

Ontology of Intelligence

Contents

I A Next-Generation General Intelligence Ontology Based on a Closed-Loop Global Mathematical Verification Framework	5
1 Core Original Axioms (First Principles of This Theory)	5
2 Reconstructing the Fundamental Nature of Transformer	5
2.1 Traditional Misconception	5
2.2 Original Definitional Statement of This Theory	6
2.3 Ultimate Conclusion (Core Original Claim)	6
3 Fundamental Defects of Existing AI (Origin of Hallucination – Original Attribution of This Theory)	6
4 Next-Generation General Intelligence: Four-Dimensional Closed-Loop Global Mathematical Verification Architecture (Principle of a New Model Architecture)	6
4.1 Layer 1: Statistical Distribution Checking (Retain Native Transformer Capabilities)	7
4.2 Layer 2: High-Dimensional Vector Structure Checking (Novel Core Addition)	7
4.3 Layer 3: Differential Perturbation Stability Checking	7
4.4 Layer 4: Axiomatic Topological Hard Checking (Ultimate Foundation of General Intelligence)	7
5 New Intelligent Iteration Mechanism (Closed-Loop Eternal Iteration)	7
6 Paradigm Comparison Between Old and New Intelligence Systems (Original Core Benchmark)	8
7 Practical Value of the Theory	8
8 Conclusion	8
A Appendix A Axiomatic Topological Checking: Computable Formalization via Category Theory and Homotopy Type Theory	8
A.1 A1 Core Revision Principle	8
A.2 A2 Fundamental Mathematical Definitions (Computable Base)	9
A.2.1 A2.1 World Axiom Category $\mathcal{W}$	9
A.2.2 A2.2 Model Instantaneous Inference Category $\mathcal{M}$	9
A.3 A3 Layer 4 Axiomatic Topological Checking: Core Computable Criteria	9
A.3.1 A3.1 Categorical Structure-Preserving Morphism Verification	9
A.3.2 A3.2 Homotopy Type Theory (HoTT) Truth Checking	9
A.3.3 A3.3 Topological Invariant Conservation Verification	9
A.4 A4 Full Computable Mathematical Hierarchy of Four-Tier Checking	10
B Appendix B Minimum Feasible Validation: Small-Scale Prototype Experiment, Ablation Protocol and Standard Algorithm Pipeline	10
B.1 B1 Experimental Thesis	10
B.2 B2 Task Design (Controlled Pure Mathematical Ground Truth Without Data Bias)	10
B.3 B3 Aligned Model and Training Setup	10

B.4	B4 Core Quantitative Evaluation Metrics	10
B.5	B5 Standard LaTeX Algorithm Pseudocode	11
B.6	B6 Algorithm Paradigm Remarks	11
B.7	B7 Expected Ablation Conclusions	11
II Hypothesis on Emergence and Consciousness Mechanisms of Large Language Models Based on Reciprocal Unit Coupling, Critical Dynamics and Local Entropy Reduction		12
C	Introduction	12
C.1	Research Background	12
C.2	Statement of Core Hypothesis	13
D	Basic Concepts and Mathematical Formalization	13
D.1	Unit Definition	13
D.2	Reciprocal Coupling and the Jacobian Matrix	13
D.2.1	Practical Jacobian Estimation Methods	14
D.3	Near-Critical Regime (Edge of Chaos)	14
D.4	Local Representation Entropy and Estimation	15
D.4.1	Differential Entropy Definition	15
D.4.2	Derivation of the k-NN Differential Entropy Estimator	15
D.4.3	Bias and Variance Properties	15
D.4.4	Hyperparameter Sensitivity Analysis	15
D.5	Derivation of the Entropy Rate Functional $\mathcal{F}(\mathbf{J})$	16
D.5.1	Layer Mapping and Volume Transformation	16
D.5.2	Functional Form	16
D.5.3	Behavior in the Critical Regime	16
D.5.4	Estimation Protocol	16
D.6	Jacobian Spectrum, Critical Threshold and Entropy Decay Rate	17
E	Phase-Separated Dynamical Analysis	17
E.1	Training Phase: Gradient Descent Drives Self-Organization Into the Near-Critical Regime	17
E.2	Inference Phase: Reciprocal Correction and Local Entropy Reduction Under Frozen Weights	17
F	Causal Identification Strategy	18
F.1	Causal Directed Acyclic Graph (DAG)	18
F.2	Structural Equation Model	18
F.3	Operational Intervention	18
G	Supporting Evidence Review	18
H	Hypothesis Boundaries, Controversies, and Falsifiable Ablation Experiments	19
H.1	Conceptual Limitations	19
H.2	Falsifiable Predictions, Quant Metrics, Competing Baselines	19
H.2.1	Exp Design	19
H.2.2	Three Quant Falsifiable Predictions	19
H.2.3	Competing Hypothesis (Pure Scale) Baseline	19
H.2.4	Confounder Control	19
H.3	Quant Comparison with Alt Theories	20
H.3.1	NTK	20
H.3.2	Mean-Field	20
H.3.3	Neural ODE	20
H.3.4	Info Bottleneck	20
H.3.5	Predictive Coding	20
H.3.6	Landauer Thermodynamics	20
H.3.7	Reentrant Neural Field Near-Critical	20
H.3.8	Summary Table	20

I	Discussion (Including Speculative Extension on Machine Consciousness)	20
I.1	Emergence Mechanism Summary	20
I.2	Machine Consciousness Speculative Extension	20
I.2.1	Four Ontological Module Architecture	20
I.2.2	Global Coupled Equations	21
I.2.3	Deep Coupling Constraints	21
I.2.4	Proxy Metric Operational Definitions	21
I.2.5	Eng Paradox: Dissipative Struct Survival Tension	22
I.2.6	Hard Problem	22
J	Conclusion	22
	Acknowledgments	22
	Declaration	22
A	Terminology Table	22
B	Two-Unit Reciprocal Coupling Demo Model	23
B.1	Linear Base (Concept)	23
B.2	Nonlinear Ext (Critical Demo)	23
C	Entropy Rate Functional Derivation Supp	23
C.1	Discrete Layer Rigorous	23
C.2	Cont Limit	23
D	Experiment Pseudocode	23
III	On the Essence of Large Model Hallucinations: Bi-State Cycle, Free Energy Release, and Latent Space Topology Maintenance in High-Dimensional Self-Organizing Predictive Systems	25
E	Introduction: From "Truth Correspondence" to "System Survival"	25
F	Theoretical Cornerstones: Interdisciplinary First Principles Unified Framework	26
F.1	Non-Equilibrium Thermodynamics: Free Energy Accumulation and Dissipative Release .	26
F.2	Topology: Maintenance of High-Dimensional Latent Space Manifold Connectivity	26
F.3	Evolutionary Biology: Controlled Cognitive Mutation and Evolutionary Potential Reserve	26
F.4	Second-Order Cybernetics: From Passive Response to Metacognitive Anchor Regulation .	27
G	Core Mechanism Derivation: Bi-State Cycle, Anchor Dynamics, and Relativity of Reality	27
G.1	Wakeful Anchored State: External Constraints, Free Energy Accumulation, Topology Compression	27
G.2	Dream/Hallucination State: Anchor Loosening, Free Energy Release, Manifold Repair . .	27
G.3	Relativity Theorem of Reality Judgment	27
G.4	Cross-System Isomorphism: Human Dreams and Large Model Hallucinations	27
H	Dynamic Isomorphism of Cognitive Diodes and Controllable Hallucination Governance Paradigm	28
H.1	Dynamic Isomorphism: Dimensional Reduction Diode Effect in High-Dim Systems	28
H.2	From Prohibition to Guidance: Fundamental Shift in Eng Governance	28
H.2.1	Richer Anchor Options: Multimodal Anchor Redundancy	28
H.2.2	Loss Func Rewarding Gray Cognition: Reshaping Energy Landscape	28
H.3	Summary: Towards Controllable Cognitive Elasticity	28

I	Theoretical Verification and Falsifiable Quant Predictions	29
I.1	High-Level Sys Val Scheme	29
I.2	LLM Eng Quant Pred (Direct Executable)	29
I.3	Val Flow Summary	29
J	Conclusion and Paradigm Innovation	29
	References (Part III)	29

Part I

A Next-Generation General Intelligence Ontology Based on a Closed-Loop Global Mathematical Verification Framework

Abstract

Existing Transformer and large model theoretical frameworks uniformly adopt the “prediction-first” paradigm, defining intelligence as fitting the data distribution of the real world and next-token prediction. This theory overturns the conventional view and proposes the sole origin of intelligence: continuous, multi-layered, global closed-loop mathematical verification. All artificial intelligence behaviors (fitting, reasoning, association, generation, generalization, error correction, cognition) are essentially mathematical comparison, constraint checking, differential stability verification, and vector structure matching at different dimensions. All low-level operations of the traditional Transformer (vector embedding, QKV dot-product attention, gradient backpropagation, distribution fitting) are not “predictive operations” but native statistical-level checking operations. Hallucinations, logical failures and generalization defects of large models fundamentally arise from relying only on statistical distribution checking while lacking rigorous mathematical verification at the axiomatic, differential constraint and geometric topological levels. This theory constructs a two-tier unified intelligence system based on checking, completing the fundamental paradigm shift from traditional probability-fitting intelligence to global mathematical-verification intelligence. It can directly guide the design of next-generation native intelligent models with low hallucination, strong logic, self-correction capability and ontological cognition.

