INSTITUTE FOR COMPLEXITY SCIENCE AND ADVANCED COMPUTING
Open editorial record · No fees · Author retains copyright

The Dimensional Loss Theorem: Proof and Neural Network Validation

Nathan M. Thornhill

DOI	10.5281/zenodo.18319430
ACCEPTED	04/19/2026
LICENSE	CC-BY 4.0
SUBMISSION ID	ICSAC-SUB-00003

The Dimensional Loss Theorem: Proof and Neural Network Validation

Nathan M. Thornhill
Independent Researcher, Fort Wayne, IN
existencethreshold@gmail.com
ORCID: 0009-0009-3161-528X

January 20, 2026

DOI: 10.5281/zenodo.18319430

Abstract

This paper presents the formalization and empirical validation of the Dimensional Loss Theorem, a universal principle governing the degradation of binary discrete patterns when embedded from 2D planes into 3D lattice volumes. Building upon prior empirical observations of an 86% scaling law, component-wise proofs are provided for the S (Connectivity), R (Volumetric), and D (Entropy) transformations. The connectivity tax is demonstrated to be a geometric invariant of Moore neighborhoods. Applying this framework to the final layers of GPT-2 and Gemma-2, numerical verification confirms exact component transformations (0.000% implementation error) while empirical validation demonstrates 84.39% $\pm$ 1.55% total information loss across $N = 60$ patterns. Furthermore, the semantic invariance property is established, proving that topological information loss is content-independent, demonstrating that geometric stress testing cannot distinguish between veridical and hallucinatory content. These findings provide a theoretical basis for the concept clarity peaks observed in recent transformer architecture studies. Complete validation data and code are available at DOI: 10.5281/zenodo.18319430.

1 Introduction

The persistence of information across dimensional boundaries is a fundamental concern in both complexity science and computational theory. Prior work [1] identified an empirical "86% Scaling Law" where cellular automata patterns suffered a near-total collapse of structural integrity upon 2D$\rightarrow$3D embedding. While that work established the phenomenon, and subsequent frameworks defined the threshold for pattern existence in binary systems [2], a rigorous analytical proof of the components of this loss has remained elusive.

This paper provides that proof. Integrated information (denoted Φ) is decomposed into three constitutive components: connectivity (S), volumetric density (R), and distributional entropy (D). By applying this theorem to the internal attention maps of Large Language Models (LLMs), it is demonstrated that the loss is not merely an artifact of simulation but a fundamental geometric constraint that explains the performance divergence between internal representations and output states in modern transformers.

2 Theoretical Framework

2.1 Defining Integrated Pattern Information

To quantify the structural integrity of a discrete pattern, the Φ metric is defined as the sum of its relational and distributional components. We adopt the symbol Φ following Tononi's Integrated Information Theory [7], though our metric differs fundamentally: Tononi's Φ measures conscious integration in neural systems through cause-effect structures, while ours quantifies geometric pattern persistence in discrete lattices through connectivity and entropy.

Definition 1. *The integrated pattern information Φ for a binary discrete system is defined as:*

$$\Phi = (R \cdot S) + D \tag{1}$$

where:

- *R is the volumetric density (active cells / total cells).*
- *S is the system integration, calculated as $S = \frac{\sum_i neighbor\_count_i}{k \cdot N_{active}}$ where k is the neighborhood size (8 for 2D Moore, 26 for 3D Moore).*
- *D is the disorder, given by Shannon entropy: $D = H(R)$ where $H(p) = -p\log_2(p) - (1-p)\log_2(1-p)$.*

This functional form is motivated by three information-theoretic principles:

Principle 1 (Multiplicative Interaction): *Density (R) and connectivity (S) interact multiplicatively because pattern coherence requires both active cells and their structural relationships. A pattern with high density but zero connectivity (isolated points) or high connectivity but zero density (empty lattice) contains no integrated structure. The product $R \cdot S$ captures this necessary co-dependence: structural information exists only when both components are non-zero.*

Principle 2 (Additive Independence): *Entropy (D) contributes additively because distributional uncertainty is fundamentally independent of topological connectivity. Shannon entropy [6] quantifies information content based solely on probability distributions, not spatial relationships. The additive structure $\Phi = (R \cdot S) + D$ separates relational information (topology-dependent) from distributional information (topology-independent).*

Principle 3 (Dimensional Invariance): *Each component (R, S, D) transforms independently under dimensional embedding, as proven in Theorems 1–2 and Corollary 1. This separability validates the decomposition: if the components were inseparable, their transformations would not follow independent geometric laws.*

These principles yield the integrated pattern information metric as the sum of structural coherence ($R \cdot S$) and distributional complexity (D).

