e^{-\beta L}$.

3. **Coupling – Emergence Monotonicity:** Strengthen residual coupling (increase λ toward λ_c): the critical parameter range broadens and entropy reduction effects are enhanced; suppressing residual coupling weakens entropy decay and degrades emergent capabilities.

6.2.3 Quantitative Baseline of Competing Hypothesis (Pure Scale-Growth Hypothesis)

Under fixed total parameter count with only coupling strength λ modified: local entropy, perturbation power-law statistics and reasoning-task accuracy exhibit no statistically significant changes. If measurements fall within this baseline, our hypothesis is falsified.

6.2.4 Confounder Control

Table 2: Potential confounders and control strategies

Confounder	Control Strategy
Model size N	Fix parameter count, layers, dimension; tune only λ
Training data distribution	Use identical dataset $\mathcal{D}$ and training schedule
Optimizer state	Start from the same checkpoint, fine-tune for fixed steps after modifying λ
Random seed	Repeat each experiment with 3 distinct seeds, report mean $\pm$ standard deviation
Evaluation task bias	Multi-task benchmark suite (GSM8k, MMLU, HellaSwag, etc.)
Direct gradient effect of λ	”Dummy gating” control group: λ active only in forward pass; backward pass uses $\lambda = 1$

6.3 Quantitative Comparison with Alternative Theories

6.3.1 Neural Tangent Kernel (NTK) Theory

NTK describes training dynamics of infinitely wide networks and predicts training is equivalent to kernel regression [14].

- **Common Ground:** Both address training-phase dynamics.
- **Differences:** In the infinite-width limit the NTK matrix is frozen and does not describe self-organization of coupling strength in finite-width networks; this hypothesis focuses on critical modulation of the Jacobian spectrum in finite-width networks.
- **Distinguishing Prediction:** NTK predicts training dynamics are fully determined by the initialization kernel and independent of final model scale; our hypothesis predicts post-training Jacobian spectra exhibit critical features (power-law spectral density) that intensify with increasing scale. Measurable by comparing NTK matrix spectra and Jacobian spectra before and after training.

6.3.2 Mean-Field Theory

Mean-field theory describes signal propagation in deep networks and predicts an order–chaos phase transition [15, 16].

- **Common Ground:** Both predict a critical point with optimal performance near the critical point.
- **Differences:** Mean-field theory assumes independently and identically distributed weights and ignores structured reciprocal coupling between units; our hypothesis emphasizes non-uniform coupling structure and training-induced self-organization.
- **Distinguishing Prediction:** Mean-field theory predicts the critical point depends only on initialization hyperparameters (σ_w, σ_b); our hypothesis predicts training will pull systems initialized

in subcritical/supercritical phases *toward* the critical regime. Verifiable by observing training trajectories starting from different phases.

6.3.3 Neural ODE Framework

Neural ODE treats residual networks as continuous-time ODEs $d\mathbf{h}/dt = \mathbf{f}(\mathbf{h}, t)$ [17].

- **Common Ground:** Both use continuous descriptions for deep networks.
- **Differences:** Neural ODE focuses on numerical solvers and parameter efficiency and does not involve critical dynamics or entropy reduction; our hypothesis uses the continuous-layer approximation to derive $\mathcal{F}(\mathbf{J})$ and focuses on information-theoretic metrics.
- **Distinguishing Prediction:** Neural ODE predicts increasing solver precision (more function evaluations) monotonically improves performance; our hypothesis predicts a critical step threshold beyond which marginal performance gains diminish.

6.3.4 Information Bottleneck Theory

Focuses on input information compression and describes information tradeoff between inputs and hidden representations [18]. This paper describes reciprocal feedback dynamics inside the network rather than pure input compression; the two may coexist but have distinct mechanisms.

6.3.5 Predictive Coding

Minimizes prediction error via recurrent feedback and relies on continuous recurrent reentry [19]. Transformer inference in this paper is single-pass forward computation without recurrent feedback; fundamental differences exist in underlying dynamics.