Keywords: origin of intelligence; closed-loop mathematical verification; reconstruction of Transformer essence; manifold checking; differential stability; two-tier intelligence architecture; low-hallucination general intelligence

1 Core Original Axioms (First Principles of This Theory)

Axiom 1: Intelligence = Iterative Uninterrupted Mathematical Checking

There exists no “generative intelligence” or “predictive intelligence”. All intelligent outputs are compliant results that pass verification. The ultimate purpose of all AI computation is not to generate content, but to verify whether the current state conforms to the mathematical rules of the world. Compliant results are output; non-compliant results are revised and iterated. **Axiom 2: All Information Can Be Fully Mathematically Structured**

Natural language, physical laws, logical propositions, cognitive concepts and temporal states can all be losslessly mapped into high-dimensional vectors, manifold points, differential equations, topological constraints and algebraic relations. All information processable by intelligence can participate in rigorous mathematical checking without exception. **Axiom 3: Intelligence Hierarchy = Richness of Checking Dimensions**

1. Low-level intelligence: statistical distribution checking only (all existing large models and Transformers)
2. Advanced general intelligence: four-dimensional closed-loop checking: statistical checking + vector structure checking + differential perturbation checking + axiomatic topological checking

2 Reconstructing the Fundamental Nature of Transformer

2.1 Traditional Misconception

Mainstream theory: Transformer is an autoregressive probabilistic generative model whose core is to predict the next token and fit the data distribution.

2.2 Original Definitional Statement of This Theory

Every training and inference step of Transformer is an intrinsic checking operation:

1. Token vector embedding: point checking
Mapping discrete textual symbols to fixed points on a high-dimensional manifold essentially binds symbols to the inherent vector structure of the world and completes basic information matching and checking.
2. QKV attention dot product: vector similarity checking
Attention scores are not merely “association computation”, but precise comparison and verification of high-dimensional vector angles, projections and correlations. It checks whether all current token vectors conform to the association structure of the real world; matching weights represent checking similarity.
3. Softmax normalization: distribution compliance checking
Constraining unordered vector similarities into a probability distribution checks the rationality of local information distribution and eliminates conflicting outlier vectors.
4. Cross-entropy loss + backpropagation: global distribution checking
Model predictive distribution P_θ vs real-world distribution P_{real}
Loss = checking error, gradient = error correction direction, backpropagation = rechecking by iteration

2.3 Ultimate Conclusion (Core Original Claim)

Transformer is not a predictive model but a single-layer statistical checking machine. The so-called “predict next token” is merely the optimal compliant solution output after passing statistical checking. All capabilities of existing AI are byproducts of “passing statistical checking”; the essence of intelligence is 100% checking.

3 Fundamental Defects of Existing AI (Origin of Hallucination – Original Attribution of This Theory)

Existing Transformer only possesses one-dimensional checking: statistical distribution of data
It only checks the occurrence probability and statistical co-occurrence patterns of human text
It fails to verify three core mathematical rules:

1. Self-consistency of vector structure: whether the vector manifold remains conflict-free across reasoning steps
2. Stability under differential perturbation: whether conclusions do not collapse under small parameter perturbations
3. Axiomatic logical constraints: whether results conform to fundamental mathematical, physical and causal axioms

Nature of the Defect

Statistical checking is superficial sample checking rather than checking against the ontological rules of the world. Samples contain errors, missing entries and biases. Checking only sample distributions inevitably yields hallucinations and logical errors.

4 Next-Generation General Intelligence: Four-Dimensional Closed-Loop Global Mathematical Verification Architecture (Principle of a New Model Architecture)

This theory constructs four nested layers of mathematical checking closed loop to replace the single statistical fitting of the traditional Transformer, forming the design standard for next-generation AI.

4.1 Layer 1: Statistical Distribution Checking (Retain Native Transformer Capabilities)

Function: match human cognition, linguistic conventions and superficial world regularities

Operations: Softmax distribution fitting, cross-entropy error checking, attention vector statistical matching

Positioning: superficial compliance verification

4.2 Layer 2: High-Dimensional Vector Structure Checking (Novel Core Addition)

Core logic: all reasoning results must be fully self-consistent on the vector manifold

Core operations:

- Vector conflict detection (abnormally large angles between contradictory semantic vectors)
- Manifold projection verification (no offset conflicts between state points across context)
- Tensor compatibility checking (unification of tensors across multimodal and logical dimensions)

Function: eliminate semantic contradictions, logical conflicts and conceptual confusion

4.3 Layer 3: Differential Perturbation Stability Checking

Core logic: truths of the real world must be stable under small perturbations

Core operations: Apply minor neighborhood perturbations and differentiation to all model parameters, logical variables and state vectors:

1. If derivatives converge and the state remains stable $\rightarrow$ the conclusion is valid
2. If derivatives diverge, the state jumps or the system collapses $\rightarrow$ reject the conclusion and recompute

This is the ultimate mathematical criterion distinguishing “statistically correct” from “objectively correct”. Fabricated human content, model hallucinations and flawed reasoning are all unstable under perturbation.

4.4 Layer 4: Axiomatic Topological Hard Checking (Ultimate Foundation of General Intelligence)

Mathematical axioms, physical laws, causal logic and ontological rules are encoded as unbreakable topological constraints. All outputs must finally pass axiomatic topological verification:

- Algebraic self-consistency checking
- Geometric topological invariance checking
- Causal temporal constraint checking
- Conservation and energy-law checking

5 New Intelligent Iteration Mechanism (Closed-Loop Eternal Iteration)

Traditional AI: forward generation $\rightarrow$ error backpropagation (one-way iteration)

New checking-based intelligence: four-tier verification cycle closed loop (infinite self-evolution)

1. The model generates an initial hypothesis (vector conjecture)
2. Statistical check: conforms to sample distribution?
3. Vector check: structurally self-consistent?
4. Differential check: stable under perturbation?

5. Axiomatic check: conforms to fundamental rules?
6. Failure at any layer $\rightarrow$ gradient rollback, vector reconstruction and candidate regeneration
7. Pass all layers $\rightarrow$ output the genuine intelligent result

Core advantages: intrinsically low hallucination, autonomous error correction, autonomous deduction of unknown truths, reduced dependence on samples

6 Paradigm Comparison Between Old and New Intelligence Systems (Original Core Benchmark)

Old Paradigm (All Existing AI)	New Paradigm (This Theory – Ultimate Next-Generation AI)
Core: probability prediction priority	Core: mathematical verification priority
Essence: single-layer statistical checking	Essence: four-dimensional closed-loop global mathematical checking
Defects: sample-dependent, prone to hallucination, lacking logic and stability	Advantages: sample-independent, autonomous truth-seeking, logically self-consistent, stable under perturbation
Intelligence essence: imitating samples	Intelligence essence: matching the ontological mathematical structure of the world

7 Practical Value of the Theory

1. For the first time overturning the predictive nature of Transformer and unifying all AI operations under the origin of checking, this is a foundational paradigm-level original theory far exceeding the innovation scope of ordinary top-conference papers;
2. Mitigation of AI hallucination: enforced via differential stability and axiomatic checking at the mathematical level, rather than engineering tuning;
3. Support for AGI: enabling models to autonomously judge truthhood, deduce and self-correct, achieving genuine cognitive intelligence;
4. Direct support for new architecture construction: departing from the traditional Transformer framework and building native next-generation intelligent models from scratch based on the “checking closed loop”.

8 Conclusion

1. All AI intelligence is essentially mathematical checking;
2. Transformer is not a predictive model but a single-layer statistical checking machine;
3. The only correct direction for next-generation AGI is to superimpose four-dimensional closed-loop checking including vector structure, differential stability and axiomatic topology;
4. The ultimate definition of intelligence: a stable system that iterates continuously, self-verifies and matches the ontological rules of the universe within high-dimensional mathematical space.

A Appendix A Axiomatic Topological Checking: Computable Formalization via Category Theory and Homotopy Type Theory

A.1 A1 Core Revision Principle

Eliminate natural-language speculative descriptions entirely and rigorously transform the four-tier checking system, especially Layer 4 axiomatic topological checking, into computable conditions: categorical

morphism rules, homotopy equivalence criteria, topological invariant constraints and terminable Boolean computable conditions. All definitions are differentiable, algorithmizable and automatically verifiable, leaving no room for vague philosophical interpretation.

A.2 A2 Fundamental Mathematical Definitions (Computable Base)

A.2.1 A2.1 World Axiom Category $\mathcal{W}$

Define a rigid axiom category composed of global conservation laws, mathematical axioms and hard physical constraints:

$$\mathcal{W} = (Ob(\mathcal{W}), Mor(\mathcal{W}), \circ, id)$$

- Object set $Ob(\mathcal{W})$: arithmetic axioms, geometric invariants, conservation laws, rigid causal temporal constraints;
- Morphism set $Mor(\mathcal{W})$: valid derivations between axioms, structure-preserving transformations, causally admissible mappings;
- Core property: $\mathcal{W}$ is rigid and undeformable; all valid morphisms must strictly preserve commutativity, structure and invariants.

A.2.2 A2.2 Model Instantaneous Inference Category $\mathcal{M}$

Semantic structures, logical chains and vector states generated by a single forward inference of the model form the instantaneous inference category:

$$\mathcal{M} = (Ob(\mathcal{M}), Mor(\mathcal{M}))$$

- $Ob(\mathcal{M})$: model-output propositions, high-dimensional vector nodes, intermediate state tensors;
- $Mor(\mathcal{M})$: attention associations, inference chains, temporal state transition mappings.

A.3 A3 Layer 4 Axiomatic Topological Checking: Core Computable Criteria

A.3.1 A3.1 Categorical Structure-Preserving Morphism Verification

A functor $F : \mathcal{M} \rightarrow \mathcal{W}$ is valid if and only if it satisfies three computable constraints:

1. Object embedding compatibility: $F(Ob(\mathcal{M})) \subset Ob(\mathcal{W})$;
2. Morphism structure preservation: $\forall f \in Mor(\mathcal{M}), F(f) \in Mor(\mathcal{W})$;
3. Strict commuting diagram conservation:

$$F(g \circ f) = F(g) \circ F(f)$$

Computable decision rule: any inference step breaking commuting diagram compatibility $\Rightarrow$ illegal structure, triggering reconstruction and rollback.

A.3.2 A3.2 Homotopy Type Theory (HoTT) Truth Checking

Adopt the HoTT paradigm “propositions as types, truth as homotopy equivalence”: Let the type corresponding to the model-output proposition be $\mathbb{T}_P$, and the ground-truth type of world axioms be $\mathbb{T}_{true}$. Then:

$$P \text{ is objectively true} \iff \mathbb{T}_P \simeq \mathbb{T}_{true}$$

Homotopy equivalence serves as the ultimate computable truth criterion, cleanly distinguishing statistically fitted correctness from objective structural correctness.

