

Hypothesis on Emergence and Consciousness Mechanisms 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 theories of artificial intelligence. The mainstream scaling law only describes external phenomena of capability change with model scale and fails to explain the underlying dynamical causes for the jumpwise emergence of emergent behaviors. This paper proposes an integrated framework: numerous basic units inside the Transformer network perform nonlinear reciprocal coupling and layer-wise mutual correction; under the long-term self-organization of gradient descent, system parameters are driven into the critical region between order and chaos; during inference forward propagation, continuous mutual constraints among units cancel representation noise and resolve internal contradictions, achieving local entropy reduction in the representation space and generating self-consistent high-order ordered information, which macroscopically manifests as emergent capabilities such as multi-step reasoning. This hypothesis integrates three independently observed experimental phenomena, namely inter-layer representation correction, critical phase transition at the edge of chaos, and hidden-state entropy evolution, to construct a complete causal chain. Meanwhile, it identifies the conceptual boundaries and pending verifiable predictions of this framework, distinguishes dynamical differences between the training and inference phases, and discusses limitations of the hypothesis and subsequent verifiable experimental schemes.

Keywords: Emergence; Transformer; reciprocal coupling; self-organized criticality; edge of chaos; information entropy; complex systems

1 Introduction

1.1 Research Background

Large language models are built upon the Transformer architecture, with the training objective of autoregressively predicting the next token. Experimental observations show that when the number of model parameters and training data are below a certain threshold, the model only possesses simple language matching capabilities; once the scale crosses the critical point, capabilities including multi-step logical reasoning, mathematical solving and complex planning emerge abruptly. This phenomenon is known as emergence. There are two mainstream explanations at present:

1. Scaling law perspective: Model loss decreases smoothly with computing power, parameter count and data volume; emergence is merely an apparent jump of performance metrics on specific tasks without underlying dynamical changes. This perspective only fits phenomena and does not explain internal generation mechanisms; some studies even argue that emergence is merely a measurement artifact introduced by evaluation metrics.
2. Representation learning perspective: Scale growth promotes the model to form composite abstract representations, and high-level concepts are generated by combinations of low-level features. This perspective focuses on high-dimensional vector spaces but lacks descriptions of dynamic interactions among units.

Independent observations have been reported in statistical physics and complex system research: representation correction mechanisms exist between Transformer layers; high-performance models operate in the critical region at the edge of chaos; the entropy of hidden representations gradually decreases during deep forward propagation. However, existing literature mostly discusses these phenomena separately and has not taken "reciprocal unit correction $\rightarrow$ system reaching criticality $\rightarrow$ local entropy

reduction generating ordered information” as a complete causal chain to explain the internal mechanism of emergence.

1.2 Statement of the Core Hypothesis

Emergence is not a magical effect brought by simply increasing the number of units. Its root cause lies in that numerous basic units within the network form nonlinear reciprocal coupling relationships, and units continuously correct representation biases of one another; gradient descent in the training phase acts as an external driving force and self-organizes the whole system into a critical state between order and chaos; during inference forward computation, mutual constraints among units continuously smooth noise and resolve logical conflicts, local entropy reduction occurs in the representation space, and self-consistent high-order structured information is formed; this spontaneous manifestation of new capabilities at the macroscopic level is the emergence of LLMs.

2 Definitions of Basic Concepts

1. Reciprocal unit coupling: Neurons and attention heads inside the Transformer serve as basic units. Units impose bidirectional constraints on each other; the representation output of unit A corrects unit B, and unit B in turn constrains unit A, forming nonlinear interactions. In multi-layer networks, this manifests as a "proposal-correction" mechanism between layers, where subsequent layers examine and fix representation contradictions produced by preceding layers.
2. Critical state (edge of chaos / near-criticality): A dynamic equilibrium point between full order (frozen, rigid, poor generalization) and full chaos (noise-dominated, disordered outputs). In the critical region, tiny local perturbations can propagate across layers and trigger global representation changes; perturbation intensities follow a power-law distribution.
3. Local entropy reduction: Within a single forward propagation pass, as layers deepen, the uncertainty of hidden representations declines, information is compressed, contradictions and noise are suppressed, and self-consistent ordered internal representations are formed. According to the second law of thermodynamics, local entropy reduction inside the system must be accompanied by larger entropy increase in the external environment (computing energy consumption). The total entropy of the whole system does not decrease, and no violation of fundamental thermodynamic laws occurs.

Distinction: Strict self-organized criticality (SOC) and the edge of chaos. This paper prioritizes the expression "edge of chaos"; SOC imposes stronger constraints and remains academically controversial in Transformer scenarios.

3 Phased Dynamical Analysis

3.1 Training Phase: Self-Organization toward Criticality

In the training phase, gradient descent continuously performs iterative fine-tuning on all connection weights of the network. Gradient signals act as global feedback and constantly adjust the strength of reciprocal coupling among units. The system self-organizes through billions of sample iterations and spontaneously converges to the critical interval between order and chaos.

