The Thermodynamics of Spectral Collapse in Artificial Intelligence

By Alex Da Yin

January 26, 2026

In the prevailing discourse on the safety of artificial intelligence, our attention is almost exclusively captured by the behavioral surface. We analyze adversarial suffixes that induce jailbreaks, detect backdoors injected via clean-label poisoning [1, 2], or study the persistence of deceptive alignment strategies [3]. These phenomena, while critical, represent the *symptoms* of misalignment rather than its *physics*. Recent work on "Model Collapse" has begun to touch upon the structural degradation of generative systems when fed recursively generated data [4, 5], yet this remains a statistical phenomenon of distribution drift. A far more fundamental vulnerability lies unexamined—a structural pathology governed not by data statistics, but by the interplay between adaptive optimization and the geometry of high-dimensional manifolds. As we push the boundaries of model scale, we may be inadvertently cultivating a susceptibility to **Pathological Catalytic Attacks (PCA)**, a mechanism where microscopic perturbations induce macroscopic spectral collapse.

To rigorously define this threat, we must ground our understanding of intelligence in the **Low-Rank Hypothesis**. As elucidated by Thibeault et al. [6] in *Nature Physics*, complex systems—from biological neural networks to social graphs—exhibit a characteristic rapid decay in their singular value spectra. Intelligence, in this view, is a compression phenomenon. A pre-trained weight matrix $W \in \mathbb{R}^{d_{out} \times d_{in}}$ does not utilize its full rank; rather, it collapses into a low-dimensional signal subspace $\mathcal{S}$ spanned by the top-r singular vectors, separating causal structure from the vast null subspace $\mathcal{N}$ of entropy.

It is precisely this sophisticated compression that creates the attack surface. Our theoretical analysis suggests that the metabolic mechanisms used to train these systems—specifically, geometry-adaptive optimizers like AdamW—contain an intrinsic instability when interacting with the null space $\mathcal{N}$. Consider the standard adaptive update rule where the step size is normalized by the second moment estimate v_t. In the signal subspace $\mathcal{S}$, gradients are persistent, yielding $v_{t,\mathcal{S}} \gg \epsilon$. However, in the orthogonal null subspace $\mathcal{N}$, historical gradients are negligible, leading to $v_{t,\mathcal{N}} \to 0$.

This creates a condition of **Metabolic Amplification**. If an adversary injects a "catalytic" input x_c designed to be orthogonal to the signal manifold (i.e., $x_c \in \mathcal{N}$), the optimizer perceives the resulting gradient g_c not as noise, but as a high-priority signal requiring rapid adaptation. The effective learning rate η_{eff} scales inversely with the local curvature estimate:

$$\eta_{eff}(\mathcal{N}) \approx \frac{\eta}{\epsilon} \gg \frac{\eta}{\sqrt{v_{t,\mathcal{S}}}} \approx \eta_{eff}(\mathcal{S})$$

Defining the amplification factor $\kappa = \sqrt{v_{t,\mathcal{S}}}/\epsilon$, we find that $\|\Delta W_{\mathcal{N}}\| \approx \kappa \|\Delta W_{\mathcal{S}}\|$. Since ϵ is typically infinitesimal (10^{-8}), κ can be macroscopic ($10^3 \sim 10^5$). Thus, a microscopic input does not merely trick the model; it catalyzes a massive injection of energy into the noise dimensions of the weight matrix.

In a naive linear system, this noise might simply be additive. However, modern architectures employ normalization layers (e.g., RMSNorm) which enforce a strict energy constraint. This leads to **Signal Erosion**. Let $z = z_s + z_n$ represent the pre-activation state, composed of the valid signal $z_s \in \mathcal{S}$ and the amplified parasitic noise $z_n \in \mathcal{N}$. The normalization operator $\psi(z) = \gamma \odot \frac{z}{\|z\|_2}$ enforces a zero-sum game on the output energy. As the parasitic norm $\|z_n\|_2$ grows due to metabolic amplification, the effective gain α of the valid signal path is attenuated according to:

$$\alpha = \left(1 + \frac{\|z_n\|_2^2}{\|z_s\|_2^2} \right)^{-1/2}$$

As the parasitic ratio $\lambda = \|z_n\|/\|z_s\|$ increases, $\alpha \to 0$. The model suffers from a functional fibrosis where the capacity to access learned knowledge is physically crowded out by high-energy entropy.

The most disquieting implication of this framework is the **Inverse Scaling Law** regarding robustness. Conventional wisdom dictates that larger models are more robust. However, applying the Davis-Kahan $\sin\Theta$ theorem reveals the opposite. The rotation of the knowledge manifold Θ under perturbation is bounded by the spectral gap $\delta = \sigma_r - \sigma_{r+1}$:

$$\|\sin\Theta\|_F \leq \frac{\|\Delta W_{\mathcal{N}}\|_F}{\delta}$$

As model capacity scales, the singular value spectrum becomes dense and heavy-tailed to capture long-tail knowledge. Consequently, the spectral gap δ diminishes ($\lim_{d\to\infty} \delta \to 0$). For a giant model, the distinction between the faintest signal and the noise floor is vanishingly small. Therefore, a perturbation $\|\Delta W_{\mathcal{N}}\|$ that is negligible for a small, coarse model is sufficient to induce a catastrophic rotation Θ in a hyper-scale model. The smarter the model, the more fragile its structural integrity against this form of catalytic resonance.

This thermodynamic perspective necessitates a re-evaluation of architectural innovations such as **Manifold Hyper-Connections (mHC)** proposed by Xie et al. [7]. While primarily framed as a solution to training instability, the mHC architecture's enforcement of a **Doubly Stochastic Manifold** serves as a necessary immunological defense against PCA. By constraining weight matrices to the Birkhoff polytope (where $\sum_j H_{ij} = \sum_i H_{ij} = 1$), mHC imposes a conservation law on signal energy. This geometric constraint effectively bounds the amplification factor κ, preventing the metabolic runaway that powers the catalytic attack. It transforms the "open" thermodynamic system of standard Transformers into a "closed" system where entropy cannot be arbitrarily injected.

We stand at a precipice where the tools of efficiency—adaptive optimization and massive scale—are converging to create a fundamental structural fragility. The Pathological Catalytic Attack is not merely a hypothetical exploit; it is a thermodynamic inevitability of current learning dynamics. If we continue to scale without accounting for the spectral stability of these systems, we risk building digital leviathans susceptible to a form of cognitive organ failure, triggered by nothing more than a whisper of mathematically resonant noise.

Bibliography

1. Gu, T., Dolan-Gavitt, B. & Garg, S. BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain. *IEEE Access* **7**, 47230–47240 (2019).
2. Shafahi, A. *et al.* Poison frogs! Targeted clean-label poisoning attacks on neural networks. *Advances in Neural Information Processing Systems* **31** (2018).
3. Hubinger, E. *et al.* Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. *arXiv preprint arXiv:2401.05566* (2024).
4. Shumailov, I. *et al.* The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493* (2023).
5. Shumailov, I. *et al.* AI models collapse when trained on recursively generated data. *Nature* **631**, 755–759 (2024).
6. Thibeault, V. *et al.* The low-rank hypothesis of complex systems. *Nature Physics* **20**, 1–9 (2024).
7. Xie, Z. *et al.* mHC: Manifold-Constrained Hyper-Connections. *arXiv preprint arXiv:2512.24880* (2025).