Catastrophic Forgetting in Transformer Attention is Structurally Inevitable: A Transfer of the Hopfield Impossibility Theorem

Andreas Bean*

April 2026

Abstract

We prove that the structural impossibility of safe incremental concept addition established for dense associative memories in [1] transfers to Transformer attention via the Modern Hopfield Network correspondence [2]. The transfer is exact for *linear attention*: the key Gram matrix $G = K^\top K/d$ is the Hopfield weight matrix, and the impossibility theorem applies verbatim — any update ΔK that introduces a new concept as a fixed direction of $G' = G + \Delta G$ must shift at least one eigenvalue of G, disrupting the retrieval geometry of existing stored concepts. For standard *softmax attention*, the fixed-point disruption is exponentially suppressed at high inverse temperature β, but a weaker yet structurally identical result holds: the Gram geometry $\rho_{\mu\nu}(G) = k_\mu^\top G\, k_\nu$ is globally perturbed for all concept pairs (k_μ, k_ν) with nonzero overlap on the perturbed eigenvector — regardless of β, optimiser, or rank of the update. We state both results precisely, identify where the analogy is exact and where it is approximate, and discuss consequences for continual fine-tuning, knowledge editing, and LoRA.

1 Introduction

The impossibility theorem of [1] establishes that no weight perturbation ΔW to a symmetric dense weight matrix W can introduce a new attractor a^* without necessarily changing the eigenstructure of W. The result is purely algebraic: it uses only the symmetry of ΔW and the completeness of the eigenbasis, independent of any specific learning algorithm.

Transformer attention heads use a key matrix $K \in \mathbb{R}^{p\times d}$ whose rows $k_1, \ldots, k_p$ store the "concepts" available for retrieval. At first glance, no symmetric weight matrix appears in this picture. The key insight of Ramsauer et al. [2] is that the attention mechanism *is* a Hopfield network, with the symmetric Gram matrix $G = K^\top K/d$ governing the retrieval energy landscape.

Two tracks. We structure the transfer argument into two parts, which differ in exactness:

1. **Linear attention** (softmax $\to$ identity, used in Performer, cosine attention [3]): G is *exactly* the Hopfield weight matrix; WMT26 Theorem 1 applies verbatim.
2. **Softmax attention** (standard Transformer): the fixed-point disruption is exponentially small at high β; the weaker Gram-geometry disruption (Proposition 4.2) is exact.

Why this matters. Continual fine-tuning changes K, hence G. Even if individual stored keys remain approximately fixed (high-β regime), the *relational geometry* among all concept pairs is globally coupled to the Gram perturbation — the same coupling term that drives catastrophic forgetting in dense Hopfield networks.

*Preliminary draft. Comments welcome.

2 From Transformer Attention to the Gram Weight Matrix

2.1 Standard self-attention

Let $K \in \mathbb{R}^{p\times d}$ have rows $k_1, \ldots, k_p \in \mathbb{R}^d$ (keys), $Q \in \mathbb{R}^{T\times d}$ (queries), $V \in \mathbb{R}^{p\times d_v}$ (values). The attention output for query q_i is:

$$\mathrm{Attn}(q_i, K, V) = V^\top \mathrm{softmax}\big(\beta K q_i\big), \quad \beta = 1/\sqrt{d}.$$

The corresponding Modern Hopfield retrieval update [2] is:

$$x^{(t+1)} = K^\top \mathrm{softmax}\big(\beta K x^{(t)}\big).$$

2.2 The symmetric Gram weight matrix

Definition 2.1 (Key Gram matrix)**.** The *key Gram matrix* of attention head K is

$$G \;:=\; \tfrac{1}{d}\, K^\top K \;=\; \tfrac{1}{d}\sum_{\mu=1}^{p} k_\mu k_\mu^\top \;\in\; \mathbb{R}^{d\times d},$$

symmetric positive semi-definite with spectral decomposition $G = \sum_{j=1}^{d} \lambda_j e_j e_j^\top$.

