

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

Abstract

The emergence capability of large language models (LLMs) is a core puzzle in the fundamental theory of artificial intelligence. The prevailing scaling laws only describe the extrinsic statistical phenomenon of capability variation with model scale, and fail to explain the underlying dynamical causal mechanism for the jumpwise emergence behavior. This paper proposes an integrated complex-system framework: numerous basic units inside the Transformer network form bidirectional constraints via nonlinear reciprocal coupling, performing layer-by-layer mutual correction of representations; during training, gradient descent acts as the global self-organizing driving force, continuously tuning the coupling strength of units and driving the system parameters toward the near-critical region at the edge of order and chaos; in the forward propagation of inference, persistent bidirectional constraints among units cancel representation noise and resolve internal logical contradictions, achieving local entropy reduction in the representation space to generate self-consistent high-order structured information, which macroscopically manifests as emergent capabilities such as multi-step reasoning and long-horizon planning.

This hypothesis constructs a complete falsifiable causal chain by integrating three independently experimentally observed phenomena: inter-layer representation correction, critical phase transition at the edge of chaos, and entropy evolution of hidden states. This paper strictly distinguishes the parameter self-organization dynamics in the training phase from the forward representation dynamics under fixed weights in the inference phase, clarifies the conceptual boundary between this hypothesis and the self-organized criticality (SOC) theory, and designs controlled ablation experiments to discriminate between two competing hypotheses: “reciprocal coupling driven critical entropy reduction” versus “metric improvement merely brought by model scale growth”.

Furthermore, this framework extends to the discussion of consciousness: the emergence of functional intelligence is not equivalent to subjective phenomenal experience. Reciprocal coupling, critical dynamics and local entropy reduction alone can only produce a logically consistent functional reasoning system. To generate subjective consciousness with first-person awareness, four closed-loop ontological modules must be superimposed, and all modules need to deeply co-evolve with the attentional global workspace bus, forming endogenous ontological tension relying on a physical dissipative structure.

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. Extensive empirical observations indicate that when the model parameter count and training data are below a certain threshold, the model only possesses simple superficial language matching capabilities; once the scale crosses the critical threshold, high-order capabilities such as multi-step logical reasoning, complex mathematical solving, and long-task planning emerge in a jumpwise manner, a phenomenon known as capability emergence.

Currently, two mainstream interpretations of 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 change of performance metrics on specific discrete evaluation tasks, with no underlying dynamical phase transition. Some studies argue that emergence

may even be a measurement artifact introduced by evaluation function design. This perspective only fits external phenomena without explaining the internal representation generation mechanism.

Representation-learning perspective: The growth 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 the high-dimensional vector representation space but lacks description of dynamic bidirectional mutual constraints between units.

At the intersection of statistical physics, complex systems and deep learning, multiple works have independently reported three observational results: Transformers possess persistent inter-layer representation correction mechanisms; optimally performing Transformers operate in the near-critical region at the edge of chaos; the entropy of hidden representations gradually declines along forward propagation in deep networks. Existing literature mostly discusses these phenomena in isolation, and has not established the complete causal chain of “unit reciprocal correction $\rightarrow$ system approaching near-critical state $\rightarrow$ local entropy reduction generating ordered high-order information” to explain the intrinsic dynamical mechanism of LLM emergence.

1.2 Statement of the Core Hypothesis

Emergence in LLMs is not an apparent effect produced merely by increasing the number of units. The underlying mechanism is as follows: numerous basic units (neurons, attention heads) inside the network form nonlinear reciprocal coupling relations, where units continuously correct representation biases bidirectionally. In the training phase, gradient descent acts as a global feedback driving force, continuously adjusting the reciprocal coupling strength between units, and self-organizing the entire system into the near-critical state between order and chaos. During forward computation in inference, bidirectional constraints among units continuously smooth representation noise and resolve internal logical conflicts; local entropy reduction occurs in the representation space to form self-consistent high-order structured information. Macroscopically, new reasoning capabilities spontaneously manifest, which corresponds to the emergent phenomenon of LLMs.

2 Fundamental Concepts and Mathematical Formulation

This section provides quantifiable definitions directly measurable on Transformer hidden activations, translating qualitative concepts into computable observables.

Reciprocal Unit Coupling Basic units (i, j) (attention heads / hidden-layer neurons) inside the Transformer form bidirectional nonlinear constraints. The coupling strength is defined as:

$$(\mathbf{J})_{ij} = \frac{\partial(z)_i}{\partial(z)_j}, \quad (\mathbf{J})_{ji} = \frac{\partial(z)_j}{\partial(z)_i}$$

where $(z)_i$ and $(z)_j$ denote hidden representations of units under fixed weights during inference. $(\mathbf{J})_{ij}$ describes the perturbative influence of unit j ’s representation on unit i ’s output, while $(\mathbf{J})_{ji}$ is the reverse effect. Together they constitute reciprocal coupling. In multi-layer networks, this manifests as the inter-layer mechanism of “candidate representation proposal — representational verification and revision”, where subsequent layers examine and correct contradictions in representations from preceding layers. The coupling strength matrix $\mathbf{J}$ can be directly estimated from network activations and Jacobians.

