Architecture-Independent Geometric Memory Failure

A synthesis of Thornhill 2026b/c (January) and Barman, Starenky, Bodnar, Narasimhan, Gopinath 2026 (March)

Nathan M. Thornhill

05/15/2026

Institute for Complexity Science and Advanced Computing (ICSAC)
Fort Wayne, Indiana, United States
ORCID: 0009-0009-3161-528X
Contact: nthornhill@icsacinstitute.org · research@nathanthornhill.com

Deposited: 05/15/2026 · License: CC-BY-4.0

Abstract

In January 2026 two papers were deposited on Zenodo establishing that information loss at dimensional boundaries in discrete systems is a geometric phenomenon with an architecture-independent magnitude: 86.01% ± 2.39% in cellular automata across 1,500 patterns (Thornhill 2026b, DOI 10.5281/zenodo.18262424, 01/14/2026), and 84.39% ± 1.55% on transformer hidden states (GPT-2, Gemma-2), supported by a formal proof of the component transformations S, R, and D (Thornhill 2026c, DOI 10.5281/zenodo.18319430, 01/20/2026). It was predicted, in the closing discussion of Thornhill 2026c, that the geometric account should hold across substrates wherever density dilution and neighborhood-structure expansion occur together at a representational boundary.

In March 2026, Barman, Starenky, Bodnar, Narasimhan, and Gopinath independently published two arXiv preprints (arXiv:2603.27116 and arXiv:2604.06222) reporting that production retrieval embedding models — MiniLM-L6-v2, BGE-base, BGE-large — concentrate their variance into approximately 16 effective dimensions regardless of nominal dimensionality (384, 768, 1024), and that this concentration places those models in an interference-vulnerable geometric regime that reproduces quantitative signatures of human memory failure (power-law forgetting with exponent $b = 0.460 \pm 0.183$, Deese–Roediger–McDermott false-alarm rate of 0.583, spacing-effect ordering, tip-of-tongue behavior). They establish a parallel theorem — the No-Escape Theorem — characterizing what cannot be repaired within semantically continuous kernel-threshold memory systems.

The two bodies of work are methodologically distinct. They use different metrics ($\Phi = R \cdot S + D$ vs. participation ratio), study different substrates (cellular automata and transformer hidden states vs. pretrained retrieval embeddings), and report different specific quantities (an 86% loss

constant in Φ vs. a fixed point at ~16 effective dimensions across nominal sizes). They also reach the same broader conclusion from independent directions: that representational memory failure is a geometric property of the embedding operation, not a property of any particular architecture, training regime, or biological substrate.

The present note records the chronology of the two lines of evidence in a single citable document, summarizes the methodological differences, and identifies the substantive convergence: an architecture-independent geometric fixed point as the principal explanatory mechanism for representational memory failure in the systems studied.

---

1. Background

1.0 Chronology of the Public Record

Date	Event	Venue / Identifier
01/14/2026	**Thornhill 2026b** — *Pattern Loss at Dimensional Boundaries: The 86% Scaling Law* — first quantification of an architecture-independent dimensional-loss constant (86.01% ± 2.39%) across 1,500 cellular-automata patterns, three dimensional transitions, five grid sizes, and two distinct rule sets	Zenodo, DOI 10.5281/zenodo.18262424
01/20/2026	**Thornhill 2026c** — *The Dimensional Loss Theorem: Proof and Neural Network Validation* — formal proof of the component transformations $S \to (4/13) \cdot S$, $R \to R/N$, $D \to H(R/N)$, and empirical replication on GPT-2 and Gemma-2 hidden states at 84.39% ± 1.55%; closing discussion predicts substrate-universality of the geometric account	Zenodo, DOI 10.5281/zenodo.18319430
03/28/2026	**Barman, Starenky, Bodnar, Narasimhan, Gopinath** — *The Price of Meaning: Why Every Semantic Memory System Forgets* — No-Escape Theorem for interference-driven forgetting in	arXiv:2603.27116

Date	Event	Venue / Identifier
	semantically continuous, kernel-threshold memory systems	
03/27/2026	**Barman, Starenky, Bodnar, Narasimhan, Gopinath** — *The Geometry of Forgetting* — empirical observation that production embedding models (MiniLM, BGE-base, BGE-large) concentrate variance to a fixed point of ~16 effective dimensions regardless of nominal size; quantitative reproduction of human memory phenomena (Ebbinghaus power law, DRM false memories, spacing, tip-of-tongue) from the geometry of pretrained embedding space	arXiv:2604.06222
05/15/2026	**Present note** — synthesis of the two parallel lines of evidence; identifies the substantive convergence as architecture-independent geometric memory failure across distinct substrates and methodologies	Zenodo (this deposit)

The two bodies of work proceeded independently. Barman et al. do not cite the Zenodo deposits; their reference list of 24 entries draws from the cognitive-science (Ebbinghaus, Roediger & McDermott, Anderson, Baddeley, Wixted) and ML-embedding (Sentence-BERT, BGE C-Pack, attention-is-all-you-need, catastrophic-forgetting-in-neural-networks) traditions. The present note records both lines of evidence in a single citable document.