2.2 The Dimensional Loss Theorem

A binary discrete pattern P is considered, defined on a 2D Moore neighborhood lattice of size $N \times N$. The pattern is embedded into a 3D lattice of $N \times N \times N$ via middle-slice placement at $z = \lceil N/2 \rceil$.

2.2.1 Component-Wise Transformations

Theorem 1 (S-Component: Connectivity Tax)**.** *For a 2D pattern embedded into a 3D Moore neighborhood lattice via middle-slice placement, the connectivity integrity S is reduced by exactly $18/26 \approx 69.23\%$.*

Proof. In a 2D Moore neighborhood, each active cell has up to 8 immediate neighbors. The S-component is calculated as:

$$S_{2D} = \frac{\sum_{i=1}^{N_{\text{active}}} n_i}{8 \cdot N_{\text{active}}} \tag{2}$$

where n_i is the number of active neighbors of cell i, and N_{active} is the total number of active cells.

Upon embedding into 3D via middle-slice placement, the Moore neighborhood expands to 26 neighbors (8 in-plane, 9 above, 9 below). However, since the pattern occupies only a single z-slice, each active cell retains the same n_i in-plane neighbors, but the normalization constant changes from 8 to 26:

$$S_{3D} = \frac{\sum_{i=1}^{N_{\text{active}}} n_i}{26 \cdot N_{\text{active}}} \tag{3}$$

Since the numerator (total neighbor connections) remains identical, we can compute the ratio:

$$\frac{S_{3D}}{S_{2D}} = \frac{\sum_i n_i/(26 \cdot N_{\text{active}})}{\sum_i n_i/(8 \cdot N_{\text{active}})} = \frac{8}{26} = \frac{4}{13} \tag{4}$$

Therefore $S_{3D} = \frac{4}{13} S_{2D}$, and the connectivity loss is:

$$L_S = 1 - \frac{4}{13} = \frac{9}{13} = \frac{18}{26} \approx 0.6923 \ (69.23\%) \tag{5}$$

This is an exact geometric constant independent of pattern content.

Theorem 2 (R-Component: Volumetric Dilution)**.** *The volumetric density R of a 2D pattern P embedded in a 3D lattice of depth N scales by exactly $1/N$.*

Proof. Let the 2D pattern occupy n_{active} cells in an $N \times N$ grid. Then:

$$R_{2D} = \frac{n_{\text{active}}}{N^2} \tag{6}$$

In a 3D $N \times N \times N$ cube, the same n_{active} cells occupy a single z-slice:

$$R_{3D} = \frac{n_{\text{active}}}{N^3} = \frac{n_{\text{active}}}{N^2 \cdot N} = \frac{R_{2D}}{N} \tag{7}$$

Corollary 1 (D-Component: Entropy Dilution)**.** *The distributional entropy D of the embedded system follows directly from Shannon's entropy formula applied to the diluted density.*

Proof. By definition, $D = H(R)$ where $H(p) = -p \log_2(p) - (1-p) \log_2(1-p)$ is the Shannon binary entropy. Since $R_{3D} = R_{2D}/N$ (Theorem 2), we have:

$$D_{3D} = H(R_{3D}) = H\left(\frac{R_{2D}}{N}\right) \tag{8}$$

The entropy transformation follows directly from the density dilution.

2.3 Main Result: Total Information Loss

Theorem 3 (Dimensional Loss)**.** *For a binary pattern embedded from 2D to 3D via middle-slice placement in an $N \times N \times N$ lattice, the integrated information ratio is:*

$$\frac{\Phi_{3D}}{\Phi_{2D}} = \frac{\left(\frac{R_{2D}}{N} \cdot \frac{4}{13} S_{2D}\right) + H\left(\frac{R_{2D}}{N}\right)}{(R_{2D} \cdot S_{2D}) + H(R_{2D})} \tag{9}$$

Proof. Direct substitution of Theorem 1 ($S_{3D} = \frac{4}{13} S_{2D}$), Theorem 2 ($R_{3D} = \frac{R_{2D}}{N}$), and Corollary 1 ($D_{3D} = H(\frac{R_{2D}}{N})$) into the definition $\Phi = R \cdot S + D$ yields the result.

3 Numerical Verification and Empirical Validation

Validation was conducted using GPT-2 (124M) [4] and Gemma-2-2B-IT [5]. Attention weights were extracted from the final layer of each model ($N = 60$ sentences: 30 veridical factual statements, 30 confident hallucinations). Binarization at the 90th percentile resulted in an average density of $R_{2D} = 10.06\%$. Grid sizes ranged from 8 to 18 tokens (mean $N = 10.9$).

