WHEN IS SOFTMAX ATTENTION A GRADIENT? EXACT CHARACTERIZATION AND CONDITIONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Energy and potential representations provide structural explanations for the fixed points and dynamics of tied or specially structured attention maps. Fixed-memory softmax attention can be a gradient even when its keys and effective values are untied, but this depends on the additive geometry of the keys. For $F(q) = U^\top \mathrm{softmax}(\beta K q)$, we give a finite necessary-and-sufficient criterion for a global potential by grouping equal key-pair sums. Generic memories force a common scaling and translation of the keys; binary Cartesian memories admit exactly independent coordinate scalings. We characterize constant-metric convex representations and derive two-sided finite-query bounds on the distance to the conservative value space from measured Jacobian asymmetry, with explicit conditioning and observation-error dependence. A near-Cartesian family shows why conditioning is essential: curl tends to zero while both the parameter distance and the best field approximation within the same-key conservative class remain bounded away from zero. Deterministic numerical checks validate the exact characterization and finite-query certificates, and illustrate the predicted loss of conditioning near pair-sum collisions. Finally, an analytic bifurcation construction shows that arbitrarily small incompatible perturbations can produce oscillation near a reference map's stability boundary. These results distinguish exact compatibility, reliable approximate certification, and dynamical stability.

1 INTRODUCTION

Fixed-memory softmax attention updates a state by weighting effective values according to query–key similarities (Vaswani et al., 2017). When such an update is iterated, a potential or energy representation can provide a structural explanation of its fixed points and dynamics, and under suitable convexity assumptions can further imply descent guarantees. This perspective builds on energy-based associative memory (Hopfield, 1982) and its dense and high-capacity extensions (Krotov and Hopfield, 2016; Demircigil et al., 2017; Krotov and Hopfield, 2021). Modern Hopfield networks give an important example in which the vectors defining the attention weights are aligned with those driving the state update (Ramsauer et al., 2021; Santos et al., 2025). General attention need not preserve this alignment: separate values and return projections introduce additional feedback. This raises a basic structural question: for a specified fixed-memory attention map, when does a global potential still exist, and what features of the memory determine the answer?

Existing work provides both energy-compatible constructions and differential conditions for conservativeness. Related approaches construct energy-based Transformers (Hoover et al., 2023), derive optimization interpretations of Transformer layers (Yang et al., 2022), give exact query-conditioned gradient-step representations of RoPE-softmax forward passes (Huang et al., 2026), or obtain a Wasserstein gradient flow through doubly stochastic attention (Sander et al., 2022). Jacobian symmetry is a classical criterion and has been used for untied auto-encoders (Im et al., 2016); more recently, Smart et al. (2025) derived the corresponding query-dependent constraint for softmax attention. The unresolved issue is not merely whether such a condition can be written, but whether it can be reduced to an explicit, query-independent characterization for a given memory. The softmax weights accessible from a fixed set of keys form a constrained query-dependent family rather than arbitrary probability vectors, so replacing the actual query family by a larger simplex can impose unnecessarily strong conditions. A useful characterization should therefore respect the attainable family while eliminating the dependence on the query itself.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We resolve this gap by deriving an exact finite characterization of energy compatibility for fixed-memory softmax attention, reducing the query-dependent symmetry condition to algebraic constraints determined by the additive geometry of the keys. The characterization shows that generic memories impose a rigid relation between keys and effective values, while structured memories can admit broader compatible families. We further extend the analysis beyond Euclidean gradients by characterizing when a constant positive-definite metric restores a convex potential representation, together with corresponding certificates, obstructions, and Euclidean repair guarantees.

Having characterized Euclidean repair, we then ask what approximate compatibility can actually certify. Finite query measurements can be used to estimate distance to the conservative class, but this inference becomes ill-conditioned near pair-sum collisions. In particular, vanishing Jacobian asymmetry need not imply a small same-memory Euclidean repair without a quantitative conditioning guarantee. This distinction also matters dynamically. Representation error preserves contraction when a positive stability margin is available, but structural proximity alone does not control behavior near criticality. A softmax Neimark–Sacker construction shows that arbitrarily small incompatible perturbations can produce oscillatory instability near the stability boundary.

The resulting picture is that memory geometry determines the admissible energy-compatible structure, while conditioning and stability margins determine how far that structure can be used quantitatively and dynamically. Deterministic numerical checks support these mechanisms, and the supplementary pretrained and retrieval studies delineate the scope of the fixed-memory theory.