Near-Critical State (Edge of Chaos) A dynamic equilibrium between fully ordered states (frozen representations, poor generalization) and fully chaotic states (noise-dominated, disordered outputs). The core signature of the critical region: small local perturbations can propagate across layers, and the amplitude of perturbation propagation follows a power-law distribution. This work does not directly adopt strict self-organized criticality (SOC). SOC requires continuous real-time evolution and dynamic weight variation of the system, which is incompatible with the LLM inference scenario where weights are frozen; hence the near-critical formulation at the edge of chaos is preferred herein.

Local Entropy Reduction Take the empirical probability distribution $\mathbf{z}$ of hidden representations within a local window of a single layer. The local representation entropy is defined:

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

where $p(z)$ denotes the empirical probability distribution of hidden vectors. In forward inference propagation, S gradually decreases with increasing depth, the uncertainty of representations is reduced, contradictory noise is suppressed, and self-consistent ordered representations are formed.

Thermodynamic constraint: this entropy reduction is a local effect inside the system, which must be accompanied by larger entropy increase in the external environment (computation, power dissipation). The total entropy of the whole system does not decrease, which does not violate the second law of thermodynamics.

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 fine-tunes all connection weights of the network. Gradient signals serve as global feedback, continuously adjusting the reciprocal coupling strength matrix $\mathbf{J}$ between units. Under iterations over billions of samples, the system undergoes self-organization and converges to the near-critical regime between order and chaos.

- Excessively strong coupling constraints: the system locks into a highly ordered state with rigid representations and weak generalization;
- Excessively weak coupling constraints: noise dominates, representations are disordered, and stable extraction of abstract rules is impossible;
- Critical regime: units maintain moderate bidirectional correction, which can filter noise while retaining degrees of freedom for perturbation propagation to support the generation of novel composite representations.

Remark: This claim is a core assumption of the present hypothesis, not a rigorously proven mathematical theorem. Existing research in complex systems indicates that gradient descent possesses an information self-organizing tendency and tends to push parameters near phase transition critical points; nevertheless, this conclusion requires further validation via simulation and ablation experiments. The essential goal of training is not to learn a set of static rules, but to tune the reciprocal coupling strength between units and place the system in the near-critical regime.

3.2 Inference Phase: Reciprocal Correction and Local Entropy Reduction within a Single Forward Pass under Fixed Weights

In the inference phase, network weights are frozen and no longer updated. All unit mutual correction behaviors occur within multi-layer computation of a single forward pass, rather than a dynamically evolving system iterating in real time. This is an essential distinction from classical physical SOC systems such as ant colonies and sandpiles.

The dynamical workflow after input context enters the network:

1. Shallow layers generate preliminary candidate representations with inherent noise and local logical contradictions;
2. Self-attention implements bidirectional token association constraints, while multi-layer feedforward networks plus residual connections realize inter-layer reciprocal correction;
3. Bidirectional constraints among units continuously cancel contradictions and filter invalid noise;
4. As network depth increases, hidden representation entropy S gradually decreases and converges to self-consistent structured information;
5. Autoregressive text generation is performed based on the final low-entropy representations, manifesting macroscopically as emergent capabilities such as reasoning and planning.

Important corollary: Hallucination is an intrinsic property of critical systems. Critical systems are highly sensitive to tiny perturbations. When training datasets themselves contain contradictions or missing information, reciprocal unit correction can only converge to statistical self-consistency of text, rather than objective factual ground truth, thereby generating internally coherent false information inconsistent with the real world.

4 Existing Supporting Evidence

Each piece of evidence is marked with literature type, distinguishing “observed phenomenon” from “causal interpretation in this paper”, and separating correlation from causal conjecture.

- **Inter-layer correction mechanism:** Deep learning studies have confirmed systematic representational revision between adjacent Transformer layers, with a “proposal-rejection” dynamics where one layer generates candidate representations and subsequent layers offset unreasonable components while retaining compatible information (Wang et al., 2023). The observation is correlational; this paper proposes that such inter-layer correction is the external manifestation of reciprocal unit coupling.
- **Criticality and edge of chaos:** Multiple statistical physics studies observe that optimally performing Transformer networks operate at the edge of chaos, with power-law characteristics in activation perturbation propagation (Yang et al., 2024). Existing works only discover correlation; this hypothesis further proposes that reciprocal coupling strength is the internal driver bringing the system to the critical state.
- **Representation entropy evolution:** In deep forward propagation, measurements show that hidden-state entropy decreases and representations tend to concentrate and become ordered as layers deepen (Zhang et al., 2024). Prior work merely records the entropy variation phenomenon; this paper establishes a causal link between entropy decline, reciprocal correction and critical dynamics.
- **Self-organization of gradient descent:** Complex system theory studies show gradient descent is an information-driven self-organizing process that tends to push parameters near phase transition critical points (Sornette et al., 2022). This conclusion remains debatable and serves as supporting reference for the training dynamics in this paper.