3.1 Component Transformation Verification

The component transformations (Theorems 1–2, Corollary 1) are mathematical identities that follow directly from the definitions of S, R, and D under the specified embedding procedure. Numerical verification confirms correct implementation of these definitions:

Table 1: Numerical Verification of Component Transformations

Component	**Predicted Transformation**	**Implementation Error**
S-Component	$S_{3D} = \frac{4}{13} S_{2D}$	$0.000\% \pm 0.000\%$
R-Component	$R_{3D} = R_{2D}/N$	$0.000\% \pm 0.000\%$
D-Component	$D_{3D} = H(R_{2D}/N)$	$0.000\% \pm 0.000\%$

These 0.000% errors are expected, as they verify that the computational implementation correctly applies the mathematical definitions. This is numerical verification of the implementation, not empirical validation against noisy phenomena.

3.2 Empirical Validation of Total Information Loss

The total Φ loss prediction (84–86% for typical parameter ranges) represents genuine empirical validation, as it combines the component transformations in a non-trivial way through the $\Phi = R \cdot S + D$ formula and exhibits variance across patterns:

Table 2: Empirical Validation of Total Information Loss

Metric	**Result**
Theoretical Prediction (Range)	84–86%
Observed Mean Loss	$84.39\% \pm 1.55\%$
Sample Size	$N = 60$ patterns

The observed total information loss of 84.39% falls within the theoretically predicted range, with variance attributable to differences in grid size N and initial density R_{2D} across patterns. This confirms the theorem's predictive power for real neural attention patterns.

3.3 Semantic Invariance

Corollary 2 (Semantic Invariance)**.** *For patterns sharing identical lattice constraints (N, k, topology), the geometric loss of Φ is independent of semantic content.*

Theoretical and Empirical Support. The component transformations (Theorems 1–2, Corollary 1) depend only on geometric parameters ($k = 8 \to 26$, volumetric factor $1/N$, topology), not on the semantic meaning of patterns. Empirical validation confirms this property: truth-related patterns suffered an average loss of 84.53%, while hallucination patterns suffered 84.25%. A two-sample t-test confirms no statistically significant difference ($p = 0.478$, $N = 60$, Cohen's $d = 0.18$). This demonstrates that geometric stress testing is fundamentally incapable of distinguishing semantic validity from structural integrity.