2 Setup and Potential Structure

Let the rows of $K \in \mathbb{R}^{P\times d}$ and $V \in \mathbb{R}^{P\times m}$ be the keys and values. A linear return $C \in \mathbb{R}^{d\times m}$ gives

$$F(q) = CV^\top \operatorname{softmax}(\beta Kq) = U^\top \operatorname{softmax}(\beta Kq), \qquad U = VC^\top, \quad \beta > 0. \tag{1}$$

The rows $u_i^\top$ of U are the effective values in query coordinates. The memory and return are fixed during iteration. Put $s(q) = \operatorname{softmax}(\beta Kq)$, $J_s = \operatorname{diag}(s) - ss^\top$, and $Z(q) = \sum_i e^{\beta k_i^\top q}$. The standard softmax derivative and exponential-family covariance identity (Gao and Pavel, 2017; Wainwright and Jordan, 2008) give $DF(q) = \beta U^\top J_s K$.

We call F *conservative* if $F = \nabla\Phi$ for a scalar $\Phi \in C^2(\mathbb{R}^d)$. Discrete convergence of $q^+ = F(q)$ requires further conditions. If Φ is also convex, $E(q) = \|q\|^2/2 - \Phi(q)$ has the standard concave–convex descent bound (Yuille and Rangarajan, 2003). More generally, a *constant-metric convex representation* means $F = H^{-1}\nabla\Phi$ for $H = H^\top \succ 0$ and convex Φ. Our analysis concerns these two specified energy classes.

The return matters independently of notation. For invertible $S \in \mathbb{R}^{m\times m}$, $V' = VS$ and $C' = CS^{-\top}$ leave U and F unchanged. Imposing an identity return afresh after changing value coordinates generally changes both; Appendix A.1 gives a scalar example.

3 Exact Characterization on Actual Softmax Queries

Let $\mathcal{A}_K$ denote the set of distinct unordered pair sums, and let $\mathcal{P}_a$ collect the index pairs with sum a:

$$\mathcal{A}_K = \{k_i + k_j : 1 \le i < j \le P\}, \mathcal{P}_a = \{(i,j) : i < j,\ k_i + k_j = a\}, \qquad a \in \mathcal{A}_K. \tag{2}$$

Define

$$B_a(K,U) = \sum_{(i,j)\in\mathcal{P}_a} \left[(u_i - u_j)(k_i - k_j)^\top - (k_i - k_j)(u_i - u_j)^\top\right]. \tag{3}$$

These matrices are linear in U for fixed K.

Theorem 1 (Finite softmax criterion)**.** *Let $K, U \in \mathbb{R}^{P\times d}$, $P, d \ge 1$, and $\beta > 0$. For the map (1), the following statements are equivalent:*

1. *$DF(q)$ is symmetric on a nonempty open subset of $\mathbb{R}^d$.*
2. *$DF(q)$ is symmetric for every $q \in \mathbb{R}^d$.*

3. $B_a(K, U) = 0$ *for every* $a \in \mathcal{A}_K$.

4. *There is a* $\Phi \in C^2(\mathbb{R}^d)$ *with* $F = \nabla\Phi$.

When these conditions hold, one such potential is $\Phi(q) = \int_0^1 F(tq)^\top q \, dt$.

Proof. If $P = 1$, then $F = u_1$ and all four statements hold, with potential $u_1^\top q$. Assume henceforth that $P \geq 2$. The pairwise covariance identity gives

$$DF(q) = \beta \sum_{i<j} s_i(q) s_j(q)(u_i - u_j)(k_i - k_j)^\top. \tag{4}$$

Therefore

$$\frac{Z(q)^2}{\beta}\big(DF(q) - DF(q)^\top\big) = \sum_{a \in \mathcal{A}_K} e^{\beta a^\top q} B_a(K, U). \tag{5}$$

Choose v such that $(a - a')^\top v \neq 0$ for every distinct a, a'; finitely many proper hyperplanes are excluded. If the left side vanishes on an open set, choose q_0 there and restrict to $q = q_0 + tv$ on a small interval. The exponential rates $\beta a^\top v$ are distinct. Differentiating at $t = 0$ through order $|\mathcal{A}_K| - 1$ gives an invertible Vandermonde system, entry by entry. Thus $e^{\beta a^\top q_0} B_a = 0$ for every a. This proves $1 \Rightarrow 3$; (5) gives $3 \Rightarrow 2$, and $2 \Rightarrow 1$ is immediate.