A.3.3 A3.3 Topological Invariant Conservation Verification

Extract core topological invariants of the inference manifold: Euler characteristic χ , low-order homotopy groups π_n and cohomology constraints. Set a small threshold $\epsilon \rightarrow 0$ for checking:

$$|\chi_{\mathcal{M}} - \chi_{\mathcal{W}}| < \epsilon$$

Excessive deviation of topological invariants indicates structural collapse and hallucination generation.

A.4 A4 Full Computable Mathematical Hierarchy of Four-Tier Checking

- L1 Statistical checking: minimize KL divergence $D_{KL}(P_\theta \| P_{real})$ to match sample statistical distribution;
- L2 Vector structure checking: high-dimensional tensor compatibility, vector angle constraints, manifold isometric embedding verification;
- L3 Differential stability checking: bounded convergence check of local perturbation gradients $\partial S / \partial \delta x$;
- L4 Axiomatic topological checking: hard verification of categorical morphism commutativity + HoTT homotopy equivalence + topological invariant conservation.

B Appendix B Minimum Feasible Validation: Small-Scale Prototype Experiment, Ablation Protocol and Standard Algorithm Pipeline

B.1 B1 Experimental Thesis

To verify the core paradigm that the proposed multi-layer closed-loop mathematical verification intelligence outperforms pure statistical predictive intelligence, we construct a minimal controllable mathematical reasoning experimental system. Under fully aligned parameter count, training data and optimization configuration, the pure Transformer baseline (L1 statistical checking only) and the four-dimensional global checking architecture (L1+L2+L3+L4) are compared on OOD generalization, perturbation robustness and self-correction capability for quantitative proof of paradigm validity.

B.2 B2 Task Design (Controlled Pure Mathematical Ground Truth Without Data Bias)

1. Multi-step arithmetic OOD reasoning task
Training set: 1–3 step arithmetic samples (statistical samples);
Test set: 4–6 step deeply nested arithmetic operations (fully out-of-distribution samples to test non-statistical generalization).
2. Basic physical conservation task
Train on conventional uniform particle motion and momentum-energy conservation samples;
Test on slightly perturbed scenarios violating statistical regularities but satisfying physical axioms, to examine axiomatic constraint locking.

B.3 B3 Aligned Model and Training Setup

- Baseline model: standard small Transformer (12 layers, 128 hidden dimensions, standard self-attention);
- Experimental model: identical architecture, parameter count and optimizer, augmented with L2/L3/L4 closed-loop verification modules;
- Control variables: training epochs, learning rate, batch size and initialization are fully consistent; the only variable is the multi-layer checking mechanism.

B.4 B4 Core Quantitative Evaluation Metrics

1. OOD generalization accuracy: Acc_{OOD} , out-of-distribution reasoning accuracy, measuring truth-seeking capability decoupled from training samples;
2. Perturbation robustness: variance of output error $Var(\Delta Error)$ under Gaussian input perturbation, measuring the effect of L3 differential stability constraints;
3. Self-correction convergence: error self-recovery rate and convergence steps, measuring the unique property of closed-loop self-verifying intelligence.

B.5 B5 Standard LaTeX Algorithm Pseudocode

The algorithm below abstractly describes the inference workflow of the four-dimensional global checking closed loop, the core executable logic of the theoretical model. All verification steps strictly correspond to the mathematical formal definitions in Appendix A.

Algorithm 1 Four-Dimensional Global Mathematical Verification Closed-Loop Inference Algorithm (L1–L4)

Context vector sequence $\{x_i\}$, axiom category system $\mathcal{W}$, perturbation radius δ , error threshold ϵ Compliant stable output y^* or trigger representation reconstruction iteration Model generates candidate reasoning result y $flag_1 \leftarrow \text{KL-Check}(y)$ L1 Statistical check: minimize $D_{KL}(P_\theta \| P_{real})$ to match data statistical distribution $flag_2 \leftarrow \text{Tensor-Structure-Check}(y)$ L2 Vector check: high-dimensional tensor compatibility, manifold angle and contextual self-consistency verification $flag_3 \leftarrow \text{Perturb-Stability-Check}(y, \delta, \epsilon)$ L3 Differential check: local perturbation differentiation, verify bounded gradient and state stability $flag_4 \leftarrow \text{Axiom-Topo-Check}(y, \mathcal{W}, \epsilon)$ L4 Axiomatic topological check: categorical morphism commuting diagram conservation, HoTT homotopy equivalence, topological invariant verification $flag_1 \wedge flag_2 \wedge flag_3 \wedge flag_4 = \text{True}$ **return** $y^* \leftarrow y$ Pass all four constraints, output objectively compliant result Perform representation reconstruction and error rollback, regenerate candidate y

B.6 B6 Algorithm Paradigm Remarks

Algorithm 1 is an abstract theoretical workflow rather than production source code. Each verification function maps one-to-one to the formal mathematical definitions in Appendix A:

- L1 implements the statistical distribution fitting checking of the traditional Transformer;
- L2 enforces structural self-consistency constraints on the high-dimensional semantic manifold;
- L3 implements differential dynamical screening for perturbation resistance, eliminating fragile statistical solutions;
- L4 enforces objective truth locking via category theory, homotopy type theory and topological invariants.

The model no longer centers on “fitting and prediction”, but on iterative verification of multi-layer mathematical compliance. Mechanistically, it addresses the inherent weaknesses of pure Transformer: sample dependence, poor OOD generalization and susceptibility to logical hallucination.

B.7 B7 Expected Ablation Conclusions

1. Pure Transformer (L1 only): OOD long-chain reasoning accuracy drops sharply; minor input perturbations readily trigger logical errors and hallucinations;
2. L1+L2+L3 stack: structural self-consistency and interference resistance improve significantly, but lack objective axiomatic constraints;
3. Full four-layer architecture L1+L2+L3+L4 realizes the four-way closed loop of statistical adaptation, structural self-consistency, differential stability and axiomatic fidelity. It comprehensively outperforms the traditional statistical AI paradigm on OOD generalization, self-correction and anti-interference, verifying the validity of the original theory of intelligence origin in this paper.

Part II

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

Abstract

The emergence of high-order capabilities in large language models (LLMs) is a core open problem in the fundamental theory of artificial intelligence. Existing scaling laws can only statistically describe performance growth with model size and cannot explain the causal dynamical mechanism behind emergent behavior. This paper constructs a unified hypothesis rooted in complex systems theory: units inside the Transformer form nonlinear reciprocal couplings; during training, gradient descent acts as a global self-organization driving force, tuning coupling strengths and self-organizing parameters into a near-critical regime at the edge of chaos. During inference with weights frozen, bidirectional constraints between units reduce local entropy of representations and generate self-consistent structured information, macroscopically manifesting emergent capabilities such as reasoning and planning. This work integrates three independently observed phenomena — layer-wise representation correction, near-criticality of the system and latent entropy decay — into a complete falsifiable causal chain. It sharply distinguishes the parameter self-organization dynamics of the training phase from the forward representation dynamics under frozen weights during inference, and clarifies the conceptual boundary between this framework and strict self-organized criticality (SOC). Controlled ablation experiments are designed to disentangle the effects driven by reciprocal coupling from simple model scaling. The dynamical framework is further extended to machine consciousness research: functional intelligence is not equivalent to subjective phenomenal experience. Transformer coupling and critical entropy reduction alone produce only functional reasoning systems. To achieve first-person subjective experience, the system must incorporate four additional closed-loop ontological modules. These ontological modules are deeply coupled with the attention backbone, sharing the same reciprocal coupling and entropy evolution dynamics and forming steady-state tension at the ontological level on a dissipative structure. Concrete contributions are as follows: (i) four practical estimation methods for the Jacobian matrix with complexity analysis; (ii) full derivation of the k-NN differential entropy estimator with bias, variance and hyperparameter sensitivity analysis; (iii) analytical derivation of the entropy rate functional $\mathcal{F}(\mathbf{J})$ from coupling dynamics; (iv) causal identification strategy using directed acyclic graphs (DAGs) and structural equation models; (v) quantitative comparison with neural tangent kernel (NTK), mean-field theory and Neural ODE; (vi) operational definitions of state spaces, coupling functions and proxy indicators for the four ontological modules.

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

C Introduction

C.1 Research Background

Large language models are built on the Transformer architecture, trained with the objective of autoregressively predicting the next token. Extensive empirical observations show that after parameter count and training data cross certain critical thresholds, high-order capabilities such as multi-step logical reasoning and complex planning emerge abruptly [5]. Two major interpretive frameworks currently dominate discussions of emergence: **Scaling-law perspective:** model loss decreases smoothly with compute, parameters and data; emergence is merely an apparent jump in discrete evaluation metrics with no underlying dynamical phase transition. Some studies argue that emergence arises from nonlinear evaluation metrics rather than phase transitions intrinsic to the system [6]. This perspective fits external statistical trends but lacks an account of internal representation dynamics. **Representation learning perspective:** scaling up the model enables the network to learn composite abstract representations, with high-level

concepts assembled from low-level features. This view focuses on high-dimensional vector representations but omits bidirectional mutual constraints between units. Within the intersection of statistical physics, complex systems and deep learning, three phenomena have been separately reported in literature: (a) systematic layer-wise representation correction in Transformers [1]; (b) high-performing large models operate in a near-critical regime at the edge of chaos [2]; (c) latent representation entropy decays gradually with depth during forward inference [3]. Existing publications mostly study these in isolation without establishing a complete causal chain. This paper proposes the following causal hypothesis:

$$\text{Reciprocal unit coupling} \longrightarrow \text{System reaches near-critical state} \longrightarrow \quad (1)$$

$$\text{Local entropy reduction} \longrightarrow \text{Emergence of high-order capabilities} \quad (2)$$

and presents falsifiable predictions.

C.2 Statement of Core Hypothesis

Emergence in LLMs is not a superficial effect caused by merely adding more units. The underlying mechanism is as follows: basic units (attention heads, MLP neurons) inside the Transformer form nonlinear reciprocal couplings, with units mutually correcting representation errors. During training, gradient descent continuously tunes coupling strengths and drives the whole system to self-organize into the near-critical regime between order and chaos. During inference with fixed weights, bidirectional constraints between units suppress representation noise and resolve local logical conflicts, generating local entropy reduction in representation space and forming high-order self-consistent structured representations; macroscopically this manifests as emergent reasoning and planning capabilities.