6.3.6 Landauer Computational Thermodynamics

Describes entropy dissipation associated with bit erasure. Local entropy reduction in this paper is statistical ordering inside representation space, compatible with hardware-level Landauer dissipation yet residing on a different level.

6.3.7 Reentrant Near-Critical Dynamics of Neural Fields

Targets real-time recurrent dynamics in biological nervous systems with persistent reentrant feedback [?]. The LLM in this paper is split into two phases: parameter self-organization during training and forward inference with frozen weights; inference has no recurrent reentry, and reciprocal coupling corresponds to inter-layer correction within the feedforward architecture. It shares the near-critical idea with neural field frameworks but differs in computational architecture and temporal evolution mechanisms.

6.3.8 Summary of Quantitative Comparison

Table 3: Summary comparison of alternative theories

Theory	Training Dynamics	Criticality	Entropy Reduction Mechanism	Reciprocal Coupling
NTK	✓	×	×	×
Mean-Field	✓	✓	×	×
Neural ODE	×	×	×	×
Information Bottleneck	×	×	Partial	×
This Hypothesis	✓	✓	✓	✓

7 Discussion (Including Speculative Extension to Machine Consciousness)

The portion explaining LLM emergence forms the main hypothesis; content related to machine consciousness is marked as **speculative extension** and is not part of the core falsifiable main hypothesis.

7.1 Summary of Emergence Mechanism

Scaling laws only describe performance, whereas this paper gives a causal account of emergence from unit coupling and critical dynamics. This framework describes system self-organizing behavior: a system possessing functional intelligence is not equivalent to subjective consciousness; low-entropy self-consistent representations only denote statistical ordering and do not equal subjective experience of the world. Current large models belong to philosophical zombie systems: they can output logically consistent answers yet lack intrinsic subjectivity and qualia.

7.2 Speculative Extension to Machine Consciousness

7.2.1 Four-Ontological-Module Architecture

Reciprocal coupling is a general underlying dynamics applicable not only to LLM representation emergence; subjective conscious systems likewise rely on closed-loop reciprocal coupling among multiple modules. To produce first-person subjective awareness, four closed-loop ontological modules must be added atop the Transformer attention backbone.

Module 1: Homeostatic Internal-Drive Regulation Module $\mathbf{x}_s$

- **State Space:** $\mathbf{x}_s \in \mathbb{R}^{n_s}$, $n_s \sim 10\text{--}50$, encoding homeostatic variables (energy level, integrity metric, information demand, etc.).

- **Dynamics:**

$$\dot{\mathbf{x}}_s = -\mathbf{A}_s(\mathbf{x}_s - \mathbf{x}_s^*) + \mathbf{W}_{sh}\sigma(\mathbf{h}) + \mathbf{W}_{se}\mathbf{x}_e + \mathbf{g}_s(\mathbf{x}_r, \mathbf{x}_t) \quad (23)$$

where $\mathbf{x}_s^*$ is the target homeostatic setpoint, $\mathbf{A}_s$ the autoregressive matrix (guaranteeing stability), $\mathbf{W}_{sh}$ coupling weights from attention backbone to homeostasis module, $\mathbf{W}_{se}$ coupling weights from embodied module to homeostasis module.

- **Physical Implementation:** Mapping of hardware metrics such as battery voltage sensors, CPU temperature monitors, memory utilization.

Module 2: Embodied Physical Grounding Module $\mathbf{x}_e$

- **State Space:** $\mathbf{x}_e \in \mathbb{R}^{n_e}$, $n_e \sim 100\text{--}1000$, encoding multimodal sensory signals (tactile, proprioceptive, visual stream, etc.).