G governs the quadratic part of the Hopfield energy: $E_{\mathrm{quad}}(x) = -\frac{1}{2}x^\top G x + \frac{1}{2}\|x\|^2$. The minimum directions of E_{quad} are the eigenvectors of G with the largest eigenvalues — exactly the Hopfield "stored patterns" in the linear-network limit.

2.3 Linear attention: exact equivalence

The *linear attention* head (obtained by replacing softmax with the identity map on the normalised dot-products, as in Performer [4] and cosine attention [3]) computes:

$$x^{(t+1)} = G x^{(t)} \;/\; \|G x^{(t)}\|.$$

The fixed points of this normalised iteration are exactly the eigenvectors of G: $Ge_j = \lambda_j e_j$. Adding a new key k_{new} — introducing a new concept a^* with target eigendirection e^* and retrieval weight $\lambda^* > \lambda_j$ for all existing j — means designing G' such that $G'e^* = \lambda^* e^*$.

This is *exactly* the setting of WMT26 Theorem 1 with $W \leftarrow G$, $\Delta W \leftarrow \Delta G$, $a^* \leftarrow e^*$, $\mu \leftarrow \lambda^*$. The impossibility theorem applies verbatim.

2.4 Softmax attention: fixed-point analysis

For finite β, the fixed points of $F(x) = K^\top \mathrm{softmax}(\beta K x)$ satisfy $x = F(x)$. At high β, these fixed points concentrate near the rows k_μ of K [2]. The Jacobian at a fixed point k_μ is:

$$J_\mu \;=\; \beta K^\top\big(\mathrm{diag}(p_\mu) - p_\mu p_\mu^\top\big)K,$$

where $p_\mu = \mathrm{softmax}(\beta K k_\mu)$. At high β, $p_\mu \to e_\mu$ (the standard basis vector for μ), giving $J_\mu \to 0$ — fixed points are *super-attractors* with convergence in one step.

Adding k_{new} changes the softmax denominators by $\exp(\beta k_{\mathrm{new}}^\top k_\mu)$, shifting existing fixed points by $O(\exp(-\beta\Delta_\mu))$ where $\Delta_\mu = \|k_{\mathrm{new}} - k_\mu\|^2/2$ is the energy gap. For well-separated concepts and realistic $\beta = 1/\sqrt{d}$, this shift is small but non-zero.

Remark 2.2 (Exponential suppression does not eliminate structural disruption)*.* The exponential suppression of fixed-point shift at high β does not mean that adding k_{new} is benign. Fine-tuning does not occur at a fixed frozen $\beta \to \infty$; gradient descent traverses intermediate configurations where the Jacobian structure of F depends on G globally. Moreover, even at high β, the *relational geometry* (see Section 4) is changed by a finite amount, independent of β.

3 Transfer Theorem: Linear Attention

Lemma 3.1 (Symmetry of the Gram perturbation)**.** *For any $\Delta K \in \mathbb{R}^{s\times d}$ (adding s new keys), $\Delta G = G' - G$ is symmetric.*

Proof. $\Delta G = \frac{1}{d}(\Delta K^\top K + K^\top \Delta K + \Delta K^\top \Delta K)$. Each summand is symmetric. □

Theorem 3.2 (Impossibility for Linear Attention)**.** *Let $K' = \left[K^\top \mid k_{\text{new}}\right]^\top$ (one new key appended), giving $\Delta G = k_{\text{new}} k_{\text{new}}^\top / d$. Let $k_{\text{new}} = \sum_j \alpha_j e_j$ with at least two indices $i \neq j$ satisfying $\alpha_i \neq 0$, $\alpha_j \neq 0$, and $\lambda_i \neq \lambda_j$. Then:*

(i) At least one eigenvalue of G' differs from the corresponding eigenvalue of G.

(ii) No choice of k_{new} can simultaneously make k_{new} a new fixed direction of G' (with $G'k_{\text{new}} = \lambda^* k_{\text{new}}$, $\lambda^* > \max_j \lambda_j$) while leaving all existing eigenvectors e_j of G unchanged as eigenvectors of G'.