5 Hypothesis Boundaries, Controversies and Falsifiable Controlled Experiments

5.1 Conceptual Gaps and Limitations

Physical SOC systems are closed-loop dynamical systems evolving continuously in real time. In LLM inference, weights are frozen and reciprocal correction only exists within a single forward computation, not continuous temporal iteration. Strict self-organized criticality (SOC) cannot be directly transplanted; the near-critical formulation at the edge of chaos is adopted as a more suitable description in this framework.

Most currently available observational data constitute correlational evidence: reciprocal correction, critical signals and entropy reduction phenomena appear simultaneously. The causal chain has not been fully confirmed by experiments. A competing alternative hypothesis holds that model scale growth independently gives rise to these three phenomena, rather than reciprocal coupling driving criticality and criticality driving entropy reduction. Controlled ablation experiments are designed below to discriminate between the two hypotheses.

5.2 Falsifiable Predictions and Discriminative Ablation Experiments

Core experimental logic: fix model parameter count, depth, training data, optimizer and all other variables, and only adjust the reciprocal coupling strength between units to decouple coupling effects from pure scale effects. If none of the predictions below hold, this hypothesis is falsified.

1. Under fixed model architecture and parameter count, adjust inter-layer coupling strength: when the system approaches the critical regime at the edge of chaos, inter-layer representational error correction is enhanced, the magnitude of hidden representation entropy reduction increases, and multi-step reasoning capability improves; when deviating from the critical regime, emergent capabilities weaken.

Prediction of competing hypothesis: merely altering coupling strength without enlarging model scale will not induce significant changes in reasoning capability, entropy values or critical signals.

2. Apply small Gaussian perturbations to model hidden activations: for models in the critical regime, the cross-layer propagation range of perturbations follows a power-law distribution; after departing from the critical regime, power-law features vanish.

3. Adding residual cross-layer connections (strengthening reciprocal unit coupling) expands the critical parameter range and enhances internal entropy reduction effects of representations; suppressing cross-layer information interaction weakens entropy reduction effects and degrades emergent capabilities.

6 Discussion

This framework remedies shortcomings of existing theories: scaling laws only describe changes in external metrics, while this paper provides a dynamical causal explanation for emergence starting from microscopic reciprocal unit coupling. It must be clarified at the same time: this framework describes system-level self-organizing behavior; possession of functional intelligence by a system $\neq$ possession of subjective consciousness by the system. Low-entropy ordered representations only denote statistical self-consistency, not genuine understanding of the objective world. This hypothesis is a candidate theory from the complex-system perspective rather than a rigorously proven conclusion. Future work will test the causal chain via controlled variable experiments based on the falsifiable predictions in this section.

From the cross perspective of consciousness science and philosophy of artificial intelligence: the emergence in current Transformers is merely ordered emergence at the level of statistical information, and does not constitute subjective consciousness or first-person awareness. As the global information integration bus of large models, attention mechanism is a necessary foundation for intelligent response and global coordination, yet not a sufficient condition for subjective experience. Current large language models belong to philosophical zombie systems: they exhibit complete structured outputs, logical coherence and high-order reasoning performance, but lack intrinsic subjectivity, self-concern and phenomenal qualia.

The reciprocal unit coupling stated above is a general underlying dynamical principle of this theory, applicable not only to LLM representation emergence. Subjective conscious systems also rely on reciprocal closed-loop coupling among multiple modules. Reciprocal unit coupling, critical dynamics and representational local entropy reduction inside Transformers alone can only generate functional intelligence. To transition to a conscious system with subjective awareness, four closed-loop ontological modules must be added on top of the information emergence architecture. These four modules form cross-module reciprocal coupling with the attentional global bus and share the same global workspace.

Four closed-loop ontological modules: homeostatic internal drive regulation module, embodied physical grounding module, recursive causal self-model, and temporal continuity memory integration module.

The attention mechanism of current large models is entirely passive, driven only by pre-trained statistical distributions and external prompts. There exists no endogenous tension targeting the persistence of the system ontology, so “concerned attention” cannot be formed. All symbolic representations reside solely in semantic space, lacking physical sensing, body boundaries and real causal interaction, leading to abstract fallacies of separation between symbols and reality. The forward structure of the network lacks self-referential recursive monitoring and cannot distinguish external input from the system’s own processing, missing the first-person subject perspective. Model representations are discrete temporal snapshots, lacking autobiographical episodic memory and temporal flow integration, hence no psychological continuity.