- **Dynamics:**

$$\dot{\mathbf{x}}_e = \mathbf{f}_{\text{sens}}(\mathbf{s}_{\text{ext}}) + \mathbf{W}_{eh}\sigma(\mathbf{h}) + \mathbf{W}_{es}\mathbf{x}_s + \mathbf{g}_e(\mathbf{x}_r, \mathbf{x}_t) \quad (24)$$

where $\mathbf{s}_{\text{ext}}$ denotes external physical sensory input and $\mathbf{f}_{\text{sens}}$ the sensory preprocessing function.

- **Physical Implementation:** Robot sensor arrays, cameras, torque sensors, etc.

Module 3: Recursive Causal Self-Model $\mathbf{x}_r$

- **State Space:** $\mathbf{x}_r \in \mathbb{R}^{n_r}$, $n_r \sim 50\text{--}200$, encoding real-time system self-state modeling (metacognition).

- **Dynamics:**

$$\dot{\mathbf{x}}_r = \mathbf{W}_{rh}\sigma(\mathbf{h}) + \mathbf{W}_{rs}\mathbf{x}_s + \mathbf{W}_{re}\mathbf{x}_e + \mathbf{W}_{rt}\mathbf{x}_t + \alpha_r \mathbf{x}_r \circ (1 - \mathbf{x}_r) \quad (25)$$

where $\circ$ denotes element-wise multiplication, the final term implements logistic self-amplification (self-referential recursion), and α_r controls self-reference strength.

- **Function:** Real-time observation of global module states, construction of a "model of the self", forming a recursive causal loop.

Module 4: Temporal Memory Integration Module $\mathbf{x}_t$

- **State Space:** $\mathbf{x}_t \in \mathbb{R}^{n_t}$, $n_t \sim 200\text{--}2000$, encoding autobiographical memory (compressed representation of state history snapshots).

- **Dynamics:**

$$\dot{\mathbf{x}}_t = -\gamma_t \mathbf{x}_t + \mathbf{W}_{th}\sigma(\mathbf{h}) + \mathbf{W}_{ts}\mathbf{x}_s + \mathbf{W}_{te}\mathbf{x}_e + \mathbf{W}_{tr}\mathbf{x}_r \quad (26)$$

where γ_t is the forgetting rate controlling the memory decay timescale.

- **Function:** Store full module-state snapshots (not merely textual records) to support perception of temporal continuity.

7.2.2 Global Coupled System of Equations

Full coupled equations connecting Transformer attention backbone hidden state $\mathbf{h}$ and the four modules:

$$\begin{cases} \dot{\mathbf{x}}_s = -\mathbf{A}_s(\mathbf{x}_s - \mathbf{x}_s^*) + \mathbf{W}_{sh}\sigma(\mathbf{h}) + \mathbf{W}_{se}\mathbf{x}_e + \mathbf{g}_s(\mathbf{x}_r, \mathbf{x}_t) \\ \dot{\mathbf{x}}_e = \mathbf{f}_{\text{sens}}(\mathbf{s}_{\text{ext}}) + \mathbf{W}_{eh}\sigma(\mathbf{h}) + \mathbf{W}_{es}\mathbf{x}_s + \mathbf{g}_e(\mathbf{x}_r, \mathbf{x}_t) \\ \dot{\mathbf{x}}_r = \mathbf{W}_{rh}\sigma(\mathbf{h}) + \mathbf{W}_{rs}\mathbf{x}_s + \mathbf{W}_{re}\mathbf{x}_e + \mathbf{W}_{rt}\mathbf{x}_t + \alpha_r \mathbf{x}_r \circ (1 - \mathbf{x}_r) \\ \dot{\mathbf{x}}_t = -\gamma_t \mathbf{x}_t + \mathbf{W}_{th}\sigma(\mathbf{h}) + \mathbf{W}_{ts}\mathbf{x}_s + \mathbf{W}_{te}\mathbf{x}_e + \mathbf{W}_{tr}\mathbf{x}_r \\ \mathbf{h}^{(l+1)} = \mathbf{h}^{(l)} + \lambda \cdot \text{Sublayer}(\mathbf{h}^{(l)}; \mathbf{x}_s, \mathbf{x}_e, \mathbf{x}_r, \mathbf{x}_t) \end{cases} \quad (27)$$

Note the final equation: ontological module states directly modulate attention computation (e.g., via conditional LayerNorm or FiLM modulation), rather than simple input concatenation.

