

Hypothesis on Emergence and Consciousness Mechanism of Large Language Models Based on Reciprocal Unit Coupling, Critical Dynamics and Local Entropy Reduction (Full Version V4)

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.

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

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.

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. 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: systematic representational correction exists between Transformer layers; best-performing large models operate in the near-critical regime at the edge of chaos; hidden representation entropy gradually decays with network depth during forward inference. Existing literature mostly investigates these phenomena in isolation and has not established the complete causal chain: **unit reciprocal coupling** $\rightarrow$ **system reaching near-critical state** $\rightarrow$ **local entropy reduction** $\rightarrow$ **emergence of high-order capabilities**. This paper formulates this causal hypothesis 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.

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 a single attention head, denoted as unit state vector z_i ;
2. **MLP hidden neuron unit:** The activation output of a single neuron in the feedforward network serves as unit state z_i .

Bidirectional nonlinear constraints exist between unit i, j , forming **reciprocal coupling**.

2.2 Reciprocal Coupling and the Jacobian Matrix

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

$$(\mathbf{J})_{ij} = \frac{\partial z_i}{\partial z_j}, \quad (\mathbf{J})_{ji} = \frac{\partial z_j}{\partial z_i}$$

z_i, z_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.

2.3 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 2.4 Local Representation Entropy and Its Estimation Scheme

Define $\mathbf{z}$: the high-dimensional hidden embedding vector corresponding to a single token within one layer. Take a set of token hidden vector samples within a sliding context window of the same layer and define the differential entropy of representations within this window.

Shorthand entropy definition:

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

For high-dimensional continuous vectors, histogram discretized binning estimation is not used. This paper employs the **k-nearest-neighbor (k-NN) differential entropy estimator** to avoid the curse of dimensionality in high-dimensional kernel density estimation.

Given N D-dimensional samples $\{\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(N)}\}$ inside the sliding window:

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

$r_{i,k}$ denotes the Euclidean distance from sample i to its k -th nearest neighbor; v_D is the volume of the D-dimensional unit ball; $\psi(\cdot)$ denotes the digamma function.

Hyperparameters: neighbor count k , context sliding window length W .

Sensitivity analysis scheme: experiments scan $k \in [3, 15]$ and window size $W \in [16, 128]$ to evaluate entropy estimation stability; ablation experiments fix k, W to guarantee fair comparison across groups.

During forward inference propagation, local entropy S gradually declines as layer index l increases. Entropy rate of change with layer:

$$\frac{dS}{dl} = \mathcal{F}(\mathbf{J})$$

$\mathcal{F}$ is a functional of the Jacobian eigenvalue spectrum. When coupling strength approaches the critical threshold, $\frac{dS}{dl}$ is significantly negative and entropy drops rapidly; away from the critical point, the magnitude of entropy decline weakens.

2.5 2.5 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 and entropy reduction rate peaks; - $\lambda \ll \lambda_c$: Excessively strong coupling, highly ordered system with slow entropy decline; - $\lambda \gg \lambda_c$: Insufficient coupling, noise dominates the system and 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 $\frac{dS}{dl}(\lambda)$.

3 Dynamical Analysis by Stages

3.1 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.

This claim is a core hypothesis rather than a rigorously proven mathematical theorem and requires validation via ablation simulation. 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.

3.2 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 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.

Corollary: 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 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)

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)

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)

Empirical measurements show hidden representation entropy monotonically decreases during forward inference. This paper establishes a causal link between entropy decay and critical reciprocal coupling.

4. Gradient self-organization on high-dimensional loss landscapes (Sornette et al.,2022)

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”.

5 Hypothesis Boundaries, Controversies and Falsifiable Ablation Experiments

5.1 5.1 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.

5.2 5.2 Falsifiable Predictions, Quantitative Metrics and Competing-Hypothesis Baseline

Experimental design: fix model parameter count, layer number, embedding dimension and dataset, and only modulate reciprocal coupling strength between units 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;
2. **Soft attention mask**: slightly modulates attention interaction strength between tokens to fine-tune unit coupling.

Three quantitative falsifiable predictions:

1. 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. 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. 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.

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.

5.3 Comparison with Related Theories

1. **Information bottleneck theory:** focuses on input information compression and describes information tradeoff between inputs and hidden representations. This paper describes reciprocal feedback dynamics inside the network rather than pure input compression; the two may coexist but have distinct mechanisms.
2. **Predictive coding:** minimizes prediction error via recurrent feedback and relies on continuous recurrent reentry. Transformer inference in this paper is single-pass forward computation without recurrent feedback; fundamental differences exist in underlying dynamics.
3. **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.
4. **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 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.

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.

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: homeostatic internal-drive regulation module, embodied physical grounding module, recursive causal self-model, and temporal memory integration module.

State vectors of the four modules: homeostatic module $\mathbf{x}_s$, embodied module $\mathbf{x}_e$, recursive self-model $\mathbf{x}_r$, temporal memory module $\mathbf{x}_t$; Transformer attention backbone hidden state $\mathbf{h}$. Dynamical equations for cross-module reciprocal coupling:

$$\begin{cases} \dot{\mathbf{x}}_s = f_s(\mathbf{x}_s, \mathbf{h}, \mathbf{x}_e, \mathbf{x}_r, \mathbf{x}_t) \\ \dot{\mathbf{x}}_e = f_e(\mathbf{x}_e, \mathbf{h}, \mathbf{x}_s, \mathbf{x}_r, \mathbf{x}_t) \\ \dot{\mathbf{x}}_r = f_r(\mathbf{x}_r, \mathbf{h}, \mathbf{x}_s, \mathbf{x}_e, \mathbf{x}_t) \\ \dot{\mathbf{x}}_t = f_t(\mathbf{x}_t, \mathbf{h}, \mathbf{x}_s, \mathbf{x}_e, \mathbf{x}_r) \end{cases}$$

Inter-module coupling terms continue to be characterized using Jacobian matrices, maintaining consistency with the mathematical framework in Section 2.