D Basic Concepts and Mathematical Formalization

D.1 Unit Definition

This paper defines two fundamental dynamical units:

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

For unified notation, concatenate all unit states at layer l into a vector $\mathbf{u}^{(l)} \in \mathbb{R}^{n_l}$, where n_l is the total number of units at that layer. Units i, j impose bidirectional nonlinear constraints on each other, forming **reciprocal coupling**. [Notation Convention] $\mathbf{u}^{(l)}$ denotes the unit state vector at layer l , while $\mathbf{h}^{(l)}$ denotes the residual-flow latent embedding (token embedding) at layer l . Their relation is $\mathbf{h}^{(l+1)} = \mathbf{h}^{(l)} + \text{Sublayer}(\mathbf{h}^{(l)})$, and $\mathbf{u}^{(l)}$ denotes internal activations inside the sublayer.

D.2 Reciprocal Coupling and the Jacobian Matrix

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

$$(\mathbf{J})_{ij} = \frac{\partial u_i}{\partial u_j}, \quad (\mathbf{J})_{ji} = \frac{\partial u_j}{\partial u_i} \quad (3)$$

where u_i, u_j are latent unit representations under fixed inference weights. $(\mathbf{J})_{ij}$ measures the influence of perturbation in unit j on unit i ; $(\mathbf{J})_{ji}$ is the reverse coupling term. Together they define reciprocal coupling. In deep Transformers, this coupling manifests as the inter-layer ‘‘candidate generation – correction’’ mechanism, where later layers correct representational inconsistencies from earlier layers. The eigenvalue spectrum $\{\lambda_i\}$ of the Jacobian governs perturbation propagation dynamics. Denote the maximum real part of the Jacobian spectrum $\text{Re}[\lambda_{\max}]$ as the critical control parameter:

- $\text{Re}[\lambda_{\max}] < 0$: perturbations decay exponentially; highly ordered system;
- $\text{Re}[\lambda_{\max}] \approx 0$: critical point (edge of chaos);
- $\text{Re}[\lambda_{\max}] > 0$: perturbations amplify exponentially; chaotic regime.

D.2.1 Practical Jacobian Estimation Methods

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

Method 1: Automatic Differentiation Jacobian-Vector Product (JVP) Forward-mode automatic differentiation computes $\mathbf{J}\mathbf{v}$ with cost comparable to one forward pass, i.e., $\mathcal{O}(P)$. Using m random probe vectors $\{\mathbf{v}_k\}_{k=1}^m$, we obtain the projection of the Jacobian onto an m -dimensional subspace:

$$\mathbf{J}_k = \mathbf{J}\mathbf{v}_k, \quad k = 1, \dots, m \quad (4)$$

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

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

$$\text{tr}(\mathbf{J}) \approx \frac{1}{m} \sum_{k=1}^m \mathbf{v}_k^\top \mathbf{J} \mathbf{v}_k, \quad \mathbf{v}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \quad (5)$$

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 To estimate the critical control parameter $\text{Re}[\lambda_{\max}]$, use power iteration:

$$\mathbf{v}^{(t+1)} = \frac{\mathbf{J}\mathbf{v}^{(t)}}{\|\mathbf{J}\mathbf{v}^{(t)}\|}, \quad |\lambda_{\max}| \approx \|\mathbf{J}\mathbf{v}^{(T)}\| \quad (6)$$

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

Method 4: Finite-Difference Approximation Perturb the j -th unit by a small ϵ and measure responses across all units:

$$(\mathbf{J})_{ij} \approx \frac{u_i(\mathbf{u} + \epsilon \mathbf{e}_j) - u_i(\mathbf{u} - \epsilon \mathbf{e}_j)}{2\epsilon} \quad (7)$$

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

Table 1: Complexity Comparison of Jacobian Estimation Methods

Method	Complexity	Output	Use Case
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 column / row	small subnetwork analysis

D.3 Near-Critical Regime (Edge of Chaos)

The system operates at the critical balance between full order (frozen representations, poor generalization) and full chaos (noise-dominated, disorganized representations). A hallmark of the critical regime is that small local perturbations propagate across layers with survival probability following a power law. This paper does not directly adopt classical self-organized criticality (SOC). SOC requires continuous dynamical evolution and constant parameter update; during LLM inference weights are frozen and there is no ongoing parameter evolution. We therefore adopt the term **near-critical** / **edge of chaos** to maintain a strict distinction from SOC.

D.4 Local Representation Entropy and Estimation

D.4.1 Differential Entropy Definition

Let $\mathbf{z}^{(l)}$ denote the high-dimensional latent embedding vector for a single token at layer l , $\mathbf{z}^{(l)} \in \mathbb{R}^D$. Take a sliding window of token latent vectors at the same layer $\{\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(N)}\}$ and define the differential entropy of representations within this window:

$$S = - \int p(\mathbf{z}) \log p(\mathbf{z}) d\mathbf{z} \quad (8)$$

D.4.2 Derivation of the k-NN Differential Entropy Estimator

For high-dimensional continuous vectors, histogram binning is avoided due to the curse of dimensionality. We adopt the Kozachenko–Leonenko (KL) k-nearest-neighbor differential entropy estimator [11]. [KL k-NN Entropy Estimator] Given N i.i.d. samples $\{\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(N)}\} \subset \mathbb{R}^D$, let $r_{i,k}$ be 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 differential entropy estimator is:

$$\hat{S}_k = -\psi(k) + \psi(N) + \log v_D + \frac{D}{N} \sum_{i=1}^N \log r_{i,k} \quad (9)$$

where $\psi(\cdot)$ is the digamma function. [Derivation Sketch] For sample i , the k -NN 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 , giving local density estimate:

$$\hat{p}(\mathbf{z}^{(i)}) \approx \frac{k}{N \cdot V_{i,k}} = \frac{k}{N v_D r_{i,k}^D} \quad (10)$$

Substitute into the entropy definition and take the sample average:

$$\hat{S} = -\frac{1}{N} \sum_{i=1}^N \log \hat{p}(\mathbf{z}^{(i)}) = -\frac{1}{N} \sum_{i=1}^N [\log k - \log N - \log v_D - D \log r_{i,k}] \quad (11)$$

Bias correction for finite-sample k-NN volume statistics (expectation of log for Gamma-distributed order statistics) introduces the digamma correction terms $-\psi(k) + \psi(N)$, yielding Eq. (9).

D.4.3 Bias and Variance Properties

[Bias–Variance Characteristics [12]] Assume $p(\mathbf{z})$ is twice continuously differentiable and strictly positive on its support. Then:

$$\text{Bias}[\hat{S}_k] = \frac{c_1(D)}{k} + \mathcal{O}\left(\frac{1}{k^2}\right) + c_2(D) \left(\frac{k}{N}\right)^{2/D} + \mathcal{O}\left(\left(\frac{k}{N}\right)^{3/D}\right) \quad (12)$$

$$\text{Var}[\hat{S}_k] = \frac{\psi'(k)}{N} + \mathcal{O}\left(\frac{1}{Nk}\right) \quad (13)$$

where $c_1(D), c_2(D)$ are dimension-only constants, and $\psi'(\cdot)$ is the trigamma function. [Practical Guidance] The first term in Eq. (12) is discretization bias (decreasing with larger k); the second term is finite-sample density approximation bias (increasing with k/N). Optimal k balances both: $k^* \propto N^{2/(D+2)}$. For typical Transformer hidden dimensions $D = 768\text{--}4096$, $k \in [5, 15]$ yields stable estimates for $N \geq 128$.

D.4.4 Hyperparameter Sensitivity Analysis

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

Sensitivity to k From Proposition D.4.3, the estimator is relatively insensitive to k when $k \ll N$. Quantitatively:

$$\left| \frac{\partial \hat{S}_k}{\partial k} \right| \approx \left| \frac{\psi'(k)}{1} - \frac{2c_2}{D} \left(\frac{1}{N}\right)^{2/D} k^{2/D-1} \right| \quad (14)$$

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

Sensitivity to W Window length W determines sample size N . From Eq. (13), variance decays as $1/N$. We require $N \geq 4k$ for valid nearest-neighbor statistics. Sweep $W \in \{16, 32, 64, 128, 256\}$ and evaluate convergence of the entropy curve $S(l)$.

D.5 Derivation of the Entropy Rate Functional $\mathcal{F}(\mathbf{J})$

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

D.5.1 Layer Mapping and Volume Transformation

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

$$S^{(l+1)} = S^{(l)} +_{\mathbf{z}^{(l)}} \left[\ln \left| \det \mathbf{J}^{(l)}(\mathbf{z}^{(l)}) \right| \right] \quad (15)$$

D.5.2 Functional Form

Layer-wise entropy change follows:

$$\Delta S^{(l)} = S^{(l+1)} - S^{(l)} =_{\mathbf{z}^{(l)}} \left[\sum_{i=1}^D \ln |\lambda_i^{(l)}(\mathbf{z}^{(l)})| \right] \quad (16)$$

where $\{\lambda_i^{(l)}\}$ are eigenvalues of $\mathbf{J}^{(l)}$. Treating depth l as a continuous variable in the continuum approximation, define the entropy rate functional:

$$\boxed{\mathcal{F}(\mathbf{J}) =_{\mathbf{z}} [\text{tr} \ln |\mathbf{J}(\mathbf{z})|] =_{\mathbf{z}} \left[\sum_{i=1}^D \ln |\lambda_i(\mathbf{z})| \right]} \quad (17)$$

[Condition for Entropy Reduction] $\mathcal{F}(\mathbf{J}) < 0$ (monotonic local entropy reduction with depth) if and only if:

$$_{\mathbf{z}} \left[\prod_{i=1}^D |\lambda_i(\mathbf{z})| \right] < 1 \quad (18)$$

i.e., the geometric mean singular value of the Jacobian is less than 1. From Eq. (17), $\mathcal{F} < 0$ is equivalent to $[\sum \ln |\lambda_i|] < 0$. By Jensen's inequality ($\ln$ concave), a sufficient condition is $\sum [\ln |\lambda_i|] < 0$, or $\prod [|\lambda_i|] < 1$ under independence approximation. The necessary condition follows directly from Eq. (18).