1.1 The January 2026 Work

The dimensional-loss constant was established in two papers deposited on Zenodo in January 2026, building on the persistence framework $\Phi = R \cdot S + D$ originally introduced in Thornhill 2026a (*The Existence Threshold*, DOI 10.5281/zenodo.18166974) and later generalized across substrates in Thornhill 2026d (*The Dynamic Existence Threshold*, DOI 10.5281/zenodo.18373411).

Thornhill 2026b — "Pattern Loss at Dimensional Boundaries: The 86% Scaling Law" (published 01/14/2026, DOI 10.5281/zenodo.18262424) reported a controlled experiment on 1,500 random binary patterns across three dimensional transitions (1D→2D, 2D→3D, 3D→4D), five grid resolutions (15 to 25 cells per side), and two distinct cellular-automata rule sets (Conway's Game of Life, HighLife). The integration–differentiation persistence functional $\Phi = R \cdot S + D$ was measured before and after middle-placement embedding into the next-higher-dimensional grid.

Loss across all 1,500 patterns averaged **86.01% ± 2.39%**, with a coefficient of variation of 2.8%. The mean loss varied across the three transitions by less than half a percentage point (1D→2D: 85.82%; 2D→3D: 86.09%; 3D→4D: 86.13%). Rule set, scale, and content had no measurable effect.

Decomposing Φ into its integration component ($R \cdot S$) and its differentiation component (D), the loss is structured: $R \cdot S$ collapses by 99.6%; D drops by 82–83%. The pattern's coordination is destroyed; its statistical signature is severely degraded but not erased. After the first dimensional crossing, Φ stabilizes near a floor at $\Phi \approx 0.169$ — roughly one-sixth of the maximum for random binary input.

Thornhill 2026c — "The Dimensional Loss Theorem: Proof and Neural Network Validation" (published 01/20/2026, DOI 10.5281/zenodo.18319430) provided a formal derivation of the loss structure for the 2D→3D embedding case:

$S \to (4/13) \cdot S$ (coupling normalization expands from 8 to 26 Moore neighbors)

$R \to R/N$ (density dilution as the pattern occupies one slice of N possible slices)

$D \to H(R/N)$ (Shannon entropy of the new occupancy ratio)

Empirical validation on transformer hidden states (GPT-2 and Gemma-2, N = 60 patterns) yielded **84.39% ± 1.55%** — differing from the cellular-automata constant by approximately 1.6 percentage points on a substrate with no architectural relationship to the original experiment. Numerical verification of the component transformations matched the predicted values to 0.000% implementation error, as reported in the deposited abstract.

The closing discussion of Thornhill 2026c argued that the loss mechanism — density dilution combined with neighborhood-structure expansion at a representational boundary — is geometric rather than substrate-specific, and predicted that architecture-independent geometric fixed points should be observable across other classes of system.

1.2 The March 2026 Parallel Finding

In March–April 2026, Barman, Starenky, Bodnar, Narasimhan, and Gopinath published two arXiv preprints reporting parallel work on the geometry of representational memory failure in production embedding models.

In "The Geometry of Forgetting" (arXiv:2604.06222), the authors report three principal empirical results:

1. **Power-law forgetting under interference.** Using an embedding-based memory system queried under simulated Ebbinghaus conditions (1,000 facts spanning 30 simulated days, with 10,000 distractors as competitors), the fitted forgetting exponent is $\mathbf{b = 0.460 \pm 0.183}$ (95% CI [0.354, 0.644]), close to the human value of $b \approx 0.5$. Without competitors the exponent drops to $b \approx 0.009$ — fifty times smaller. The authors conclude that interference, not temporal decay, is the dominant driver of the human-like forgetting curve in this geometry.

2. **A fixed-point effective dimensionality across the tested embedding models.** Computing the participation ratio d_eff = $(\Sigma\lambda_i)^2$ / $\Sigma\lambda_i^2$ across three production embedding models (all transformer-based dense retrieval architectures) reveals model-independent concentration:

Model	Nominal d	Effective d (participation ratio)
MiniLM-L6-v2	384	**15.7 ± 0.0**
BGE-base	768	**16.6 ± 0.1**
BGE-large	1024	**16.3 ± 0.1**

   Only 17–18 principal components are needed to account for 95% of variance regardless of nominal dimensionality. The authors term this the *dimensionality illusion* and identify it as the explanation for interference vulnerability in production retrieval systems. The result is established across the three tested production embeddings; broader generalization across architecturally distinct embedding paradigms is consistent with the data but not directly demonstrated.