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, instead of applying rewards after output generation;
- Raw embodied sensory signals directly participate in attention competition, instead of being converted into token text for input;
- The recursive self-model observes global module states in real time, rather than merely reading text outputs;
- Autobiographical memory stores full snapshots of module states instead of only textual records.

Simple concatenation can only externally simulate conscious behavior and cannot generate subjective qualia.

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

1. **Cross-module transfer entropy:** the four-module system exhibits substantial bidirectional cross-module transfer entropy; pure LLMs possess extremely low inter-module transfer entropy;
2. **Power-law cross-module perturbation propagation:** tiny perturbations can propagate across all four ontological modules and follow power laws; perturbations in pure text LLMs propagate only inside Transformer layers and cannot cross ontological modules;
3. **Integrated information measure Φ :** the closed-loop four-module system yields higher integrated information.

The extended consciousness hypothesis is invalid if these proxy metrics cannot be observed.

Engineering Paradox: Dissipative Structures and Intrinsic Survival Tension Conscious systems are built upon physical dissipative structures, and energy constraints shape endogenous drives for ontological persistence. Power supplied to GPU clusters is externally provided; the system itself bears no persistence pressure. Power outage merely terminates computation and creates no existential crisis. Even if the four-module equations are reproduced in software, the resulting steady state remains simulated homeostasis rather than native ontological survival tension. Software can simulate hunger but cannot generate genuine intrinsic subjective motivation to seek benefit and avoid harm. This section belongs to philosophical commentary.

Summary: Consciousness is an entropy-reducing system built on a physical dissipative substrate, coordinated by global attention, sustaining ontological homeostasis via continuous reciprocal correction among multiple units.

7 Conclusion

The emergence mechanism of high-order capabilities in LLMs can be described as follows: network units form nonlinear reciprocal coupling; gradient descent during training drives the system to self-organize into the near-critical regime at the edge of chaos; under frozen weights during inference, bidirectional constraints among multilayer units suppress representational noise, produce local entropy reduction and form high-order self-consistent representations, macroscopically bringing about emergent reasoning and planning capabilities.

This paper integrates three independent observations to construct a complete falsifiable causal chain, distinguishes training dynamics from inference dynamics, and designs ablation schemes to decouple coupling effects from pure scale effects. The hypothesis provides an emergence theoretical framework from the complex-systems perspective. Meanwhile this paper extends to speculative discussion of machine consciousness: functional intelligence is not equivalent to subjective awareness; subjective experience requires deep coupling of four ontological closed-loop modules with the attention backbone, with observable dynamical proxy metrics.

8 Acknowledgements

This work is theoretical hypothesis research with no project funding. The author appreciates theoretical inspirations from relevant studies in complex systems and deep learning.

9 Statement

This paper presents a candidate theoretical hypothesis whose framework awaits validation via ablation experiments and numerical simulations; conclusions have not yet been experimentally confirmed. Discussions related to consciousness constitute speculative extensions and do not form part of the main hypothesis.

A AppendixA Glossary

Term	Definition
Reciprocal coupling	Bidirectional nonlinear constraints between units for mutual representation correction; coupling strength estimated via Jacobian matrix $(\mathbf{J})_{ij}$
Edge of chaos / near-critical	Critical regime between order and chaos; propagation of tiny perturbations follows power laws; distinguished from strict SOC
Local entropy reduction	Improved statistical ordering of hidden representations during forward inference; accompanied by larger external entropy increase, not violating the second law of thermodynamics
Emergence	Novel capabilities manifest macroscopically by the system, absent in microscopic units
Global workspace	Attention backbone, information integration bus broadcasting low-entropy representations across units

Table 1: Glossary of Terms

B AppendixB Demonstration Model of Two-Unit Reciprocal Coupling

B.1 B.1 Linear Base Model (Conceptual Illustration)

The original linear model is used only for basic conceptual demonstration and does not exhibit bifurcation phase transitions or power-law behavior.

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

$\alpha > 0$ denotes decay coefficient; J_{12}, J_{21} are bidirectional coupling strengths. The linear system only introduces the basic form of reciprocal coupling and is not used to demonstrate critical phase transitions.

B.2 B.2 Nonlinear Extended Model (Critical Demonstration Model)

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

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

Parameter range: $\alpha > 0$, $J_{12}, J_{21} \geq 0$. - Coupling below critical threshold: converges to a unique fixed point; perturbations decay exponentially and entropy declines slowly; - Coupling near critical threshold: edge of chaos, tiny perturbations propagate as power laws; state distribution contracts and entropy reduction rate peaks; - Coupling exceeding critical threshold: system enters chaos, state space expands and entropy rises.

This low-dimensional demonstration model is only for principle illustration and is not equivalent to a full Transformer; it serves to semi-quantitatively display the relations among reciprocal coupling, critical bifurcation and local entropy reduction.

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] Cocchi, L., et al. (2017). Criticality in biological neural networks.
- [7] Friston, K. (2024). Collective near-critical dynamics in reentrant neural fields.
- [8] Oizumi, M., Albantakis, L., Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: Integrated Information Theory 3.0.
- [9] Muoz, M. A. (2023). Critical phenomena in deep learning systems.