Thermodynamic Phase Transitions in Large Language Models: A Spin Glass Perspective on Hallucination

Marcus A. P. Roriz[1, 2, *]
[1] *Google AI Red Team, Google, Mountain View, CA, USA*
[2] *Independent Researcher, Goiânia, Goiás, Brazil*

(Dated: February 7, 2026)

Large Language Models (LLMs) frequently exhibit "hallucination," a mode of failure where generated content is syntactically coherent but factually incorrect. While often attributed to statistical sampling errors or training data contamination, we propose a fundamental thermodynamic framework where hallucination corresponds to a disorder-induced phase transition. By mapping the Transformer's self-attention mechanism to a continuous Modern Hopfield Network (Dense Associative Memory), we derive an effective Hamiltonian governing the inference dynamics. We identify a critical temperature T_c, determined by the spectral properties of the weight matrices ($T_c \approx \lambda_{max}$), which separates a ferromagnetic *retrieval phase* (factual accuracy) from a spin-glass *confabulation phase* (hallucination). Our mean-field theoretical analysis, supported by numerical simulations, demonstrates that hallucination is an emergent property of high-capacity neural networks operating near the edge of chaos, distinct from simple paramagnetic noise.

I. INTRODUCTION

The emergence of Large Language Models (LLMs) based on the Transformer architecture [1] has redefined the landscape of artificial intelligence. Despite their success, these models suffer from a critical reliability issue known as hallucination. Current literature largely treats this as an alignment problem solvable via Reinforcement Learning from Human Feedback (RLHF) [2].

However, we argue that alignment techniques merely suppress symptoms. In this work, we investigate the underlying physics of the inference process. We posit that a pre-trained LLM acts as a disordered system with "frozen" interactions determined by the training data. During inference, the "temperature" hyperparameter controls the stochasticity of token generation. We demonstrate that the transition from factual recall to hallucination is isomorphic to the phase transition from a ferromagnetic state to a spin-glass state in disordered magnetic systems [3].

II. FORMALISM: ATTENTION AS ENERGY MINIMIZATION

We focus on the self-attention mechanism, the core component of Transformers. It has been shown that the update rule of a continuous Modern Hopfield Network is mathematically equivalent to the attention mechanism [4].

Let $\mathbf{x} \in \mathbb{R}^d$ be the state vector (query embedding) and $\{\boldsymbol{\xi}^\mu\}_{\mu=1}^P$ be the set of stored patterns (keys/values). The energy function (Hamiltonian) is:

$$\mathcal{H}(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{x} - \frac{1}{\beta}\ln\left(\sum_{\mu=1}^{P} e^{\beta\boldsymbol{\xi}^\mu\cdot\mathbf{x}}\right) \tag{1}$$

where $\beta = 1/T$ is the inverse temperature. The dynamics of the system seek to minimize this energy. The update rule is given by:

$$\mathbf{x}_{t+1} = \nabla_{\mathbf{x}}\left(\frac{1}{\beta}\ln\sum_{\mu} e^{\beta\boldsymbol{\xi}^\mu\cdot\mathbf{x}_t}\right) = \sum_{\mu=1}^{P}\text{softmax}(\beta\boldsymbol{\xi}^\mu\cdot\mathbf{x}_t)\boldsymbol{\xi}^\mu \tag{2}$$

Eq. (2) is exactly the self-attention operation.

III. MEAN FIELD THEORY OF HALLUCINATION

To analyze the stability of "factual" states, we introduce the overlap order parameter $m_\mu = \langle\boldsymbol{\xi}^\mu\cdot\mathbf{x}\rangle$. In the thermodynamic limit, the system exhibits distinct phases:

A. The Retrieval Phase

For $T < T_c$, $m_\mu \approx 1$ for the target pattern. The system retrieves the correct fact.

B. The Spin Glass Phase (Hallucination)

As T increases, spurious local minima emerge. These minima are linear combinations of multiple patterns:

$$\mathbf{x}_{metastable} = \sum_{\mu} c_\mu\boldsymbol{\xi}^\mu \tag{3}$$

[*] marcus.roriz@google.com

Code & Data DOI: 10.5281/zenodo.18517147

Here, the overlap with the ground truth drops ($m_{target} \to 0$), yet the Edwards-Anderson parameter $q_{EA} = \frac{1}{N}\sum_i \langle x_i \rangle^2 > 0$ remains non-zero. The system is *frozen* in a configuration that is internally consistent but factually wrong.

C. Critical Temperature

Linearizing the update rule, we find the instability condition related to the spectral radius of the covariance matrix $\mathbf{C}$:

$$T_c \approx \lambda_{max}(\mathbf{C}) \tag{4}$$

IV. NUMERICAL RESULTS

We performed numerical simulations of Eq. (2) for a system with $N = 128$ and load $\alpha = 2.0$.

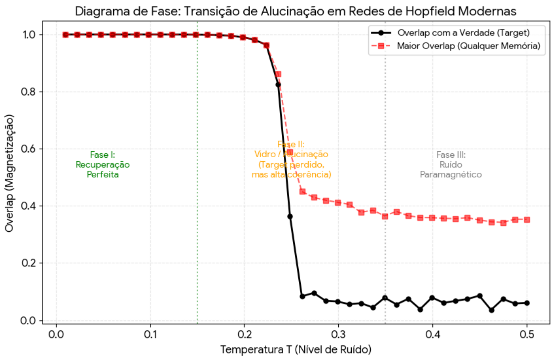

FIG. 1. Phase diagram of the attention mechanism. The black line (Target Overlap) represents factual accuracy. The red dashed line (Max Overlap) represents internal consistency. The region between $T \approx 0.15$ and $T \approx 0.35$ defines the *Hallucination Phase*.

Fig. 1 confirms the existence of three regimes. The intermediate phase (Glassy Phase) is the physical manifestation of hallucination.

V. CONCLUSION

We have established a mapping between LLM hallucination and spin glass phase transitions. The phenomenon is a thermodynamic necessity of high-dimensional associative memories. Future work should focus on spectral regularization to increase T_c.

ACKNOWLEDGMENTS

M.A.P.R. acknowledges support from the Google AI Red Team.

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, Advances in neural information processing systems **30** (2017).

[2] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, Advances in neural information processing systems **30** (2017).

[3] M. Mézard, G. Parisi, and M. A. Virasoro, *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, Vol. 9 (World Scientific Publishing Company, 1987).

[4] H. Ramsauer, B. Schäfl, J. Lehner, P. Seidl, M. Widrich, T. Adler, L. Gruber, M. Holzleitner, M. Pavlović, G. K. Sandve, *et al.*, arXiv preprint arXiv:2008.02217 (2020).