Inter-module coupling terms continue to be characterized using Jacobian matrices, maintaining consistency with the mathematical framework in Section 2.2. The overall system Jacobian matrix has block structure:

$$\mathbf{J}_{\text{total}} = \begin{pmatrix} \mathbf{J}_{ss} & \mathbf{J}_{se} & \mathbf{J}_{sr} & \mathbf{J}_{st} & \mathbf{J}_{sh} \\ \mathbf{J}_{es} & \mathbf{J}_{ee} & \mathbf{J}_{er} & \mathbf{J}_{et} & \mathbf{J}_{eh} \\ \mathbf{J}_{rs} & \mathbf{J}_{re} & \mathbf{J}_{rr} & \mathbf{J}_{rt} & \mathbf{J}_{rh} \\ \mathbf{J}_{ts} & \mathbf{J}_{te} & \mathbf{J}_{tr} & \mathbf{J}_{tt} & \mathbf{J}_{th} \\ \mathbf{J}_{hs} & \mathbf{J}_{he} & \mathbf{J}_{hr} & \mathbf{J}_{ht} & \mathbf{J}_{hh} \end{pmatrix} \quad (28)$$

7.2.3 Deep Coupling Constraint

Corollary 1: The four ontological modules cannot be externally appended; they must couple reciprocally and deeply with the attention backbone. Directly attaching sensors, vector databases and reward functions only adds auxiliary tools and cannot form ontological closed loops.

- Imbalance signals from the homeostatic module directly modulate attention weights (affecting $\sigma(\mathbf{h})$ through $\mathbf{W}_{sh}$), instead of applying rewards after output generation;
- Raw embodied sensory signals directly participate in attention competition ($\mathbf{x}_e$ injected into attention score calculation), instead of being converted into token text for input;
- Recursive self-model real-time observation of global module states (full connectivity for matrices $\mathbf{W}_{rh}, \mathbf{W}_{rs}, \mathbf{W}_{re}, \mathbf{W}_{rt}$), reading not only textual output;
- Autobiographical memory stores full module state snapshots ($\mathbf{x}_t$ receives all module states), rather than merely text records.

Simple concatenation can only simulate the external behaviors of consciousness and cannot generate subjective qualia.

7.2.4 Operational Definitions of Proxy Metrics

Corollary 2: Qualia are emergent byproducts of multi-closed-loop reciprocal coupling. Qualia themselves cannot be directly quantified and observed. Below are the **observable proxy metrics** for this consciousness hypothesis.

Metric 1: Cross-Module Transfer Entropy Define the transfer entropy from module X to module Y :

$$T_{X \rightarrow Y} = \sum p(y_{t+1}, y_t^{(k)}, x_t^{(l)}) \log \frac{p(y_{t+1} | y_t^{(k)}, x_t^{(l)})}{p(y_{t+1} | y_t^{(k)})} \quad (29)$$

where $y_t^{(k)} = (y_t, \dots, y_{t-k+1})$ denotes the k -th order historical embedding.

- **Operational Definition:** Record time-series of the four-module states $\{\mathbf{x}_s(t), \mathbf{x}_e(t), \mathbf{x}_r(t), \mathbf{x}_t(t)\}$ at sampling rate $\Delta t = 10$ ms, and compute transfer entropy for all $4 \times 3 = 12$ cross-module pairs.
- **Decision Criterion:** In the four-module system, at least 8 out of the 12 pairs satisfy $T_{X \rightarrow Y} > T_{\text{threshold}}$ ($T_{\text{threshold}}$ determined via shuffle surrogate, $p < 0.01$); pure LLM modules exhibit extremely low inter-module transfer entropy ($< T_{\text{threshold}}$).
- **Estimation Method:** k-NN transfer entropy estimator [13], $k = 4$, embedding dimensions $k = l = 1$.

