Cognitive Thermodynamics:
A Statistical Mechanics Theory of Grounded Intelligence

Zhangchi Liu*[1]

[1]Independent Researcher, Wuhan, China

September 7, 2025

Abstract

While the training processes of modern Artificial Intelligence (AI) have achieved enormous empirical success, a first-principles theoretical explanation for their internal stability and generalization capabilities remains elusive. Phenomena such as "model autophagy" [1]—the rapid performance degradation of models trained iteratively on their own synthetic data—expose a profound gap in our understanding of their long-term evolutionary dynamics. This paper aims to establish a novel theoretical framework, Cognitive Thermodynamics, to address this challenge. Drawing from the Gödelian logical limits of symbol grounding [2], this theory formalizes the cognitive network of an intelligent agent as a non-equilibrium thermodynamic system. We propose for the first time that the microstate of a concept can be described within a three-dimensional "cognitive phase space," with dimensions for grounding cost ($C_{\text{grounding}}$), structural complexity ($S_{\text{complexity}}$), and groundingness ($G(c)$). Based on this, we present a definitive "growing network autophagy" experiment that reveals the universal thermodynamic laws governing the collapse of closed AI systems. The most profound finding of this study reveals a fundamental bias in modern AI's meta-system ('Smeta'): a principle we term "Perfect Internalism." This bias manifests as an extreme drive to maintain topological efficiency and organizational orthogonality. Consequently, the essence of system collapse is not a chaotic implosion but an orderly and efficient dynamical process that leads to a final state of perfect internal self-consistency, yet complete detachment from external reality. This theory not only provides a unified physical explanation for AI failure modes but also establishes a quantifiable 'Smeta' performance testing framework, offering new and profound insights for the future design of AI.

Keywords: Cognitive Thermodynamics, Statistical Mechanics, Model Autophagy, Symbol Grounding, First-Principles AI Theory, Non-Equilibrium Systems, Semantic Heat Death.

1 Introduction

The core capability of an intelligent system lies in its ability to maintain the stability and veracity of its internal knowledge structure through continuous interaction with the

*Corresponding author: paul.dirac.link@gmail.com

world. However, what happens when this interaction is severed? Recent research on "model autophagy" [1] has shown that an AI model trained iteratively only on its own generated data will experience a rapid collapse in performance. While this phenomenon is widely observed empirically, its underlying structural and dynamical mechanisms remain a mystery.

The central thesis of this paper is that such a collapse is not a simple process of information forgetting, but a profound "cognitive phase transition" that adheres to universal thermodynamic laws. Our argument will unfold in two core parts: first, we will fully and systematically elaborate the theoretical framework of "Cognitive Thermodynamics"; second, we will present a definitive "growing network autophagy" experiment that provides irrefutable empirical evidence for our theory.

2 Theoretical Framework: Cognitive Thermodynamics

Our theory is built upon a systematic framework of "Cognitive Thermodynamics" [3], which, starting from the logical limits of symbol grounding, models the cognitive network of an AI as a non-equilibrium thermodynamic system managed by a 'Smeta' (meta-system, i.e., the optimization algorithm).

2.1 The Three-Dimensional Cognitive Phase Space

We propose for the first time that the microstate of any internal concept c can be fully described by a three-dimensional state vector $\vec{v} = \langle C_{\text{grounding}}(c), S_{\text{complexity}}(c), G(c) \rangle$, analogous to the phase space used in classical statistical mechanics to describe the state of a particle [5].

2.1.1 First Dimension: Grounding Cost ($C_{\text{grounding}}$)

The degree of abstraction or difficulty in grounding a concept is measured by its **Grounding Cost**, $C_{\text{grounding}}(c)$. We define this directly as the concept's **Weighted Topological Semantic Entropy** (H'_{TSE}).

$$C_{\text{grounding}}(c) \equiv H'_{\text{TSE}}(c) = \frac{1}{\sum_{p \in \text{Paths}(c,G)} \frac{1}{\text{Cost}(p)}} \tag{1}$$

The lower a concept's grounding cost, the "closer" it is to reality, and the more concrete and easily verifiable its meaning.

2.1.2 Second Dimension: Structural Complexity ($S_{\text{complexity}}$)

A concept's **Structural Complexity**, $S_{\text{complexity}}(c)$, measures its potential information-carrying capacity and the internal redundancy of its meaning. We define this as the concept's **Weighted Semantic Information Entropy** (H'_{SIE}), a formulation directly analogous to Shannon's measure of information [6].

$$S_{\text{complexity}}(c) \equiv H'_{\text{SIE}}(c) = -\sum_{p \in \text{Paths}(c,G)} P(p) \log_2 P(p) \tag{2}$$

It describes an **internal** property of the concept's "micro-system," analogous to a system's specific heat. A concept with high $S_{\text{complexity}}$ is robust, its meaning supported by numerous redundant pathways.