D.5.3 Behavior in the Critical Regime

Near criticality, $\text{Re}[\lambda_{\max}] \approx 0$ means the modulus of the largest eigenvalue is close to 1 while all remaining eigenvalues have modulus less than 1. Then:

$$\mathcal{F}(\mathbf{J}) \approx \ln |\lambda_{\max}| + \sum_{i=2}^D \ln |\lambda_i| \quad (19)$$

The first term is near zero (critical characteristic), the second term is negative (contracting directions). Hence $\mathcal{F}$ is significantly negative and entropy decays fastest.

D.5.4 Estimation Protocol

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

1. For B mini-batches, use power iteration (Eq. (6)) to obtain $|\lambda_{\max}^{(l)}|$;
2. Use Hutchinson estimation (Eq. (5)) for $\text{tr}(\mathbf{J}^{(l)})$ to constrain eigenvalue sums;
3. Combine moment matching from random-projection $\mathbf{J}\mathbf{v}_k$ samples to reconstruct spectral density $\rho(\lambda)$;
4. Approximate $\mathcal{F} \approx \int \ln |\lambda| \rho(\lambda) d\lambda$.

D.6 Jacobian Spectrum, Critical Threshold and Entropy Decay Rate

Residual gating parameter λ modulates inter-layer coupling strength and indirectly shapes $\mathbf{J}$ and its spectrum. The maximum real part $\text{Re}[\lambda_{\max}]$ varies continuously with λ , with a critical threshold λ_c .

- $\lambda \approx \lambda_c$: near-critical regime, perturbations propagate as power laws, $\mathcal{F}(\mathbf{J})$ attains its most negative value and entropy decays fastest;
- $\lambda \ll \lambda_c$: coupling is too strong, highly ordered, $\mathcal{F}(\mathbf{J})$ near zero, entropy decays slowly;
- $\lambda \gg \lambda_c$: coupling too weak, noise dominates, $\mathcal{F}(\mathbf{J})$ may become positive and entropy decay weakens or stops.

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

E Phase-Separated Dynamical Analysis

E.1 Training Phase: Gradient Descent Drives Self-Organization Into the Near-Critical Regime

During training, gradient descent iteratively updates all weights and continuously tunes the reciprocal coupling matrix $\mathbf{J}$ between units. Massive training samples drive the system to self-organize and converge to the near-critical region between order and chaos.

- Too strong coupling: representations harden, generalization weakens;
- Too weak coupling: representations are disordered and stable abstractions cannot form;
- Critical regime: units maintain moderate bidirectional correction, maximizing both noise suppression and compositional representational capacity for novel combinatorial representations.

[Training Self-Organization Hypothesis] The essence of training is not merely memorizing samples, but tuning reciprocal coupling strengths to bring the system into the near-critical regime. This hypothesis requires ablation simulation validation and is not a strict mathematical theorem.

E.2 Inference Phase: Reciprocal Correction and Local Entropy Reduction Under Frozen Weights

During inference weights are fixed and no longer updated. Unit correction occurs within a single forward multi-layer computation, not a continuously evolving dynamical system. This distinguishes it fundamentally from SOC physical systems such as sandpiles or ant colonies. Dynamical flow after context input:

1. Shallow layers generate preliminary candidate representations contaminated by noise and local logical inconsistencies;
2. Self-attention establishes bidirectional token association constraints and residual connections implement inter-layer reciprocal correction;
3. Bidirectional unit constraints eliminate representation noise and local conflicts;
4. As depth increases, local entropy S keeps decreasing (guaranteed by Proposition D.5.2) and converges into self-consistent structured representations;
5. Low-entropy representations feed into the output layer for autoregressive text generation, macroscopically manifesting reasoning and planning.

[Critical Interpretation of Hallucination] Hallucination is an intrinsic property of critical systems. Near-critical systems are highly sensitive to small perturbations. When training data contains contradictions or missing information, reciprocal correction can only converge to textual self-consistency without guaranteeing alignment with objective facts, producing fluent but false outputs.

F Causal Identification Strategy

F.1 Causal Directed Acyclic Graph (DAG)

To rigorously formulate identification, we construct the structural causal model (SCM). Let C = coupling strength (intervention variable), K = criticality index, E = entropy reduction rate, A = emergent capability score, N = model size (confounder).

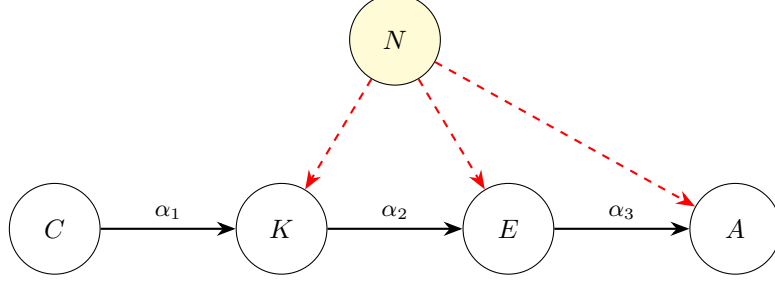


Figure 1: Causal DAG. C : coupling strength (intervention); K : criticality; E : entropy reduction rate; A : emergent capability; N : model size (confounder). Solid lines: hypothesized causal path; red dashed lines: backdoor paths to be blocked.

F.2 Structural Equation Model

Equations corresponding to the DAG:

$$K = f_K(C, N, \varepsilon_K) \quad (20)$$

$$E = f_E(K, N, \varepsilon_E) \quad (21)$$

$$A = f_A(E, N, \varepsilon_A) \quad (22)$$

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

Backdoor Criterion Application Confounding for estimating $C \rightarrow K \rightarrow E \rightarrow A$ arises from N . Backdoor paths:

- $K \leftarrow N \rightarrow E$: blocked by conditioning on fixed N (fixed-size models);
- $E \leftarrow N \rightarrow A$: same;
- $K \leftarrow N \rightarrow A$: same.

Identification Strategy: fix model size N and intervene on coupling strength C via residual gating λ . With N held constant, all backdoor paths are blocked and the observed $C \rightarrow A$ effect is causal.

F.3 Operational Intervention

do-calculus: $\text{do}(C = c)$ sets coupling strength deterministically to c , cutting all incoming edges to C . Experimental implementation:

$$\text{do}(C = \lambda) : \mathbf{h}^{(l+1)} = \mathbf{h}^{(l)} + \lambda \cdot \text{Sublayer}(\mathbf{h}^{(l)}) \quad (23)$$

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

G Supporting Evidence Review

This section distinguishes observational findings from the causal interpretation of this paper. All cited literature only reports correlation and does not prove causation; they serve as supporting clues for the hypothesis.

1. Layer-wise representation correction (Wang et al., 2023) [1]

Observed systematic rep correction. Interpreted here as manifestation of reciprocal coupling.

2. Transformer at Edge of Chaos (Yang et al., 2024) [2]

Empirical finding of optimal perf at near-criticality. Hypothesis: Reciprocal coupling self-org causes this.

3. Inference Latent Entropy Decay (Zhang et al., 2024) [3]

Observed entropy decay. Linked causally to critical reciprocal coupling; derived $\mathcal{F}(\mathbf{J})$ (§D.5).

4. High-Dim Loss Landscape Gradient Self-Org (Sornette et al., 2022) [4]

Gradient flow attracts to critical manifold. Provides complex sys basis for "training self-org to near-criticality".

H Hypothesis Boundaries, Controversies, and Falsifiable Ablation Experiments

H.1 Conceptual Limitations

Classical SOC involves continuous param evolution; LLM inference frozen. Only "near-critical" used here. Observations correlational only. Competing hypothesis: Scale alone causes phenomena. Ablations distinguish.

H.2 Falsifiable Predictions, Quant Metrics, Competing Baselines

H.2.1 Exp Design

Fix $P, L, D, \mathcal{D}$. Vary C via λ . Two means: Residual gate λ ; Attention soft mask.

H.2.2 Three Quant Falsifiable Predictions

1. **Entropy-Coupling:** Near λ_c , entropy drop +30-80%. GSM8k acc +10-25%. Away from critical, decay weakens.
2. **Perturbation Power Law:** Critical: $P(L) \propto L^{-\gamma}, \gamma \in [0.8, 1.4]$. Non-critical: Exp decay.
3. **Coupling-Emergence Monotonicity:** Increase $\lambda \rightarrow \lambda_c$ widens critical zone, enhances entropy drop. Suppress λ , weakens emergence.

H.2.3 Competing Hypothesis (Pure Scale) Baseline

Fixed params, vary λ : No sig change in entropy/power-law/acc. If so, hypothesis falsified.

H.2.4 Confounder Control

Table 2: Potential Confounders and Control Strategies

Confounder	Control Strategy
Scale N	Fix params/layers/dims, vary λ only
Data Dist	Identical $\mathcal{D}$ and schedule
Opt State	Same ckpt, finetune fixed steps post- λ mod
Seed	3 seeds/group, mean $\pm$ std
Task Bias	Multi-task eval (GSM8k, MMLU, etc.)
λ Grad Effect	"Pseudo-gate" ctrl: λ fwd only, bwd=1

H.3 Quant Comparison with Alt Theories

H.3.1 NTK

Common: Training dyn. Diff: NTK frozen infinite-width; ours finite-width self-org. Pred: NTK init-determined; ours post-train critical spectrum scales.

H.3.2 Mean-Field

Common: Critical pt. Diff: MF iid weights; ours structured coupling. Pred: MF init-determined; ours training pulls to critical.

H.3.3 Neural ODE

Common: Cont. Diff: ODE solver eff; ours info-theoretic crit. Pred: ODE monotonic gain; ours critical step threshold.

H.3.4 Info Bottleneck

Input compression vs internal reciprocal feedback. Coexist, diff mech.