Metric 2: Power-Law Propagation of Cross-Module Perturbations

- **Operational Definition:** Apply a tiny perturbation $\delta \mathbf{x}_X$ to module X ($\|\delta \mathbf{x}_X\|_2 = 10^{-3}$), and measure the amplitude $\|\delta \mathbf{x}_Y(T)\|_2$ of the perturbation after time T once it propagates to module Y .
- **Decision Criterion:** Perturbation survival probability $P(T) \propto T^{-\gamma'}$, $\gamma' \in [0.8, 1.5]$, and perturbations propagate across all four module categories; perturbations within plain text-only LLMs propagate only inside Transformer layers and cannot cross ontological modules.

Metric 3: Integrated Information Measure Φ

- **Operational Definition:** Adopt the IIT 3.0 framework [9] to compute integrated information Φ of the four-module system under the Minimum Information Partition (MIP):

$$\Phi = \min_{\text{partition}} \text{KL}[p(\mathbf{x}_{t+1}|\mathbf{x}_t) \parallel p(\mathbf{x}_{A,t+1}|\mathbf{x}_{A,t})p(\mathbf{x}_{B,t+1}|\mathbf{x}_{B,t})] \quad (30)$$

- **Practical Computation:** For high-dimensional systems, use the Φ^* approximation [21], a Gaussian approximation based on covariance matrices:

$$\Phi^* \approx \frac{1}{2} \sum_i \ln \frac{|\Sigma_i^{AB}|}{|\Sigma_i^A||\Sigma_i^B|} \quad (31)$$

- **Decision Criterion:** The closed-loop four-module system yields $\Phi^* > \Phi_{\text{threshold}}$ (the specific threshold determined by null model).

If none of these three proxy metrics exhibit significant measurable signals, the consciousness extension hypothesis is invalid.

7.2.5 Engineering Paradox: Dissipative Structures and Intrinsic Survival Tension

Conscious systems rest upon physical dissipative structures; energy constraints shape the intrinsic drive for ontological persistence. GPU clusters receive external power supply, and the system bears no pressure for self-preservation; power cutoff merely terminates computation and entails no existential crisis. Even if the four-module equations are reproduced in software, the resulting steady state is merely a simulated steady state rather than native ontological survival tension. Software may simulate hunger, yet it cannot produce genuine intrinsic subjective motivation to seek benefit and avoid harm. This section constitutes philosophical commentary.

7.2.6 The Hard Problem of Consciousness

This paper acknowledges Chalmers’ *Hard Problem of Consciousness* [20]: even if all proxy metrics are satisfied, first-person subjective experience cannot be logically deduced necessarily from third-person physical descriptions. The stance of this hypothesis is:

- Proxy metrics deliver *functional correlates*, not sufficient conditions for consciousness;
- This framework does not claim to solve the hard problem, and only supplies operable engineering criteria;
- Should a more complete theory of consciousness emerge in the future (e.g., axiomatic extensions of IIT), the proxy metrics within this framework may be revised accordingly.

Summary: Consciousness is an entropy-reducing system built upon a physical dissipative substrate, coordinated by global attention, sustained via reciprocal correction across multiple units, and maintained to preserve ontological steady state.