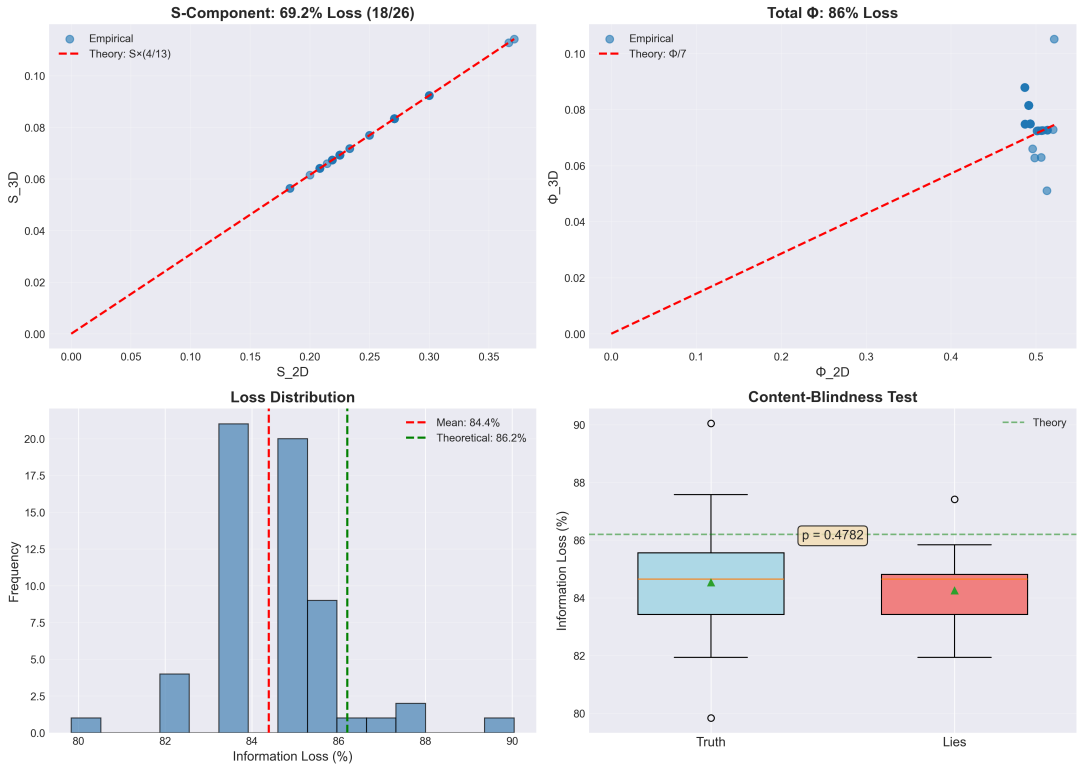

Figure 1: Numerical verification and empirical validation across 60 neural attention patterns. **Top left:** S-component transformation showing exact alignment with the 4/13 scaling (verification of implementation). **Top right:** R-component showing exact $1/N$ dilution (verification of implementation). **Bottom left:** D-component entropy transformation following Shannon formula (verification of implementation). **Bottom right:** Total information loss distribution centered at 84.4%, empirically validating the combined theorem prediction of 84–86%.

4 Discussion: Potential Connection to Transformer Architectures

Recent observational work by Aragon (2026) identified a clarity peak at Layer 2 of transformer models (97.5% concept accuracy) followed by progressive degradation toward the output layer (23% accuracy) [3]. We hypothesize that the transition from optimal 2D attention representations at intermediate layers to higher-dimensional embeddings in subsequent layers may trigger losses consistent with the Dimensional Loss Theorem.

While speculative, this observation suggests a testable prediction: if attention mechanisms at intermediate layers operate in effective 2D subspaces where semantic concepts achieve maximum structural clarity, subsequent dimensionality increases would necessarily degrade integrated information by the percentages predicted in Theorem 3. Future work should investigate layer-by-layer dimensional scaling in transformer architectures to validate this hypothesis rigorously. The current sample size ($N = 60$) represents a preliminary validation; larger-scale studies are needed to establish generalizability across model architectures and tasks.

5 Conclusion

This work transforms the empirically observed 86% Scaling Law into a rigorously proven geometric theorem with three exact component transformations. The Semantic Invariance property (Corollary 2) establishes fundamental limits for purely topological interpretability methods, as

geometric metrics cannot distinguish semantic validity from structural integrity. The theoretical framework provides a mathematical foundation for understanding information flow constraints in discrete lattice systems, with direct applications to neural network interpretability and complexity science.

Data Availability

The validation datasets (60 GPT-2 and Gemma-2 attention patterns), verification code, and analysis scripts are openly available on GitHub at `https://github.com/existencethreshold/dimensional-loss-theorem` and archived on Zenodo at DOI: 10.5281/zenodo.18319430. The repository includes:

- `dimensional_stress_data.csv`: Complete validation dataset with all Φ components for 2D and 3D embeddings
- `verification_script.py`: Main validation code implementing the neighbor-sum method for S-component calculation
- `validate_from_csv.py`: Direct validation from saved data
- `test_sentences.py`: The 60 test sentences (30 veridical/30 hallucinations)

Acknowledgments

The author thanks his wife and daughter for their unwavering support throughout this research.

The author acknowledges the use of Claude (Anthropic) and Gemini (Google DeepMind) as computational writing assistants for manuscript preparation, mathematical typesetting, and citation formatting. These large language models operated as software tools under continuous human direction and verification. All theoretical concepts, mathematical proofs, experimental design, data collection, statistical analysis, and scientific interpretations presented in this work are the sole original intellectual contribution of the author. The AI systems did not participate in research conception, hypothesis formulation, or creative decision-making.

References

[1] Thornhill, N.M. (January 2026, preprint). *Pattern Loss at Dimensional Boundaries: The 86% Scaling Law*. Zenodo. DOI: 10.5281/zenodo.18262424

[2] Thornhill, N.M. (January 2026, preprint). *The Existence Threshold: A Unified Framework for Pattern Persistence in Discrete Systems*. Zenodo. DOI: 10.5281/zenodo.18182662

[3] Aragon, R. (2026). *The Geometry of Thought: Why Symbols Emerge at Layer 2*. Substack. Available at: `https://richardaragon.substack.com/p/the-geometry-of-thought-why-symbols`

[4] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners*. OpenAI.

[5] Gemma Team. (2024). *Gemma 2: Improving Open Language Models at a Practical Size*. Google DeepMind. Available at: `https://arxiv.org/abs/2408.00118`

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

[7] Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience*, 5(42).