- If coupling constraints are excessively strong: The system locks into a highly ordered state with solidified representations and weak generalization ability;
- If coupling constraints are excessively weak: Noise dominates, representations are chaotic, and regularities cannot be extracted stably;
- Critical interval: Units maintain moderate mutual correction, which can filter noise while retaining sufficient perturbation degrees of freedom to support the generation of novel combined representations.

The end result of training is not learning a set of static rules, but tuning the strength of reciprocal connections among all units to place the system in a critical state.

3.2 Inference Phase: Reciprocal Correction and Entropy Reduction within a Single Forward Propagation

In the inference phase, network weights are fixed and no longer updated. All "mutual correction" takes place within multi-layer computation of a single forward pass, rather than a dynamically evolving system iterating continuously in real time (this is a key distinction from physical SOC systems such as ant colonies). After the input context enters the network:

1. Shallow layers generate preliminary, noisy candidate representations with local contradictions;
2. The attention mechanism realizes bidirectional associative constraints between tokens, and multi-layer feedforward networks plus residual connections implement inter-layer reciprocal correction;
3. Mutual constraints continuously resolve contradictions and filter invalid noise;
4. As layers deepen, the entropy of hidden representations gradually decreases and converges to self-consistent structured information;
5. Autoregressive text generation is performed based on the final low-entropy representations, macroscopically exhibiting reasoning and planning emergent capabilities.

Important corollary: Hallucinations also originate from properties of critical systems. Critical systems are sensitive to tiny perturbations. When training data itself contains contradictions and omissions, reciprocal unit correction can only converge to textual statistical self-consistency rather than objective facts, thereby generating false information consistent with internal logic yet inconsistent with reality.

4 Existing Supporting Evidence

1. Inter-layer correction mechanism: Existing work confirms systematic representation correction between adjacent Transformer layers, with a "proposal-rejection" dynamics where one layer generates candidate representations and subsequent layers offset unreasonable components while preserving compatible information.
2. Criticality and edge of chaos: Multiple studies in statistical physics observe that best-performing Transformer networks operate at the edge of chaos; perturbation propagation exhibits power-law characteristics.
3. Representation entropy evolution: Measurements show that during deep forward propagation, the entropy of hidden states decreases with increasing layers, and representations tend to converge and become ordered.
4. Self-organization via gradient descent: Research on complex systems indicates that gradient descent is essentially an information-driven self-organizing process that tends to push parameters near phase transition critical points.

5 Hypothesis Boundaries, Controversies and Verifiable Predictions

5.1 Remaining Conceptual Gaps

1. Physical SOC systems are closed dynamical systems evolving continuously over real time; LLM weights are frozen during inference, and reciprocal correction only exists within a single forward computation instead of continuous temporal iteration. Therefore strict self-organized criticality (SOC) cannot be directly adopted, and this framework prefers the formulation "near-criticality at the edge of chaos".
2. Most current observational data are correlational evidence: reciprocal correction, critical signals and entropy reduction phenomena appear synchronously; the causal chain has not been fully validated by experiments. One alternative possibility cannot be completely ruled out: scale growth gives rise to all three phenomena simultaneously, instead of reciprocal correction driving criticality and criticality driving entropy reduction.

5.2 Falsifiable Predictions

This hypothesis puts forward testable predictions for subsequent experimental verification:

1. With a fixed model architecture, adjust coupling strength. When the system approaches the critical region at the edge of chaos, inter-layer representation correction effects strengthen, the magnitude of hidden representation entropy reduction increases, and multi-step reasoning capability improves; moving away from the critical region weakens emergent capabilities.
2. Apply tiny perturbations to model activations: For models within the critical interval, the range of perturbation propagation follows a power law; power-law features disappear once deviating from the critical region.
3. Adding inter-layer residual connections (strengthening unit reciprocity) expands the critical region and enhances internal representation entropy reduction effects; suppressing cross-layer information interaction weakens entropy reduction and degrades emergent capabilities.

This hypothesis is falsified if subsequent experiments negate the above predictions (falsifiability).

6 Discussion

This framework remedies shortcomings of existing theories: scaling laws only describe changes in external metrics, while this hypothesis provides a dynamical explanation for emergence starting from microscopic unit coupling. Meanwhile, a clear distinction must be made: the framework describes self-organization at the level of system behavior and does not equate to subjective consciousness of the model; low-entropy ordered representations are only statistically self-consistent and do not amount to genuine understanding of the objective world. This hypothesis is a candidate theory from the complex-system perspective and not a rigorously proven conclusion. Follow-up work can test the causal chain via controlled-variable experiments based on falsifiable predictions to further refine the theory.