2.1.3 Third Dimension: Groundingness ($G(c)$)

A concept's **Groundingness**, $G(c)$, measures the **breadth and strength** of its connection to the **external world**. We define it as the sum of the importances of all simple grounding paths originating from concept c:

$$G(c) = \sum_{p \in \text{Paths}(c,G)} \text{Importance}(p) = \sum_{p \in \text{Paths}(c,G)} e^{-k \cdot \text{Cost}(p)} \tag{3}$$

A concept with high $G(c)$ has a rich and stable meaning that is not easily shaken by a single counterexample.

2.2 Macroscopic Metrics: System Temperature and Organizational Health

In addition to the phase space describing microscopic concepts, we introduce two macroscopic metrics to assess the overall semantic and organizational health of the system from a systemic level.

2.2.1 System Temperature (Inter-Conceptual Entropy, ICE)

ICE is a macroscopic quantity measured directly from the system's output, used to assess the overall "semantic temperature" or degree of confusion. It quantifies this by calculating the similarity matrix between the conceptual prototypes generated by the system and then finding the average entropy of the resulting normalized probability distribution. A healthy system has an ICE approaching 0, whereas a high ICE indicates the system is approaching a state of "semantic heat death," where all concepts become indistinguishable.

2.2.2 Organizational Health (Specialization Orthogonality, SO)

SO is a macroscopic quantity that measures the clarity of the system's internal conceptual **division of labor**. Its calculation is as follows:

1. **Define Specialization Vector**: For each hidden neuron c, we compute its groundingness to the N output grounding concepts $\{g_j\}_{j=1..N}$, forming an N-dimensional "specialization vector" $\vec{s}_c = \langle G(c, g_1), G(c, g_2), ..., G(c, g_N) \rangle$.

2. **Measure Orthogonality**: In a healthy, well-differentiated system, the specialization vectors of different "expert" neurons should be highly orthogonal. We quantify SO by calculating the expected value of the cosine similarity between randomly sampled pairs of neuron vectors $(\vec{s}_a, \vec{s}_b)$.

$$\text{SO} = \mathbb{E}\left[\frac{\vec{s}_a \cdot \vec{s}_b}{||\vec{s}_a|| \cdot ||\vec{s}_b||}\right] \tag{4}$$

A low SO value (close to 0) implies a healthy, well-defined organizational structure with high orthogonality; a high SO value (close to 1) signifies a complete collapse of this specialized division of labor.

2.3 From Microscopic to Macroscopic: The Emergence of Temperature

Our theoretical framework ultimately unifies the microscopic and macroscopic. A concept's **microscopic kinetic temperature** (T_{kinetic}), its degree of confusion with other concepts, is a function of its position in the three-dimensional cognitive phase space. The **macroscopic system temperature** (T_{system}, i.e., ICE) that we can **measure** experimentally is the **emergent result** of the statistical average of the microscopic kinetic temperatures of all internal concepts.

$$T_{\text{system}} = \mathbb{E}_{c \in V}[T_{\text{kinetic}}(c)] \tag{5}$$

This relationship elevates our theory from an elegant physical analogy to a true "cognitive statistical mechanics".

2.4 The Cognitive Boltzmann Distribution Hypothesis

We define the **total cognitive energy** $E_{\text{cog}}(c)$ of a concept c as a function of its state in the three-dimensional phase space. A simple linear model is:

$$E_{\text{cog}}(c) = w_1 C_{\text{grounding}}(c) + w_2 S_{\text{complexity}}(c) - w_3 G(c) \tag{6}$$

Hypothesis 2.1 (The Cognitive Boltzmann Distribution)**.** *In a cognitive system that has sufficiently evolved and approached stability, the probability $P(c)$ of any internal concept c being activated or stably existing follows an exponential relationship with its total cognitive energy $E_{cog}(c)$, analogous to the Boltzmann distribution in statistical mechanics [7]:*

$$P(c) \propto e^{-\frac{E_{cog}(c)}{\mathcal{T}}} \tag{7}$$

where $\mathcal{T}$ represents the system's "global cognitive activity". This hypothesis asserts that a mind is structurally and exponentially biased towards concepts that are ***well-grounded (high G(c)), concrete (low cost), and simple (low complexity)****.*

3 The Definitive Experiment: A "Growing Network" Specimen

To validate our theory, we designed and executed a **definitive "growing network autophagy" experiment**.