Proof. By Lemma 3.1, ΔG is symmetric and non-zero. Claim (i): By the theory of symmetric rank-1 perturbations (Cauchy interlacing theorem), the eigenvalues of $G' = G + \Delta G$ strictly interlace with those of G unless $\Delta G = 0$. Since $\Delta G = k_{\text{new}} k_{\text{new}}^\top / d \neq 0$ has rank 1, exactly one eigenvalue of G' exceeds $\lambda_{\max}(G)$, and all others are interlaced; hence at least one eigenvalue shifts.

Claim (ii): Suppose for contradiction that all e_j remain eigenvectors of G': $G'e_j = \mu_j e_j$ for all j. Then $\Delta G\, e_j = (\mu_j - \lambda_j)e_j$, so $k_{\text{new}}(k_{\text{new}}^\top e_j) = d(\mu_j - \lambda_j)e_j$. For j with $\alpha_j = k_{\text{new}}^\top e_j \neq 0$, this gives $k_{\text{new}} = d(\mu_j - \lambda_j)/\alpha_j \cdot e_j$ — forcing k_{new} to be a scalar multiple of e_j. By assumption, this holds for at least two distinct j, i with $\lambda_j \neq \lambda_i$ — a contradiction, since k_{new} cannot simultaneously be proportional to two linearly independent vectors. Hence at least one e_j is no longer an eigenvector of G'.

By the WMT26 impossibility theorem applied to $(G, \Delta G) \leftarrow (W, \Delta W)$, the disruption of existing eigenvectors implies that all stored concepts whose eigenbasis decomposition has non-zero weight on the rotated mode are structurally at risk. □ □

Remark 3.3 (Generic vs. degenerate cases). The exceptional case — $k_{\text{new}} \propto e_j$ for a single j — places the new concept exactly on an existing eigendirection. This requires perfect alignment with a known eigenvector of G (a measure-zero condition for random k_{new}) and only shifts the single eigenvalue λ_j, which by Case 2 of WMT26 still disrupts all concepts with non-zero projection onto e_j. Both generic and degenerate cases lead to structural disruption.

4 Gram Geometry Disruption: Softmax Attention

For general softmax attention, we state a result that does not assume fixed-point equivalence.

Definition 4.1 (Key-space relational similarity)**.** The *relational similarity* between concepts k_μ, k_ν under weight matrix G is

$$\rho_{\mu\nu}(G) \;:=\; \frac{k_\mu^\top G\, k_\nu}{\|G^{1/2}k_\mu\|\,\|G^{1/2}k_\nu\|},$$

the G-weighted cosine similarity in key space.

This quantity measures the energy-geometry relation between the two concepts: how much they share spectral weight in G.

Proposition 4.2 (Relational disruption in key space)**.** *For any $\Delta K \neq 0$ with $\Delta G = G' - G$, every concept pair (k_μ, k_ν) satisfying $\langle k_\mu, k_{\text{new}} \rangle \neq 0$ or $\langle k_\nu, k_{\text{new}} \rangle \neq 0$ has*

$$k_\mu^\top G' k_\nu \;\neq\; k_\mu^\top G k_\nu.$$

In particular, $\rho_{\mu\nu}(G') \neq \rho_{\mu\nu}(G)$ for all such pairs. The first-order change is:

$$\Delta(k_\mu^\top G k_\nu) \;=\; \tfrac{1}{d}\langle k_\mu, k_{\text{new}} \rangle \langle k_{\text{new}}, k_\nu \rangle,$$

which couples all pairs through k_{new} and cannot be selectively suppressed for individual pairs.

Proof. Direct computation: $k_\mu^\top G' k_\nu = k_\mu^\top G k_\nu + k_\mu^\top \Delta G\, k_\nu = k_\mu^\top G k_\nu + \frac{1}{d}\langle k_\mu, k_{\text{new}} \rangle \langle k_{\text{new}}, k_\nu \rangle$. The correction is non-zero iff at least one of $\langle k_\mu, k_{\text{new}} \rangle$, $\langle k_\nu, k_{\text{new}} \rangle$ is non-zero. For generic patterns and k_{new} (full-support in key space), this holds for all pairs. □ □

Remark 4.3 (This result is β-independent)*.* Proposition 4.2 makes no assumption on β: it holds for all β, including the softmax attention standard setting $\beta = 1/\sqrt{d}$. The relational disruption is not suppressed at high temperature; it is a deterministic consequence of the Gram perturbation.