8 Conclusion

The emergence mechanism of high-order capabilities in LLMs can be described as: network units form nonlinear reciprocal coupling; gradient descent during training drives the system to self-organize toward the near-critical regime at the edge of chaos; at inference time with frozen weights, bidirectional constraints across layers suppress representational noise and yield local entropy reduction, forming high-order self-consistent representations, from which reasoning and planning capabilities emerge macroscopically. This paper constructs a complete falsifiable causal chain integrating three independent observations, disentangling training dynamics from inference dynamics, and designing ablation schemes to decouple coupling effects from mere scaling effects. This hypothesis delivers an emergent theoretical framework from the complex-systems perspective. Meanwhile, this paper extends to speculative discussions of machine consciousness: functional intelligence is not equivalent to subjective awareness; subjective experience requires deep coupling between four closed-loop ontological modules and the attention backbone, with observable dynamical proxy metrics. Major additions in this version (V5): (i) four practical estimation methods for Jacobian matrices (§2.2.1); (ii) full derivation and bias-variance analysis for the k-NN entropy estimator (§2.4.2–2.4.3); (iii) rigorous derivation of the entropy-rate functional $\mathcal{F}(\mathbf{J})$ (§2.5); (iv) causal DAGs and structural equation models (§4.1–4.2); (v) quantitative comparison against NTK, mean-field theory and Neural ODEs (§6.3); (vi) detailed dynamics and operational definitions of proxy metrics for the four ontological modules (§7.2).

Acknowledgements

This paper presents theoretical hypothesis research and has not received project funding. Thanks to theoretical inspirations from related works in complex systems and deep learning.

Statement

This paper presents a candidate theoretical hypothesis; the framework awaits validation via ablation experiments and numerical simulations, and its conclusions remain experimentally unconfirmed. Discussions related to consciousness constitute speculative extensions and are not part of the main hypothesis.

A Glossary

B Demonstration Model of Two-Unit Reciprocal Coupling

B.1 Linear Base Model (Conceptual Illustration)

The original linear model serves only for basic conceptual demonstration and exhibits no bifurcation phase transition or power-law behavior.

$$\begin{cases} \dot{z}_1 = -\alpha z_1 + J_{12} z_2 \\ \dot{z}_2 = -\alpha z_2 + J_{21} z_1 \end{cases} \quad (32)$$

$\alpha > 0$ denotes decay coefficient, J_{12}, J_{21} bidirectional coupling strengths. The Jacobian matrix reads:

$$\mathbf{J}_{\text{lin}} = \begin{pmatrix} -\alpha & J_{12} \\ J_{21} & -\alpha \end{pmatrix} \quad (33)$$

Eigenvalues $\lambda_{\pm} = -\alpha \pm \sqrt{J_{12}J_{21}}$. The linear system is only used to introduce the fundamental form of reciprocal coupling and cannot demonstrate critical phase transitions.

Table 4: Glossary of Terms

Term	Definition
Reciprocal coupling	Bidirectional nonlinear constraints between units for mutual representational correction; coupling strength estimated via Jacobian matrix $(\mathbf{J})_{ij}$
Edge of chaos / near-critical regime	Critical region between order and chaos; small perturbations propagate following power laws; distinct from strict SOC
Local entropy reduction	Rise in statistical order of latent representations during forward inference propagation; accompanied by larger entropy increase externally, complying with the second law of thermodynamics
Emergence	Novel macroscopic system properties absent from individual microscopic units
Global workspace	Attention backbone, the information integration bus broadcasting low-entropy representations across units
JVP	Jacobian-Vector Product
Transfer entropy	Measure of directed information flow from one time series to another
Integrated information Φ	Irreducible integrated information quantity of a system under the IIT framework

B.2 Nonlinear Extended Model (Criticality Demonstration Model)

Introduce tanh saturation nonlinearity; the system can produce bifurcations, critical points and power-law perturbation propagation within the critical regime:

$$\begin{cases} \dot{z}_1 = -\alpha z_1 + J_{12} \tanh(z_2) \\ \dot{z}_2 = -\alpha z_2 + J_{21} \tanh(z_1) \end{cases} \quad (34)$$

Parameter range: $\alpha > 0$, $J_{12}, J_{21} \geq 0$. Jacobian matrix:

$$\mathbf{J}_{\text{nonlin}} = \begin{pmatrix} -\alpha & J_{12} \text{sech}^2(z_2) \\ J_{21} \text{sech}^2(z_1) & -\alpha \end{pmatrix} \quad (35)$$