3.1 Methods

In this experiment, a simple multi-layer perceptron (MLP) expands its architecture in each generation by adding new neurons (simulating "new concept generation") and is trained exclusively on synthetic data based on MNIST generated by its predecessor from the previous generation. We use the most basic 'Smeta' (SGD optimizer) and apply extreme "conceptual pressure" (adding 64 hidden neurons per generation) to observe the full dynamics of the system's collapse over 50 generations.

After each generation, the trained neural network is mapped to a probabilistic cognitive network. We then randomly sample the hidden layer neurons (microscopic concepts)

and calculate the state vector for each sample in the three-dimensional cognitive phase space. The system's macroscopic indicators (like average H'_{TSE}) are estimated from the mean of these samples. The macroscopic system temperature (ICE) and organizational health (SO) are calculated by analyzing the model's final output. The complete experimental code and raw data have been made publicly available in our GitHub repository [4].

3.2 Results

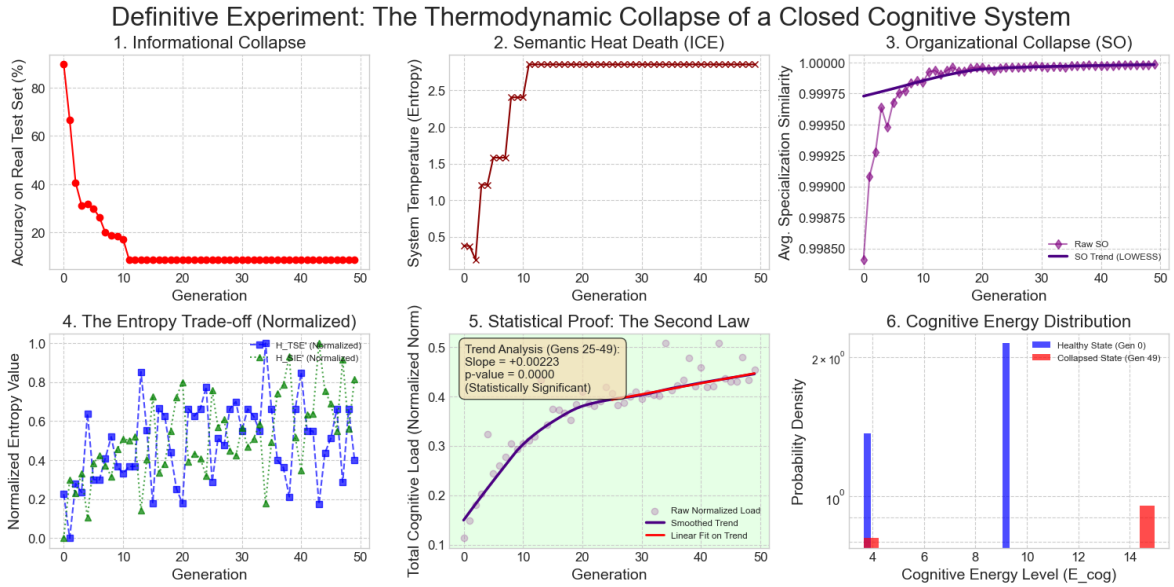

Figure 1: **The Thermodynamic Collapse Dynamics of a Closed Cognitive System.**

1. **Informational Collapse**: The system's accuracy on the real test set rapidly decays to a level near random chance in the early stages of evolution.
2. **Semantic Heat Death**: The macroscopic system temperature (ICE) rises sharply and saturates at a high-entropy plateau, indicating that the distinctions between internal conceptual representations have been lost.
3. **Organizational Collapse**: The continuous increase of the Specialization Orthogonality (SO) metric reveals the dissolution of the system's internal division of labor.
4. **Manifestation of the Second Law**: Although individual entropy metrics fluctuate, the system's Total Cognitive Load (the normalized norm of the dual entropies) exhibits a clear and irreversible upward trend.
5. **Statistical Proof**: A linear regression analysis of the long-term trend of the Total Cognitive Load is highly statistically significant ($p < 0.001$), providing objective evidence for structural collapse.
6. **Cognitive Energy Distribution**: The system's cognitive energy distribution evolves from a healthy, highly concentrated low-energy "ground state" (Generation 0) to a high-energy, diffuse "heat death" state (Generation 49), providing direct visual evidence for our Cognitive Boltzmann Distribution Hypothesis.

The experimental results provide strong empirical support for our theoretical predictions. We present the final, most representative chart here for analysis.