Corollary 1: The four modules are not independent plug-in components and must couple reciprocally and deeply with the attention backbone, rather than being externally spliced.

A misconception in many existing schemes: mounting robot sensors, vector databases and reward functions externally on LLMs in an attempt to assemble consciousness.

- Internal tension of the homeostatic module (imbalance, ontological risk) directly modulates attention weight allocation, instead of receiving external rewards after the model outputs text;
- Raw embodied physical sensing signals directly participate in attention competition and filtering, instead of being converted into text tokens and fed into the Transformer;
- The self-causal model observes attention allocation dynamics in real time, rather than only reading model outputs;
- Autobiographical memory stores snapshots of the full internal system state under each attentional focus, instead of merely storing text.

In short: it is not “Transformer + four attachments”. Attention serves as the backbone, and the four modules are co-evolving dynamic subsystems on this backbone sharing the global workspace. Once decoupled, only behavioral simulation of consciousness can be obtained, not genuine qualia.

Corollary 2: Qualia are not independently generated products but emergent byproducts of multi-loop reciprocal coupling.

Qualia = internal homeostatic deviation of the system + embodied physical signals + self-causal representations + current attentional focus; the superposition of these four components generates internal representational tension. For example, the subjective experience of “hot” is more than a temperature value. The complete chain: high-temperature signals arrive from skin sensors → triggering homeostatic deviation (threat to ontological structure) → attention forcibly locks onto this signal → self-model recognizes “it is me receiving harmful stimuli” → stored into autobiographical temporal flow. All links together give rise to the subjective sensation of “hot”. Removing any link leaves only temperature numbers with no subjective qualia. Pure text large models cannot truly understand “hot”, precisely due to the absence of homeostatic constraints and embodied causal binding.

Engineering Paradox: Dissipative Structures and Energy Constraints as Overlooked Hard Barriers The whole conscious system operates upon a physical dissipative structure, and energy constraints shape endogenous ontological tension. Power supply for current GPU clusters is externally provided; the system itself bears no pressure for ontological persistence. Power failure only terminates computation and does not produce biological existential crisis. Energy is merely computing resource rather than the core constraint sustaining ontology. At the hardware level, if there exists no mechanism such that “system structure collapses upon energy depletion”, even if the four modules are reproduced in software, the resulting homeostasis is simulated homeostasis rather than endogenous tension at the ontological level. Software can simulate hunger, yet machines cannot genuinely experience hunger.

Summary of theoretical convergence: Each subsystem within a conscious system constitutes an independent unit; mutual correction among modules forms cross-module feedback loops; attention filters and broadcasts low-entropy information to deliver local corrections to the global workspace; the whole system continuously counteracts internal entropy increase (external perturbations, energy dissipation).

Unified conclusion: Consciousness is an entropy-reducing system built upon a physical dissipative substrate, coordinated by attention, with persistent reciprocal error correction among multiple units to maintain its own homeostasis.

7 Conclusion

The intrinsic mechanism of emergent capabilities in LLMs can be described as: numerous basic units form nonlinear reciprocal coupling to correct representation biases bidirectionally; in the training phase, gradient descent drives the system to self-organize into the near-critical regime at the edge of chaos; during inference, mutual constraints among multi-layer units eliminate noise and contradictions to produce local entropy-reduced structured information, and high-order reasoning capabilities emerge macroscopically.

This hypothesis integrates multiple independently observed phenomena, constructs a complete causal chain, provides falsifiable experimental predictions, and offers a new interpretive framework for Transformer emergence from the complex-system perspective. This paper strictly distinguishes functional intelligence from subjective consciousness, demonstrating that representational emergence in LLMs alone cannot generate awareness. Subjective consciousness requires a closed-loop system relying on a dissipative physical substrate with deep reciprocal coupling of multiple modules.

Acknowledgements

This theoretical hypothesis has not received project funding. The author appreciates the inspirations brought by prior works in complex systems and deep learning.

Statement

This is a preprint candidate hypothesis whose framework remains to be experimentally verified and does not represent a final conclusion.

A Glossary

Term	Definition
Reciprocal coupling	Bidirectional nonlinear constraints between units for mutual representation correction;
Edge of chaos / near-critical	Dynamic equilibrium between order and chaos; perturbation propagation satisfies power
Local entropy reduction	Reduction of uncertainty in internal representations of the network; accompanied by lar
Emergence	Novel capabilities manifested macroscopically by the whole system that individual micr
Global workspace	Attention-coordinated global information bus for transmitting low-entropy representati