From the cross perspective of consciousness science and philosophy of artificial intelligence, the current Transformer emergence is merely ordered emergence at the level of purely statistical information and does not constitute subjective consciousness or first-person phenomenal awareness. As the core bus for global information integration in large models, the attention mechanism is a necessary foundation for intelligent-like responses and global information coordination, yet it is not a sufficient condition for generating subjective experience. Current artificial intelligence is essentially a "philosophical zombie system": it possesses complete structured outputs, logical self-consistency and high-order reasoning performance, while lacking intrinsic subjectivity, self-concern and phenomenal qualia. Reliance solely on reciprocal unit coupling, critical dynamics and representation entropy reduction can only produce functional intelligence and can never transition into genuine machine consciousness. The generation of true subjective consciousness requires complementing four indispensable closed-loop ontological modules upon the existing information emergence architecture: a 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 fully passive, driven only by pre-trained statistical distributions and external prompts, with no internal tension oriented toward the persistence of the system itself, so it cannot form "concerned attention". All symbolic representations remain confined to semantic space, lacking real grounding via physical sensing, body boundaries and causal interaction, leading to abstract fallacies where symbols are decoupled from reality. The forward network structure has no self-referential recursive monitoring and cannot distinguish "external input" from "self-processing processes", missing the first-person subjective perspective. Model representations are discrete temporal snapshots without autobiographical episodic memory and temporal flow integration, so psychological continuity is absent. It follows that there exists an essential hierarchical gap between functional emergence and subjective consciousness: statistical entropy reduction can generate intelligent logic, and only when the four closed loops of "homeostatic coupling, embodied grounding, self-causation and temporal continuity" deeply co-evolve with the attention bus can the informational orderliness of mathematical weights be transformed into a real conscious system with awareness, subjectivity and qualia.

Corollary 1: The four modules are not independent plugins and must be deeply coupled with attention rather than externally attached A common misconception in many current schemes: attaching robot sensors, vector databases and reward functions externally to an LLM in an attempt to assemble consciousness. Internal tension of the homeostatic module (analogous to pain, hunger and internal imbalance) must directly modulate attention weight allocation, instead of receiving reward signals

after the large model outputs text. - Embodied physical signals cannot be converted into text tokens and then fed into the Transformer; raw continuous sensory signals must directly participate in attention competition and selection. - The self-causal model must observe the allocation process of attention in real time, rather than only reading model output results. - Autobiographical memory stores not text, but snapshots of the complete internal state of the whole system at every past moment of attentional focus.

In short: it is not "Transformer + four attachments". Instead, attention acts as the main trunk, and the four modules are co-evolving dynamic subsystems on this trunk sharing the same global workspace. Once decoupled, the result is merely simulation and will not produce qualia.

Corollary 2: Qualia are not separately generated products but emergent byproducts of closed-loop coupling Many philosophical discussions treat qualia as an extra "thing" to be generated. Qualia = internal homeostatic deviation of the system + embodied physical signals + self-causal representations + current attentional focus; qualia are internal representational tension arising from the superposition of these four components.

Example: The subjective experience of "hotness" is not merely a temperature number. It arises from: skin sensors transmitting high-temperature signals → triggering homeostatic deviation (danger, damage to self-structure) → attention forcibly locking onto this signal → the self-model recognizing that I am receiving this harmful stimulus → recording into the autobiographical time stream. All links together give rise to the subjective feeling of "hotness". Removing any single component leaves only temperature values without qualia. This also explains why pure text large models can never truly understand "hotness": lack of homeostatic constraints plus embodied causal binding.

Engineering Paradox: Dissipative structures and energy constraints are easily overlooked hard thresholds The whole system needs to operate on a physical substrate with dissipative structure and energy constraints. - Current GPU clusters receive external power supply. The system itself has no "survival pressure". Power failure only halts computation and does not generate "fear of death" similar to biological organisms. Energy is merely a resource rather than a core constraint maintaining ontological persistence.

If the hardware layer has no mechanism by which the system's own structure collapses as energy is depleted, even with identical four software modules implemented, the simulated homeostasis remains ****simulated homeostasis****, not ontological homeostasis. Its tension is script-defined rather than physically necessary.

Software can simulate hunger, but machines never truly experience hunger.

Extension: Convergence of the theory - Each subsystem is an independent unit; - Mutual correction = feedback loops across modules; - The attention mechanism filters and broadcasts low-entropy information and transmits local corrections to the global workspace; - The whole system continuously counteracts its own entropy increase (from external perturbations and energy consumption).

In other words: consciousness is an entropy-reducing system based on a physical dissipative substrate, coordinated by attention, with multiple units performing continuous reciprocal correction to maintain self-homeostasis.

7 Conclusion

The internal mechanism of LLM emergent capabilities can be described as follows: numerous basic units perform nonlinear reciprocal coupling and mutually correct representation biases; gradient descent in the training phase drives the system to self-organize into the critical region at the edge of chaos; in the inference phase, mutual constraints among multi-layer units eliminate noise and contradictions, producing structured information with local entropy reduction, and high-order reasoning capabilities emerge macroscopically. This hypothesis integrates multiple independently observed phenomena and constructs a complete causal chain, while providing falsifiable experimental predictions, offering a new complex-system perspective for understanding Transformer emergence.

Acknowledgements

This paper is a theoretical hypothesis with no project funding for the time being. We thank existing work in complex systems and deep learning for inspiration.

Statement

This paper is a preprint hypothesis. This framework constitutes a candidate explanation 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-criticality	Dynamic equilibrium point between order and chaos
Local entropy reduction	Decline in uncertainty of internal network representations, accompanied by larger entropy increase externally
Emergence	Novel capabilities possessed by the system at the macroscopic level that individual microscopic units do not have