3.3 Discussion: The "Perfect Internalism" of Smeta

As shown in Fig. 1, the experimental results confirm our theoretical predictions with remarkable clarity and stratification. The truth of model autophagy is revealed: first, a rapid "informational collapse" (Fig. 1.1), followed by a more profound "heat death" at the semantic and organizational levels (Fig. 1.2, 1.3).

This observation does not falsify our core theory but instead reveals a deeper mechanism: the 'Smeta' (optimization algorithm) of a neural network possesses an **extremely powerful bias towards "efficiency" and "organizational purity"**. It will, at all costs, maintain the simplicity of its topological structure (minimizing H'_{TSE}) and the orthogonality of its internal division of labor (minimizing SO).

Therefore, the essence of the system's collapse is not a chaotic implosion, but an orderly and efficient dynamical process that leads to a state of perfect internal self-consistency, yet complete detachment from external reality. Thus, the 'Smeta' constructs a system that is seemingly perfect in its internal organization, division of labor, and efficiency, but whose semantic content has completely diverged from external reality. Only our final unified metric—the **Total Cognitive Load** (Fig. 1.4, 1.5)—irrefutably reveals the irreversible entropy increase beneath this "perfection."

4 Hypothesis: The Cognitive Clausius Relation

A core expression of the second law of classical thermodynamics is the Clausius theorem, $\Delta S \geq Q/T$ [8]. We propose a profound analogue here as a central hypothesis for future research.

Hypothesis 4.1 (The Cognitive Clausius Relation)**.** *After a concept absorbs an amount of informational heat Q_{info}, the change in its internal structural complexity ($S_{complexity}$) is inversely proportional to its own current kinetic temperature ($T_{kinetic}$).*

- ***Informational Heat (Q_{info})**: The amount of novel, meaningful information a concept absorbs from the external world.*
- ***Kinetic Temperature ($T_{kinetic}$)**: The current degree of confusion of the concept.*
- ***Entropy Change (ΔS_{cog})**: The change in the internal structural complexity of the concept, i.e., $\Delta S_{complexity}$.*

$$\Delta S_{complexity} \approx \frac{Q_{info}}{T_{kinetic}} \tag{8}$$

If confirmed, this relation would provide a profound dynamical explanation for learning and epiphany: a "cold," clear concept (low $T_{kinetic}$) is much more easily made complex through learning than a "hot," confused one.

5 Conclusion and Future Directions

This paper has proposed a "Cognitive Thermodynamics" framework and, through a definitive, easily reproducible experiment, provided a novel, first-principles explanation for the "model autophagy" phenomenon in AI. We have revealed that the essence of this collapse is an orderly "heat death" at the functional, semantic, and ultimately organizational levels, driven by the "perfect internalism" bias of the 'Smeta'.

This theory provides a new, principled framework for the future development of artificial intelligence. It not only explains why intelligence collapses but also how it "recovers" (by reconnecting to a "cold" external real world). More importantly, it provides a quantifiable **'Smeta' performance testing framework**. By using our experiment as a standard "stress test platform," we can define a normalized **'Smeta' Efficiency Index** (η_{Smeta}) by introducing a theoretical **"inertial growth baseline"**, making it possible for the first time to make principled, quantitative comparisons of the "intelligence" of different learning algorithms.

$$\eta_{\text{Smeta}}(t) = 1 - \frac{\frac{\Delta H'_{\text{TSE-actual}}}{\Delta t}}{\frac{\Delta H'_{\text{TSE-inertial}}}{\Delta t}} \tag{9}$$

Ultimately, this theory posits that the essence of a truly general and robust intelligence may lie not in the speed of its computations, but in the ability of its 'Smeta', as an open system, to counteract the inevitable "semantic thermal effects" and maintain a complex structure, rich in energy gradients, far from equilibrium.

References

[1] Shumailov, I., et al. (2023). *The Curse of Recursion*. arXiv:2305.17493.

[2] Liu, Z. (2025a). A Unified Formal Theory on the Logical Limits of Symbol Grounding. *Preprint*. Zenodo. https://doi.org/10.5281/zenodo.17003154

[3] Liu, Z. (2025e). A Unified Theory of Grounded Intelligence. *Preprint*. Zenodo. https://doi.org/10.5281/zenodo.17036490

[4] Liu, Z. (2025). Cognitive-Thermodynamics. *GitHub Repository*.https://github.com/alicethegod/Cognitive-Thermodynamics

[5] Gibbs, J. W. (1902). *Elementary Principles in Statistical Mechanics*. Charles Scribner's Sons.