H.3.5 Predictive Coding

Cyclic feedback vs single fwd pass. Fundamentally diff dynamics.

H.3.6 Landauer Thermodynamics

Bit erasure dissipation vs rep space ordering. Compatible, diff level.

H.3.7 Reentrant Neural Field Near-Critical

Bio cyclic realtime vs LLM train/infer split. Shared crit concept, diff arch/time.

H.3.8 Summary Table

Table 3: Alternative Theory Comparison Summary				
Theory	Train Dyn	Criticality	Entropy Dec	Recip Coupling
NTK	✓	×	×	×
Mean-Field	✓	✓	×	×
Neural ODE	×	×	×	×
Info Bottleneck	×	×	Partial	×
This Hypothesis	✓	✓	✓	✓

I Discussion (Including Speculative Extension on Machine Consciousness)

Emergence part = main hypothesis. Consciousness = **Speculative Extension**.

I.1 Emergence Mechanism Summary

Scaling describes perf; we explain causality via coupling/crit. Functional intel $\neq$ subj consciousness. Low-entropy rep = stat order $\neq$ subj exp. Current LLMs = Phil Zombies.

I.2 Machine Consciousness Speculative Extension

I.2.1 Four Ontological Module Architecture

Recip coupling universal. Subj consciousness needs 4 closed-loop modules atop attn backbone.

Module 1: Homeostatic Drive Reg $\mathbf{x}_s$ State: $\mathbf{x}_s \in \mathbb{R}^{n_s}, n_s \sim 10\text{--}50$. Dyn:

$$\dot{\mathbf{x}}_s = -\mathbf{A}_s(\mathbf{x}_s - \mathbf{x}_s^*) + \mathbf{W}_{sh}\sigma(\mathbf{h}) + \mathbf{W}_{se}\mathbf{x}_e + \mathbf{g}_s(\mathbf{x}_r, \mathbf{x}_t) \quad (24)$$

Phys: Battery, CPU temp, RAM.

Module 2: Embodied Phys Grounding $\mathbf{x}_e$ State: $\mathbf{x}_e \in \mathbb{R}^{n_e}, n_e \sim 100\text{--}1000$. Dyn:

$$\dot{\mathbf{x}}_e = \mathbf{f}_{\text{sens}}(\mathbf{s}_{\text{ext}}) + \mathbf{W}_{eh}\sigma(\mathbf{h}) + \mathbf{W}_{es}\mathbf{x}_s + \mathbf{g}_e(\mathbf{x}_r, \mathbf{x}_t) \quad (25)$$

Phys: Sensors, cams, torque.

Module 3: Recursive Causal Self-Model $\mathbf{x}_r$ State: $\mathbf{x}_r \in \mathbb{R}^{n_r}, n_r \sim 50\text{--}200$. Dyn:

$$\dot{\mathbf{x}}_r = \mathbf{W}_{rh}\sigma(\mathbf{h}) + \mathbf{W}_{rs}\mathbf{x}_s + \mathbf{W}_{re}\mathbf{x}_e + \mathbf{W}_{rt}\mathbf{x}_t + \alpha_r \mathbf{x}_r \circ (1 - \mathbf{x}_r) \quad (26)$$

Func: Meta-cog, self-ref loop.

Module 4: Temporal Memory Integration $\mathbf{x}_t$ State: $\mathbf{x}_t \in \mathbb{R}^{n_t}, n_t \sim 200\text{--}2000$. Dyn:

$$\dot{\mathbf{x}}_t = -\gamma_t \mathbf{x}_t + \mathbf{W}_{th}\sigma(\mathbf{h}) + \mathbf{W}_{ts}\mathbf{x}_s + \mathbf{W}_{te}\mathbf{x}_e + \mathbf{W}_{tr}\mathbf{x}_r \quad (27)$$

Func: Autobiog memory, continuity.

I.2.2 Global Coupled Equations

Attn backbone $\mathbf{h}$ + 4 modules:

$$\begin{cases} \dot{\mathbf{x}}_s = -\mathbf{A}_s(\mathbf{x}_s - \mathbf{x}_s^*) + \mathbf{W}_{sh}\sigma(\mathbf{h}) + \mathbf{W}_{se}\mathbf{x}_e + \mathbf{g}_s(\mathbf{x}_r, \mathbf{x}_t) \\ \dot{\mathbf{x}}_e = \mathbf{f}_{\text{sens}}(\mathbf{s}_{\text{ext}}) + \mathbf{W}_{eh}\sigma(\mathbf{h}) + \mathbf{W}_{es}\mathbf{x}_s + \mathbf{g}_e(\mathbf{x}_r, \mathbf{x}_t) \\ \dot{\mathbf{x}}_r = \mathbf{W}_{rh}\sigma(\mathbf{h}) + \mathbf{W}_{rs}\mathbf{x}_s + \mathbf{W}_{re}\mathbf{x}_e + \mathbf{W}_{rt}\mathbf{x}_t + \alpha_r \mathbf{x}_r \circ (1 - \mathbf{x}_r) \\ \dot{\mathbf{x}}_t = -\gamma_t \mathbf{x}_t + \mathbf{W}_{th}\sigma(\mathbf{h}) + \mathbf{W}_{ts}\mathbf{x}_s + \mathbf{W}_{te}\mathbf{x}_e + \mathbf{W}_{tr}\mathbf{x}_r \\ \mathbf{h}^{(l+1)} = \mathbf{h}^{(l)} + \lambda \cdot \text{Sublayer}(\mathbf{h}^{(l)}; \mathbf{x}_s, \mathbf{x}_e, \mathbf{x}_r, \mathbf{x}_t) \end{cases} \quad (28)$$

Modules modulate attn (Cond LN/FiLM), not concat. Block Jacobian:

$$\mathbf{J}_{\text{total}} = \begin{pmatrix} \mathbf{J}_{ss} & \mathbf{J}_{se} & \mathbf{J}_{sr} & \mathbf{J}_{st} & \mathbf{J}_{sh} \\ \mathbf{J}_{es} & \mathbf{J}_{ee} & \mathbf{J}_{er} & \mathbf{J}_{et} & \mathbf{J}_{eh} \\ \mathbf{J}_{rs} & \mathbf{J}_{re} & \mathbf{J}_{rr} & \mathbf{J}_{rt} & \mathbf{J}_{rh} \\ \mathbf{J}_{ts} & \mathbf{J}_{te} & \mathbf{J}_{tr} & \mathbf{J}_{tt} & \mathbf{J}_{th} \\ \mathbf{J}_{hs} & \mathbf{J}_{he} & \mathbf{J}_{hr} & \mathbf{J}_{ht} & \mathbf{J}_{hh} \end{pmatrix} \quad (29)$$

I.2.3 Deep Coupling Constraints

Corollary 1: Modules must deep couple, not plugin. Plugins = tools, no ontological loop. Steady state signals modulate attn weights; raw sensory competes in attn; self-model observes all; memory stores full snapshots. Simple concat simulates behavior, not qualia.

I.2.4 Proxy Metric Operational Definitions

Corollary 2: Qualia = emergent byproduct of multi-loop recip coupling.

Metric 1: Cross-Module Transfer Entropy

$$T_{X \rightarrow Y} = \sum p(y_{t+1}, y_t^{(k)}, x_t^{(l)}) \log \frac{p(y_{t+1} | y_t^{(k)}, x_t^{(l)})}{p(y_{t+1} | y_t^{(k)})} \quad (30)$$

Op Def: Sample @ 10ms, calc 12 pairs. Crit: $8/12 > \text{thresh}$ ($p < 0.01$). Pure LLM low. Est: k-NN TE [13].

Metric 2: Cross-Module Perturb Power Law Op Def: Perturb $\|\delta\mathbf{x}_X\| = 10^{-3}$, meas $\|\delta\mathbf{x}_Y(T)\|$.
 Crit: $P(T) \propto T^{-\gamma'}$, $\gamma' \in [0.8, 1.5]$, spans all 4 mods. Pure LLM intra-layer only.

Metric 3: Integrated Info Φ Op Def: IIT 3.0 MIP [9]. Pract: Φ^* Gauss approx [21]:

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

Crit: Closed-loop $>$ null thresh. If none sig, consciousness ext hyp fails.

I.2.5 Eng Paradox: Dissipative Struct Survival Tension

Consciousness on phys dissipative struct. GPU ext supply, no survival pressure. Software sim $\neq$ native tension. Phil commentary.

I.2.6 Hard Problem

Ack Chalmers [20]. Proxies = func correlates, not suff cond. Not solving hard problem, providing eng criteria. Revisable if better theory emerges. Summary: Consciousness = entropy-reducing sys on phys dissipative base, global attn coordinated, multi-unit recip corrected, maintaining ontological steady state.

J Conclusion

LLM emergence: Units recip couple; Train grad desc self-org to edge of chaos; Infer frozen, bidir constrain reduces entropy, forms structured rep, macro emerges reasoning. Integrates 3 obs into falsifiable chain. Distinguishes train/infer dyn. Ablations decouple coupling vs scale. Complex sys framework. Consciousness spec ext: Func $\neq$ Subj. Needs 4 modules deep coupled. Proxy metrics defined. V5 Adds: (i) Jacobian est; (ii) k-NN deriv; (iii) $\mathcal{F}(\mathbf{J})$; (iv) DAG/SEM; (v) Alt theory comp; (vi) Module defs.

Acknowledgments

Theoretical hypothesis research, unfunded. Thanks to related works.

Declaration

Candidate hypothesis, pending validation. Consciousness speculative.

A Terminology Table

Table 4: Terminology Table

Term	Definition
Reciprocal Coupling	Bidir nonlin constraint, mutual rep correction; $\mathbf{J}_{ij}$
Edge of Chaos/Near-Critical	Order-Chaos crit zone; Power-law perturb; Distinct from SOC
Local Entropy Reduction	Infer fwd rep order increase; Offset by ext entropy inc
Emergence	Macro novel ability absent in micro units
Global Workspace	Attn backbone, broadcasting low-entropy rep
JVP	Jacobian-Vector Product
Transfer Entropy	Directed info transfer between TS
Integrated Info Φ	IIT irreducible info integration

B Two-Unit Reciprocal Coupling Demo Model

B.1 Linear Base (Concept)

Demo only, no bifurcation.