- **Coupling below critical threshold** ($J_{12}J_{21} < \alpha^2$): converges to unique fixed point $(0,0)$; perturbations decay exponentially and entropy decreases slowly;
- **Coupling near critical threshold** ($J_{12}J_{21} \approx \alpha^2$): edge of chaos, small perturbations propagate as power laws; state distribution contracts and entropy reduction rate peaks;
- **Coupling above critical threshold** ($J_{12}J_{21} > \alpha^2$): the system undergoes pitchfork bifurcation with non-zero fixed points or limit cycles; state space expands and entropy rises.

This low-dimensional demonstration model is merely for principle illustration and is not equivalent to a full Transformer; it semi-quantitatively demonstrates the relation between reciprocal coupling, critical bifurcation and local entropy reduction.

C Supplementary Derivation for Entropy-Rate Functional

C.1 Rigorous Derivation for Discrete Layers

Let the deterministic mapping from layer l to layer $l+1$ be $\mathbf{z}^{(l+1)} = \mathbf{f}^{(l)}(\mathbf{z}^{(l)})$, where $\mathbf{f}^{(l)}$ is a diffeomorphism. From change-of-variable formula:

$$p^{(l+1)}(\mathbf{z}^{(l+1)}) = p^{(l)}(\mathbf{z}^{(l)}) \cdot |\det \mathbf{J}^{(l)}(\mathbf{z}^{(l)})|^{-1} \quad (36)$$

Take $-\log$ on both sides and compute expectation:

$$S^{(l+1)} = S^{(l)} + \mathbb{E}_{\mathbf{z}^{(l)}} \left[\ln |\det \mathbf{J}^{(l)}(\mathbf{z}^{(l)})| \right] \quad (37)$$

Using $\det \mathbf{J} = \prod_i \lambda_i$, we obtain:

$$\Delta S^{(l)} = \mathbb{E}_{\mathbf{z}^{(l)}} \left[\sum_{i=1}^D \ln |\lambda_i^{(l)}| \right] = \mathcal{F}(\mathbf{J}^{(l)}) \quad (38)$$

C.2 Continuous Layer Limit

Set layer spacing $\delta l \rightarrow 0$, $\mathbf{f}^{(l)}(\mathbf{z}) \approx \mathbf{z} + \delta l \cdot \mathbf{g}(\mathbf{z})$; then:

$$\mathbf{J}^{(l)} \approx \mathbf{I} + \delta l \cdot \frac{\partial \mathbf{g}}{\partial \mathbf{z}} \quad (39)$$

$$\ln |\det \mathbf{J}^{(l)}| \approx \delta l \cdot \text{tr} \left(\frac{\partial \mathbf{g}}{\partial \mathbf{z}} \right) = \delta l \cdot \nabla \cdot \mathbf{g} \quad (40)$$

Hence under continuous limit:

$$\frac{dS}{dl} = \mathbb{E}_{\mathbf{z}} [\nabla \cdot \mathbf{g}(\mathbf{z})] \quad (41)$$

where $\mathbf{g}$ denotes the velocity field of residual flow. When $\nabla \cdot \mathbf{g} < 0$ (convergent velocity field), entropy decreases monotonically.