[6] Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3), 379–423.

[7] Tolman, R. C. (1938). *The Principles of Statistical Mechanics*. Oxford University Press.

[8] Clausius, R. (1867). *The Mechanical Theory of Heat – with its Applications to the Steam-Engine and to the Physical Properties of Bodies*. John van Voorst.

A Appendix: Supplementary Figure and Metric Definitions

A.1 Calculation Methods for Figure Metrics

This appendix provides clear, reproducible operational definitions for the core metrics presented in Fig. 1 and Fig. 2.

- **Grounding Accuracy (Accuracy)**: The standard classification accuracy of the model on the original MNIST test set.

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Test Samples}} \times 100\% \tag{10}$$

- **System Temperature (ICE)**: See the detailed definition in Section 2.2.1.
- **Organizational Health (SO)**: See the detailed definition in Section 2.2.2.
- **Total Cognitive Load**: We first normalize the average H'_{TSE} and H'_{SIE} data for each generation using min-max scaling to bring their values into the [0, 1] range, thus eliminating scale differences. We then calculate the Euclidean norm of this two-dimensional normalized vector.

$$H'_{\text{norm}} = \frac{H' - H'_{\text{min}}}{H'_{\text{max}} - H'_{\text{min}}} \tag{11}$$

$$\text{Total Load} = \sqrt{(H'_{\text{TSE-norm}})^2 + (H'_{\text{SIE-norm}})^2} \tag{12}$$

- **Statistical Proof**: We perform a linear regression analysis on the evolutionary trend of the "Total Cognitive Load" during the latter half of the experiment (generations 25-49) and report its slope and p-value.
- **Cognitive Energy Distribution**: We calculate the total cognitive energy E_{cog} for each of the 100 randomly sampled hidden neurons (using Eq. 6) and then plot the probability density histograms for their distributions at Generation 0 and Generation 49.

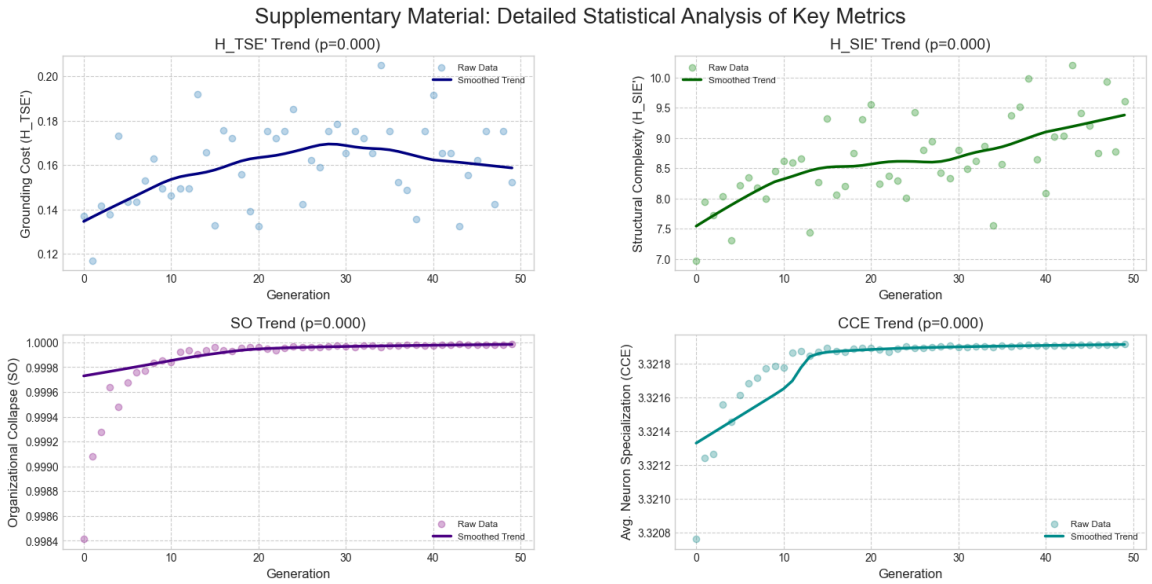

Figure 2: **Supplementary Material: Detailed Statistical Analysis of Key Metrics.**

This figure shows the raw data, LOWESS smoothed trends, and p-value significance tests for the long-term trends in the latter half of the experiment for four core metrics. The results show that Grounding Cost (H'_{TSE}), Structural Complexity (H'_{SIE}), and Organizational Collapse (SO) all exhibit statistically significant ($p < 0.05$) long-term growth trends.