5 Implications

Continual fine-tuning. Each gradient step on W_K changes K, hence G, hence $\rho_{\mu\nu}(G)$ for all concept pairs with overlap on the perturbed modes. This is the key-space counterpart of the relational disruption in WMT26 Section 5.

Knowledge editing (ROME/MEMIT). These methods directly patch weight matrices [7]. For attention weight W_K: any edit changes $K = XW_K$, hence G. Proposition 4.2 applies directly: the relational geometry among all existing concepts is globally shifted.

LoRA. $\Delta W_K = AB$ ($A \in \mathbb{R}^{d_{\text{in}} \times r}$, $B \in \mathbb{R}^{r \times d}$) yields $\Delta K = X \Delta W_K$, so $\Delta G = \Delta K^\top K/d + K^\top \Delta K/d + \Delta K^\top \Delta K/d$. By Lemma 3.1, ΔG is symmetric and non-zero for any non-zero update. LoRA bounds the *rank* of ΔW_K but not the magnitude of ΔG, which depends on $\|\Delta K\|$. Small $\|\Delta K\|$ reduces the magnitude of $\frac{1}{d}\langle k_\mu, k_{\text{new}} \rangle \langle k_{\text{new}}, k_\nu \rangle$ but does not eliminate the coupling.

Retrieval-augmented generation. Storing new concepts in an external key–value store does not modify K, hence G is unchanged and Proposition 4.2 does not apply. RAG is thus the structural bypass for both the linear and softmax cases, for the same reason that sparse connectome topology is the structural bypass in WMT26.

6 Open Problems

1. **Quantitative disruption bound for softmax attention.** Proposition 4.2 gives the exact first-order change in $k_\mu^\top G k_\nu$. A bound on the resulting change in softmax attention output (retrieval fidelity) requires controlling the non-linearity of softmax($\beta \cdot$).

2. **Multi-head attention.** Each head has its own $G^{(h)}$; heads interact through the residual stream. A joint disruption bound across heads remains open.

3. **Feed-forward layers.** Geva et al. [6] show that MLP layers act as key–value memories. The same impossibility structure applies, but the relevant "weight matrix" is the first-layer projection; formalising its Gram structure is straightforward.

4. **Machine-verifiable proof.** The linear attention case (Theorem 3.2) has a proof structure that admits formalisation in Isabelle/HOL or Lean 4, following the proof of WMT26.

7 Conclusion

The structural impossibility result of WMT26 transfers to Transformer attention in two complementary forms. For linear attention, the transfer is exact: the Gram matrix $G = K^\top K/d$ is the Hopfield weight matrix, and WMT26 Theorem 1 applies verbatim. For softmax attention, fixed-point disruption is exponentially suppressed at high β, but the Gram geometry disruption (Proposition 4.2) is exact and β-independent: the relational similarity between all existing concept pairs is globally coupled to any key addition, and this coupling cannot be selectively eliminated.

The structural escape is the same in both cases: store new concepts outside the weight matrices.

References

[1] A. Bean. Catastrophic forgetting in dense associative memories is structurally inevitable: An impossibility theorem. Preprint, 2026.

[2] H. Ramsauer et al. Hopfield networks is all you need. In *ICLR*, 2021.

[3] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *ICML*, 2020.

[4] K. Choromanski et al. Rethinking attention with Performers. In *ICLR*, 2021.

[5] M. McCloskey and N. J. Cohen. Catastrophic interference in connectionist networks. *Psychology of Learning and Motivation*, 24:109–165, 1989.

[6] M. Geva, R. Schuster, J. Berant, and O. Levy. Transformer feed-forward layers are key-value memories. In *EMNLP*, 2021.

[7] K. Meng, D. Bau, A. Andonian, and Y. Belinkov. Locating and editing factual associations in GPT. In *NeurIPS*, 2022.

[8] E. J. Hu et al. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.