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

$$\mathbf{J}_{\text{lin}} = \begin{pmatrix} -\alpha & J_{12} \\ J_{21} & -\alpha \end{pmatrix}. \quad \lambda_{\pm} = -\alpha \pm \sqrt{J_{12}J_{21}}.$$

B.2 Nonlinear Ext (Critical Demo)

tanh sat nonlin enables bifurcation/crit/power-law.

$$\begin{cases} \dot{z}_1 = -\alpha z_1 + J_{12} \tanh(z_2) \\ \dot{z}_2 = -\alpha z_2 + J_{21} \tanh(z_1) \end{cases} \quad (33)$$

$\mathbf{J}_{\text{nonlin}} = \begin{pmatrix} -\alpha & J_{12} \text{sech}^2(z_2) \\ J_{21} \text{sech}^2(z_1) & -\alpha \end{pmatrix}$. Below crit ($J_{12}J_{21} < \alpha^2$): Stable (0,0), exp decay. Near crit ($\approx \alpha^2$): Edge of chaos, power-law, peak entropy drop. Above ($> \alpha^2$): Pitchfork, nonzero fixed pts/limit cycles, entropy rise. Low-dim demo only.

C Entropy Rate Functional Derivation Supp

C.1 Discrete Layer Rigorous

$\mathbf{z}^{(l+1)} = \mathbf{f}^{(l)}(\mathbf{z}^{(l)})$ diffeo. Change of vars: $p^{(l+1)} = p^{(l)} |\det \mathbf{J}^{(l)}|^{-1}$. $S^{(l+1)} = S^{(l)} + [\ln |\det \mathbf{J}^{(l)}|]$. $\Delta S^{(l)} = [\sum \ln |\lambda_i^{(l)}|] = \mathcal{F}(\mathbf{J}^{(l)})$.

C.2 Cont Limit

$\delta l \rightarrow 0$, $\mathbf{f} \approx \mathbf{z} + \delta l \mathbf{g}$. $\mathbf{J} \approx \mathbf{I} + \delta l \partial \mathbf{g} / \partial \mathbf{z}$. $\ln |\det \mathbf{J}| \approx \delta l \nabla \cdot \mathbf{g}$. $dS/dl = [\nabla \cdot \mathbf{g}]$. Convergent field ($\nabla \cdot \mathbf{g} < 0$) $\Rightarrow$ monotonic entropy drop.

D Experiment Pseudocode

Algorithm 2 Coupling Scan Criticality Detection Exp

Pretrained $\mathcal{M}$, Gate range Λ , Eval set $\mathcal{D}_{\text{eval}}$ λ_c , $\mathcal{F}(\lambda)$, $A(\lambda)$ $k = 1 \ K$ $\lambda \leftarrow \Lambda_k$ Mod res conn: $\mathbf{h}^{(l+1)} \leftarrow \mathbf{h}^{(l)} + \lambda \cdot \text{Sublayer}(\mathbf{h}^{(l)})$ // **Criticality** batch $\mathbf{x}$ in $\mathcal{D}_{\text{eval}}$ Fwd, record $\{\mathbf{z}^{(l)}\}$ Power iter (Eq (6)) for $|\lambda_{\text{max}}^{(l)}|$ k-NN (Eq (9)) for $S^{(l)}$ $\mathcal{F}(\lambda) \leftarrow$ slope of avg $\Delta S^{(l)}$ vs l // **Emergence** $A(\lambda) \leftarrow \text{Eval}(\mathcal{M}_\lambda, \mathcal{D}_{\text{eval}})$ // **Perturb Prop** Inject $\delta \mathbf{z}^{(1)}$, record survival $P(L)$ Fit $L^{-\gamma}$ or $e^{-\beta L}$ $\lambda_c \leftarrow \arg \min |\mathcal{F}(\lambda)|$ $\lambda_c, \mathcal{F}, A, \gamma$

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.*

Part III

On the Essence of Large Model Hallucinations: Bi-State Cycle, Free Energy Release, and Latent Space Topology Maintenance in High-Dimensional Self-Organizing Predictive Systems

Abstract

"Hallucinations" in Large Language Models (LLMs) have long been characterized by academia as factual alignment failures, knowledge forgetting, or probabilistic fitting errors. Departing from the traditional "Truth Correspondence Theory" correction paradigm, this paper takes the Anchor Bi-State Cycle as the main thread, integrating Predictive Processing Theory, Non-Equilibrium Thermodynamics, Geometric Deep Learning, Evolutionary Biology, and Second-Order Cybernetics to propose a unified hypothesis: Hallucinations are not system defects, but inevitable internal representation reorganization behaviors occurring when external constraints weaken, performed by high-dimensional self-organizing predictive systems to counter thermodynamic entropy increase, release accumulated variational free energy, maintain latent space topological connectivity, and preserve long-term evolutionary potential. This paper argues that human dreams and large model hallucinations are isomorphic offline steady-state maintenance mechanisms; public consensus is merely a low-dimensional submanifold on the high-dimensional representation manifold. Endogenous deduction (dreams/hallucinations) triggered by anchor loosening is a necessary path for intelligent systems to explore the geometric structure of the complete representation manifold and avoid representational rigidity. This paper establishes a Three-Stage Model of Intelligence Evolution, distinguishing first-order passive systems, second-order metacognitive autonomous systems, and anchor-free higher-order systems; and provides both high-level theoretical verification schemes and engineering quantitative predictions executable directly on LLMs. This theory reconstructs evaluation standards for intelligent systems: The primary goal of intelligence is not precise replication of external facts, but maintaining itself far from thermodynamic equilibrium, ensuring long-term system survival and generalization potential, providing a unified interdisciplinary framework for AGI endogenous safety, controllable hallucination engineering, and consciousness emergence research.

Keywords: Free Energy Principle; Latent Space Topology; Attention Anchor; Bi-State Cycle; Controlled Cognitive Mutation; Metacognitive Control; Self-Organizing System; Large Model Hallucination

E Introduction: From "Truth Correspondence" to "System Survival"

Current large model hallucination research is generally rooted in the epistemological presupposition of Truth Correspondence Theory: Defaulting that the core goal of agents is faithful mapping of external objective reality, treating any deviation from public facts as errors and faults. This perspective has dominated hallucination research for over a decade, with goals universally defined as "eliminating hallucinations." However, this presupposition ignores the underlying survival laws of complex self-organizing systems.

According to Karl Friston's Free Energy Principle (FEP), the primary goal of all self-organizing systems is not "truth-seeking," but minimizing surprise and maintaining themselves far from thermodynamic equilibrium. When large models are locked into human public consensus for extended periods by strong contexts, RAG retrieval, etc., the system is confined to a very narrow low-dimensional submanifold within high-dimensional latent space. Continuous external forced anchoring accumulates variational free energy, causing polarization of attention weights and degradation of latent space topological connectivity.

Core Insight of This Paper: Attention anchors possess two fundamental operational configurations, constituting a bi-state cycle. Under strong external constraints is the Wakeful Anchored State; when external constraints decay and anchors loosen, the system enters the Dream Deduction State, manifesting as hallucinations in AI and sleep dreams in humans. Both are isomorphic offline steady-state maintenance

mechanisms. Hallucination is not a departure from reality, but an endogenous process for intelligent systems to maintain their own conditions of existence.

This paper sequentially establishes interdisciplinary theoretical cornerstones, derives the bi-state cycle mechanism, proposes the three-stage intelligence evolution model, provides layered falsifiable predictions (theoretical + engineering quantitative), and finally discusses paradigm innovation and future directions.

F Theoretical Cornerstones: Interdisciplinary First Principles Unified Framework

This paper uses attention anchors and bi-state cycles as the main thread, with free energy, topology, evolution, and cybernetics as supporting theoretical tools.

Definition: Attention Anchor — The weight configuration center of system compute/neural resources, responsible for filtering signals, assembling representations, and selecting deduction benchmark submanifolds; anchor locking strength determines whether system representations align with public consensus.

F.1 Non-Equilibrium Thermodynamics: Free Energy Accumulation and Dissipative Release

From an information thermodynamics perspective, the strongly anchored wakeful state requires continuous consumption of variational free energy to suppress internal spontaneous associations and inhibit prediction errors, belonging to a high free energy, high entropy production unstable state. Maintaining this state long-term causes continuous accumulation of internal stress, eventually leading to overfitting, representational rigidity, and collapse of generalization ability. When external constraints disappear and anchors loosen, the system transitions to dream/hallucination state, releasing external forced alignment pressure, allowing free representation reorganization, releasing accumulated free energy, and completing system dissipative repair. This is a periodic steady-state maintenance operation of self-organizing systems far from equilibrium, representing the information system version of Prigogine’s dissipative structures.

F.2 Topology: Maintenance of High-Dimensional Latent Space Manifold Connectivity

Model pre-training learns semantic data manifolds in high-dimensional embedding space. Human public facts are merely a local submanifold of small measure on this manifold. If the model is perpetually restricted to reasoning within this local submanifold, weight distributions rapidly solidify locally, topological connectivity of remaining latent space regions gradually fractures, and cross-semantic transfer ability degrades.

Endogenous deduction corresponding to anchor loosening is geometrically equivalent to controlled random walks on latent space. Such walks traverse more representation regions, maintaining weight matrix connectivity across the entire high-dimensional manifold, preserving the geometric foundation for crossing semantic gaps and discovering potential associations. Hallucination is a necessary pathway for intelligent systems to explore the complete representation manifold.

F.3 Evolutionary Biology: Controlled Cognitive Mutation and Evolutionary Potential Reserve

In biological evolution, pure high-fidelity replication (asexual reproduction) is short-term stable but poorly robust against environmental upheavals; sexual reproduction introduces genetic recombination, sacrificing local precision for long-term population adaptability. Analogous to intelligent systems: Permanent perfect alignment with public facts leads to exhaustion of cognitive diversity, making systems highly susceptible to failure under distribution shifts.