3. **Quantitative reproduction of human false-memory phenomena.** The Deese–Roediger–McDermott paradigm, applied to a 1024-dimensional retrieval model across all 24 published DRM word lists, yields a critical-lure false alarm rate of **0.583** (human ~0.55) with zero parameter tuning. Spacing effects and tip-of-tongue behaviors also emerge without phenomenon-specific engineering.

In "The Price of Meaning" (arXiv:2603.27116), the same authors establish what they term the **No-Escape Theorem**: that interference-driven forgetting and false recall cannot be eliminated within semantically continuous, kernel-threshold memory systems without abandoning semantic organization or adding external symbolic structure. They test the theorem across five memory-system implementations within the semantic kernel-threshold paradigm and find consistent expression of the vulnerability.

---

2. Comparison of the Two Lines of Evidence

The two bodies of work differ in metric, substrate, and specific reported quantity. They share an architecture-independence claim and a geometric explanatory mechanism. The differences and overlaps are summarized below.

	Thornhill 2026b / 2026c (January 2026)	Barman et al. 2026 (March–April 2026)
Metric	$\Phi = R \cdot S + D$ (integration plus differentiation persistence)	Participation ratio d_eff = $(\Sigma\lambda_i)^2$ / $\Sigma\lambda_i^2$
Substrate		

	Thornhill 2026b / 2026c (January 2026)	Barman et al. 2026 (March–April 2026)
	Cellular automata; transformer hidden states (GPT-2, Gemma-2)	Pretrained retrieval embedding models (MiniLM, BGE-base, BGE-large)
Sample / scale	n = 1,500 CA patterns + n = 60 transformer encodings	3 production embedding models + simulated retrieval over 1,000–50,000 items
Reported invariant	86.01% ± 2.39% Φ loss at every dimensional embedding boundary; floor at $\Phi \approx 0.169$	d_eff ≈ 16 across nominal sizes 384–1,024; b ≈ 0.460 under interference
Formal result	Dimensional Loss Theorem: component-wise proof of $S \to (4/13) \cdot S$, $R \to R/N$, $D \to H(R/N)$	No-Escape Theorem: interference cannot be repaired within semantic kernel-threshold systems
Architecture independence	Rule set, scale, and grid size produce no measurable difference; GPT-2 and Gemma-2 fall within one SD of the CA result	Nominal dimensionality varies by 2.67× across three models; effective dimensionality is essentially constant
Substrate independence	Demonstrated across two structurally unrelated substrates (CA, transformer hidden states)	Demonstrated across three pretrained embedding architectures of differing nominal size

The two specific quantities — an 86% Φ-loss constant and a 16-effective-dimensional fixed point — are not numerically equivalent under any straightforward conversion, and the present note does not claim they are. They are measured by different functionals on different objects, and the magnitudes (86% Φ loss = 14% retention, vs. $16/1024 \approx 1.6\%$ retention of nominal dimensions) do not coincide.

What the two findings share, and what is substantively new in the combined record, is the form of the result: an **architecture-independent geometric fixed point** that holds across structurally unrelated systems and is invariant to scale, training regime, and rule set. Both works arrive at this form from independent methodological directions.

---

3. The Geometric Account

The Dimensional Loss Theorem (Thornhill 2026c) and the No-Escape Theorem (Barman et al. 2026) make different formal claims and operate on different objects. They are complementary rather than competing.

The Dimensional Loss Theorem describes what happens to organizational information at a dimensional boundary. For the 2D→3D case, the connectivity normalization expands from 8 to 26 Moore neighbors, dropping the coupling component S by the factor 4/13; the density component R falls by 1/N as the pattern occupies one slice of N; the differentiation component D contracts via Shannon entropy of the new occupancy ratio. The theorem is mechanistic and predictive — it tells the reader why the loss occurs, how much loss to expect, and which components of the organizational structure are damaged most severely.

The No-Escape Theorem describes what cannot be repaired within a semantic kernel-threshold memory system once the loss has occurred. It is a constraint on architectural exits: it tells the system designer that interference-driven forgetting and false recall cannot be eliminated without leaving the semantic-similarity paradigm. The theorem is descriptive and prescriptive — it tells the reader what is and is not architecturally available.

Read together the two theorems characterize complementary aspects of a single failure mode: the Dimensional Loss Theorem establishes the geometric cost of admission into an embedding space, and the No-Escape Theorem establishes the cost of remaining inside one. To the present author's knowledge, neither theorem subsumes the other. A complete account of representational memory failure requires both.