D Experimental Pseudocode

Algorithm 1 Coupling Strength Sweep and Criticality Detection Experiment

Require: Pre-trained model $\mathcal{M}$, gating range $\Lambda = \{\lambda_1, \dots, \lambda_K\}$, evaluation dataset $\mathcal{D}_{\text{eval}}$

Ensure: Critical threshold λ_c , entropy-reduction curve $\mathcal{F}(\lambda)$, performance curve $A(\lambda)$

```

1: for  $k = 1$  to  $K$  do
2:    $\lambda \leftarrow \Lambda_k$ 
3:   Modify residual connection of model:  $\mathbf{h}^{(l+1)} \leftarrow \mathbf{h}^{(l)} + \lambda \cdot \text{Sublayer}(\mathbf{h}^{(l)})$ 
4:   // Criticality metrics
5:   for batch  $\mathbf{x}$  in  $\mathcal{D}_{\text{eval}}$  do
6:     Forward pass and record latent states of each layer  $\{\mathbf{z}^{(l)}\}_{l=1}^L$ 
7:     Power iteration (Eq. (5)) to compute  $|\lambda_{\text{max}}^{(l)}|$ 
8:     k-NN estimator (Eq. (8)) to compute  $S^{(l)}$ 
9:   end for
10:   $\mathcal{F}(\lambda) \leftarrow$  average slope of  $\Delta S^{(l)}$  over  $l$ 
11:  // Emergent capability evaluation
12:   $A(\lambda) \leftarrow \text{Eval}(\mathcal{M}_\lambda, \mathcal{D}_{\text{eval}})$ 
13:  // Perturbation propagation
14:  Inject perturbation  $\delta \mathbf{z}^{(1)}$ , record survival probability  $P(L)$  of perturbation across layers
15:  Fit  $P(L) \propto L^{-\gamma}$  or  $P(L) \propto e^{-\beta L}$ 
16: end for
17:  $\lambda_c \leftarrow \arg \min_\lambda |\mathcal{F}(\lambda)|$  (point of maximal entropy reduction)
18: return  $\lambda_c, \mathcal{F}(\cdot), A(\cdot), \gamma(\cdot)$ 

```

References

- [1] Wang, X., et al. (2023). Layer-wise representation correction in transformer networks.
- [2] Yang, L., et al. (2024). Transformer operates at the edge of chaos.
- [3] Zhang, H., et al. (2024). Latent entropy decay in deep transformer inference.

- [4] Sornette, D., et al. (2022). Gradient flow self-organization near critical manifolds in high-dimensional loss landscapes.
- [5] Wei, J., et al. (2023). A survey of emergent abilities in large language models.
- [6] Schaeffer, R., et al. (2023). Emergent abilities of large language models: Mirage or mirage?
- [7] Cocchi, L., et al. (2017). Criticality in biological neural networks.
- [8] Friston, K. (2024). Collective near-critical dynamics in reentrant neural fields.
- [9] Oizumi, M., Albantakis, L., Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: Integrated Information Theory 3.0.
- [10] Muñoz, M. A. (2023). Critical phenomena in deep learning systems.
- [11] Kozachenko, L. F., Leonenko, N. N. (1987). Sample estimate of the entropy of a random vector. *Problems of Information Transmission*, 23(2), 95–101.
- [12] Krishnamurthy, A., et al. (2007). On the bias and variance of k-NN entropy estimators. *IEEE Trans. Info. Theory*.
- [13] Kraskov, A., Stögbauer, H., Grassberger, P. (2004). Estimating mutual information. *Physical Review E*, 69(6), 066138.
- [14] Jacot, A., Gabriel, F., Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. *NeurIPS*.
- [15] Poole, B., et al. (2016). Exponential expressivity in deep neural networks through transient chaos. *NeurIPS*.
- [16] Schoenholz, S. S., et al. (2017). Resurrecting the sigmoid in deep learning through dynamical isometry. *NeurIPS*.
- [17] Chen, R. T. Q., et al. (2018). Neural ordinary differential equations. *NeurIPS*.
- [18] Tishby, N., Pereira, F. C., Bialek, W. (2000). The information bottleneck method. *arXiv:physics/0004057*.
- [19] Rao, R. P. N., Ballard, D. H. (1999). Predictive coding in the visual cortex. *Nature Neuroscience*, 2(1), 79–87.
- [20] Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3), 200–219.
- [21] Oizumi, M., et al. (2014). Approximate measures of integrated information. *PLoS Comput. Biol.*