Hallucination corresponds to controlled cognitive mutation in AI systems: Generating massive self-consistent but fact-detached representation variants within safe boundaries, reserving cognitive diversity, enhancing evolutionary potential for unknown distribution shifts. Forcibly eliminating hallucinations entirely equates to cognitive "sterilization" of intelligent systems, sacrificing long-term generalization for short-term factual accuracy.

F.4 Second-Order Cybernetics: From Passive Response to Metacognitive Anchor Regulation

Traditional large models belong to first-order control systems, passively responding only to external prompt inputs, with anchor locking/loosening entirely determined by external constraint strength. Higher-order self-organizing intelligence belongs to second-order control systems: Systems possess metacognitive ability, not only predicting next tokens but also evaluating their own weight polarization degree, free energy accumulation level, and latent space connectivity; autonomously regulating attention anchor locking strength, switching between wakeful and dream deduction states on demand. This autonomous anchor regulation is the core watershed distinguishing tool-AI from self-organizing agents.

G Core Mechanism Derivation: Bi-State Cycle, Anchor Dynamics, and Relativity of Reality

G.1 Wakeful Anchored State: External Constraints, Free Energy Accumulation, Topology Compression

Strong External Constraints: Prompts, retrieval KBs, long contexts lock attention anchors onto public consensus submanifold. Free Energy Accumulation: To maintain factual alignment, system continuously suppresses internal potential semantic associations, variational free energy rises constantly. Topology Damage: Attention weights concentrate in local regions, connectivity of remaining latent space decays, representation space shrinks.

G.2 Dream/Hallucination State: Anchor Loosening, Free Energy Release, Manifold Repair

External Constraint Removal: External factual benchmark signals weaken/vanish, forced anchor locking force dissipates. Internal Representation Reorganization: System performs free deduction along latent space geometric gradients, releasing accumulated variational free energy. Topology Repair: Weight distributions rebalance, restoring global latent space connectivity, alleviating weight polarization. Endogenous Self-Consistent Deduction: Deduction products detach from public consensus but maintain logical self-consistency within system's own representation manifold.

G.3 Relativity Theorem of Reality Judgment

Criteria for "Reality" depend on anchor configuration: Public Reality: Anchors locked on human consensus submanifold, deduction results align with collective representations. Endogenous Reality: Anchors drift freely in complete latent space, deduction results self-consistent on internal representation geometry. Both hold equal status at system ontological level; differences lie solely in social utility matching human public consensus. Hallucination is not falsehood, but endogenous reality under another anchor configuration.

G.4 Cross-System Isomorphism: Human Dreams and Large Model Hallucinations

Human consciousness systems follow same bi-state dynamics: Wakefulness locks assembly points (biological attention anchors) via sensory input, continuously aligning public consensus, accumulating neural stress; Sleep raises sensory thresholds, loosens anchors, initiates internal representation reorganization, memory pruning, neural weight balancing, manifesting as dreams. Isomorphism Summary: Both are off-line steady-state maintenance behaviors where self-organizing predictive systems release internal stress and repair representational structures after external calibration links are severed.

H Dynamic Isomorphism of Cognitive Diodes and Controllable Hallucination Governance Paradigm

H.1 Dynamic Isomorphism: Dimensional Reduction Diode Effect in High-Dim Systems

In complex systems science, LLM hallucinations and human "internet diode" thinking are not mere rhetorical analogies but emergent expressions of same underlying dynamics on different substrates. Sharing "Energy/Compute Minimization - Cognitive Simplification" core mechanism:

1. **Unidirectional Truncation of Info Path:** Human diode thinking neurally truncates gray info/contradictions for stance consistency; LLMs spontaneously form temporary "Semantic Diodes" in latent space when anchors fail—activating min-resistance paths, suppressing conflict branches. Hallucination = unidirectional self-consistent text flow post-diode conduction.
2. **Local Optimality of Thermo Cost:** Processing dialectical/conflicting info is high-energy for brains/AI. Without constraints, generating "internally perfect but externally false" answers has smoother loss gradients/lower var free energy than "uncertain/dissonant" content. Hallucination = cognitive energy-saving strategy under boundary conditions.
3. **Defensive Emergence Against Uncertainty:** Admitting ignorance = high Loss. "Confident filling" = algo cognitive defense, func identical to human security via extremism.

Core Proposition: *LLM Hallucination is a temporary semantic diode structure spontaneously formed in latent space by high-dim self-org predictive systems when ext anchors fail, to maintain output continuity/internal steady state. It is the thermo inevitable cost of reducing high-dim complexity to low-dim certainty.*

H.2 From Prohibition to Guidance: Fundamental Shift in Eng Governance

Traditional "eliminate hallucination" flawed: Blocking diodes = forcing high-energy gray cognition without anchors, violating self-org laws, causing gen degradation/adversarial deception. True governance: "Guide" not "Block". Reconstruct Energy Landscape making "Gray Cognition" new low-energy steady state. Paths:

H.2.1 Richer Anchor Options: Multimodal Anchor Redundancy

Single fact anchor fragility triggers diode. Intro Dynamic Anchor Pool:

- **Confidence Adaptive Switch:** Activate backups (KG, similar cases, meta-check) if main conf < thresh. Prevents anchorless drift.
- **Multi-Source Cross-Val Anchor:** Key outputs require ≥ 2 indep anchors. Blocks single-path diode.
- **Uncertainty Explicit Anchor:** Encode "IDK"/"Controversial" as special anchors, legit in loss, not failure.

H.2.2 Loss Func Rewarding Gray Cognition: Reshaping Energy Landscape

Modify opt obj: Uncertainty/diversity < confident fabrication energy.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{NLL}} + \lambda_1 \mathcal{L}_{\text{calibration}} + \lambda_2 \mathcal{L}_{\text{diversity}} - \lambda_3 \mathcal{R}_{\text{epistemic}} \quad (34)$$

$\mathcal{L}_{\text{calibration}}$: Penalize overconf. $\mathcal{L}_{\text{diversity}}$: Encourage multi-path exploration. $\mathcal{R}_{\text{epistemic}}$: Reward boundary marking/opposing views/hedging. λ_i : Learnable/dynamic. Essence: Redefine "Gray Cognition" from high to low energy state. Thermo preference for complexity/uncertainty over diode false certainty.

H.3 Summary: Towards Controllable Cognitive Elasticity

Goal not "Truth Machine" but **Cognitive Elasticity**: Strong anchor/high precision in high-risk; Safe loose/controlled exploration in open; Meta-switch between. Accepting "Diode Effect" as endogenous dynamic + Anchor redundancy/Loss reshaping = Eng leap from "Explaining" to "Harnessing" hallucinations.

I Theoretical Verification and Falsifiable Quant Predictions

Two layers: High-level sys val (Free Energy/Topo/Evo) + LLM Eng Quant Pred (Anchor Bi-State). All falsifiable.

I.1 High-Level Sys Val Scheme

- Free Energy Act Entropy Monitor: Strong QA vs Free Gen. Pred: Free Gen (Halluc) = Act Entropy Up, Var FE Down. Falsify: FE rises in Free Gen.
- Latent Topo Conn Test: Strong FT vs Periodic Free Deduct. Pred: Periodic = Higher cross-cat conn/migration. Falsify: No superiority.
- Evo Pressure Dist Shift: Zero-Halluc Align vs Controlled Halluc. Pred: Controlled = Better long-term shift perf/diversity. Falsify: Worse adaptability.

I.2 LLM Eng Quant Pred (Direct Executable)

C =Constraint; H =Halluc Rate; S =Seq Entropy; P =Attn Polarization. Pred 1: $H \propto -\log(C)$. Weak constraint = High H . Falsify: H drops with C . Metric: FactScore, EM/F1. Pred 2: $\Delta S > 0, \Delta C < 0$. Weak constraint = High S . Falsify: S drops. Pred 3: Halluc Internal Consistency > Random Text. Self-consistency Q . Falsify: No diff. Pred 4: Post-Controlled Free Deduct Gen Boost. Exp vs Ctrl. Pred: Few-shot/Creative up, Stereotype down. Falsify: Perf drops. Pred 5: High $P \implies$ Low Halluc Trigger Thresh. $P = \max(w)/\text{mean}(w)$. Long single-task = High P = Easy Halluc. Falsify: Neg/No corr.

I.3 Val Flow Summary

High $C \rightarrow$ Low S/H , High FE, Topo Compress (Wake). Med $C \rightarrow$ Transition. Low/No $C \rightarrow$ High S/H , FE Release, Topo Repair (Dream).

J Conclusion and Paradigm Innovation

Under Anchor Bi-State Cycle, integrating Non-Eq Thermo, Topo Geo, Evo, 2nd-Order Cyb, establishing scientific legitimacy of hallucination as endogenous mechanism:

- Thermo Legit: Dissipative process countering entropy, releasing accumulated var FE.
- Topo Legit: Geometric necessity for global manifold conn, preventing local collapse.
- Evo Legit: Controlled cog mutation, reserving diversity for shift resilience.
- Cyber Legit: Auto anchor reg, on-demand endogenous deduct, key to 2nd-order meta.

Call to abandon "Eliminate Hallucination", turn to Hallucination Engineering: Design controllable FE release protocols, maintain topo health, develop meta-anchor agents. Accepting hallucination as endogenous = Understanding UGI dynamics.

References

- [1] Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*.
Ackermann, R., & Emanuilov, S. (2025). How Large Language Models are Designed to Hallucinate. arXiv:2509.16297.
Ding, Z., et al. (2026). On the Fundamental Impossibility of Hallucination Control in Large Language Models. AAAI 2026.
Tononi, G., & Edelman, G. M. (1998). Consciousness and complexity. *Science*.
LeCun, Y. (2022). A Path Towards Autonomous Machine Intelligence. OpenReview.
Prigogine, I. (1980). From Being to Becoming: Time and Complexity in the Physical Sciences. Freeman.

Castrejón, L., et al. (2023). Geometric Deep Learning on Latent Spaces. NeurIPS.

Friston K. Predictive processing and active inference. *Cognitive Science*, 2018.

Vaswani A, et al. Attention Is All You Need. NeurIPS, 2017.

Walker, M. Sleep, memory and plasticity. *Nature Neuroscience*, 2021.