The empirical findings sit within this combined frame. Thornhill 2026b/c measure the loss as it occurs during the embedding operation; Barman et al. measure the residue from accumulated embedding operations in the static geometry of pretrained models, where the variance distribution has already been concentrated. Both observations are consistent with a geometric process that yields architecture-independent fixed points — though, as noted in §2, the specific quantities they report are not the same number under any direct conversion.

---

4. Discussion

4.1 What the Two Lines of Evidence Establish Together

The combined record supports a broader claim than either body of work establishes alone: that **representational memory failure is principally driven by geometric rather than architectural or substrate-specific factors**, and that the relevant geometric structures produce architecture-independent fixed points whose existence is robust across measurement choice.

This is a claim about the structure of the answer, not about the answer's specific value. Each body of work contributes a different specific quantity (86% Φ loss; 16-dimensional effective fixed point), and those quantities are likely measuring different aspects of an underlying geometric process. The convergence is at the level of *form — there is a substrate-independent geometric fixed point* — rather than at the level of *exact magnitude*.

Future work that bridges the two metrics — for example, computing participation ratio on the cellular-automata and transformer-hidden-state data from Thornhill 2026b/c, and computing $\Phi = R \cdot S + D$ on the embedding-space data from Barman et al. — would establish whether the underlying constants are related by a closed-form transformation, or whether the two findings measure genuinely distinct geometric properties that happen to share an architecture-independence character. The present note does not undertake that analysis.

4.2 Falsifiability

The combined architecture-independence claim is straightforwardly falsifiable. Each line of evidence is falsifiable on its own terms — by finding a substrate in which Φ loss at dimensional embedding deviates meaningfully from the 86% band, or by finding pretrained embedding models whose effective dimensionality scales with nominal dimensionality. The combined claim is falsifiable by finding a representational system in which neither geometric fixed point appears.

Both confirming and disconfirming evidence are informative. The geometric account is precise enough that careful measurement on any candidate substrate moves the result forward.

4.3 Implications for Retrieval-Augmented Systems

The Barman et al. measurement of effective dimensionality has direct implications for retrieval-augmented generation systems built on dense embedding similarity. The "dimensionality illusion" they identify — that production models advertise hundreds or thousands of dimensions but operate on far fewer — means such systems are operating in an interference-vulnerable geometric regime by construction.

This is consistent with the geometric account argued in Thornhill 2026c: that representational memory failure is not an engineering artifact to be patched at the implementation level, but a consequence of the geometry of the embedding operation itself. Engineering mitigations exist — hybrid sparse-dense retrieval, structured external memory, reranking, metadata constraints — but they operate by adding information that does not live in the embedding space alone. The two theorems considered jointly account for why this is the only available class of solution.

---

5. Acknowledgments

The careful empirical work on production embedding models reported by Barman, Starenky, Bodnar, Narasimhan, and Gopinath (Sentra) in their March–April 2026 arXiv preprints is acknowledged. The architecture-independence observation across three retrieval models, arrived at through a methodology distinct from the one used in Thornhill 2026b/c, strengthens the broader case for a geometric account of representational memory failure beyond what either line of evidence establishes alone.

The early members of ICSAC are also acknowledged for feedback on the framing of substrate universality and for review of the formal proof in Thornhill 2026c during the institute's open review pipeline.

References

Barman, S. R., Starenky, A., Bodnar, S., Narasimhan, N., & Gopinath, A. (2026). *The Price of Meaning: Why Every Semantic Memory System Forgets.* arXiv:2603.27116 [cs.LG]. Submitted 03/28/2026.

Barman, S. R., Starenky, A., Bodnar, S., Narasimhan, N., & Gopinath, A. (2026). *The Geometry of Forgetting.* arXiv:2604.06222 [q-bio.NC]. Submitted 03/27/2026.

Thornhill, N. M. (2026a). *The Existence Threshold: A Framework for Pattern Persistence in Binary Discrete Systems.* Zenodo. https://doi.org/10.5281/zenodo.18166974

Thornhill, N. M. (2026b). *Pattern Loss at Dimensional Boundaries: The 86% Scaling Law.* Zenodo. Published 01/14/2026. https://doi.org/10.5281/zenodo.18262424

Thornhill, N. M. (2026c). *The Dimensional Loss Theorem: Proof and Neural Network Validation.* Zenodo. Published 01/20/2026. https://doi.org/10.5281/zenodo.18319430

Thornhill, N. M. (2026d). *The Dynamic Existence Threshold: Integration-Differentiation Balance Predicts System State Across Substrates.* Zenodo. https://doi.org/10.5281/zenodo.18373411

Manuscript prepared at the Institute for Complexity Science and Advanced Computing (ICSAC), Fort Wayne, Indiana, United States.

Corresponding author: Nathan M. Thornhill — nthornhill@icsacinstitute.org · research@nathanthornhill.com — ORCID 0009-0009-3161-528X