For symmetric DF, differentiation under the stated integral gives

$$\partial_j \Phi(q) = \int_0^1 \left[F_j(tq) + t \sum_\ell q_\ell \partial_j F_\ell(tq) \right] dt = \int_0^1 \frac{d}{dt}\big[t F_j(tq)\big] \, dt = F_j(q). \tag{6}$$

Conversely, a C^2 potential has a symmetric Hessian. □

The theorem permits repeated or collinear keys and tests actual queries regardless of simplex attainability. It is independent of the positive temperature: the same (K, U) is conservative for every $\beta > 0$ if it is conservative for one. The equivalence concerns symmetry on open subsets of $\mathbb{R}^d$; lower-dimensional query manifolds need a manifold-specific criterion. The symmetric-Jacobian equivalence is classical (Im et al., 2016); the explicit grouped coefficient condition is the softmax-specific characterization.

3.1 Generic memories force a common scale

We use the following hypotheses only for the results that explicitly invoke them:

$$\begin{gathered} k_i \neq k_j \ (i \neq j), \qquad \dim \operatorname{aff}\{k_i\} \geq 2, \\ k_i + k_j \neq k_\ell + k_m \quad \text{for distinct unordered pairs } \{i, j\}, \{\ell, m\}. \end{gathered} \tag{7}$$

Corollary 2 (Generic rigidity)**.** *Let* K *satisfy* (7)*, let* $U \in \mathbb{R}^{P \times d}$*, and let* $\beta > 0$*. Then* F *is conservative if and only if*

$$U = \alpha K + \mathbf{1} b^\top, \qquad \alpha \in \mathbb{R}, \quad b \in \mathbb{R}^d. \tag{8}$$

It has a convex potential if and only if (8) *holds with* $\alpha \geq 0$*. The vector space of conservative effective values has dimension* $d + 1$.

Appendix A.2 proves the common-scale necessity by comparing noncollinear triangles; the converse follows from $DF = \alpha\beta K^\top J_s K$.

For $P \geq 3, d \geq 2$, (7) holds almost surely under an absolutely continuous joint distribution of all key entries. Pair collisions impose proper linear constraints; affine-span failures also have measure zero. This remains true when $P > d + 1$. This almost-sure statement is qualitative and does not provide a quantitative margin for nearly colliding pair sums. For arbitrary distinct, noncollinear keys, (8) characterizes symmetry for *every simplex distribution*; Appendix A.3 states that stronger-quantifier result.

3.2 CARTESIAN MEMORIES ADMIT EXACTLY COORDINATEWISE SCALES

The collision exception can be solved completely for a structured memory. Let $P = 2^d$ and index the rows of K by all sign vectors $s \in \{-1, 1\}^d$, with $k_s = s$. The softmax distribution factorizes over coordinates.

Theorem 3 (Binary Cartesian classification)**.** *Let K be the full binary Cartesian memory in $\mathbb{R}^d$, let $U \in \mathbb{R}^{2^d \times d}$, and let $\beta > 0$. Then F is conservative if and only if*

$$U = K\ \mathrm{diag}(a_1, \dots, a_d) + \mathbf{1}b^\top, \qquad a, b \in \mathbb{R}^d. \tag{9}$$

Its potential is convex if and only if $a_j \geq 0$ for every j. The conservative value space has dimension $2d$, and one potential is

$$\Phi(q) = \sum_{j=1}^{d} \frac{a_j}{\beta} \log\cosh(\beta q_j) + b^\top q. \tag{10}$$

The proof writes each field coordinate as a multi-affine polynomial in $\tanh(\beta q_j)$. Symmetry eliminates every cross-coordinate term (Appendix A.4).

Already in two dimensions, the opposite-corner pairs share sum zero. Their nonzero antisymmetric coefficients can cancel when $a_1 \neq a_2$. Such a map is conservative on every actual query and violates the all-simplex symmetry condition. The theorem classifies the entire structured family, including every parameter choice. An invertible linear image of the binary memory has the corresponding classification in Appendix A.5.

4 CONSTANT-METRIC CERTIFICATES AND EXPLICIT OBSTRUCTIONS

For a specified $H = H^\top \succ 0$, $HF = (UH)^\top s$. Theorem 1 therefore gives, for *every* memory,

$$HF \text{ is conservative} \iff B_a(K, UH) = 0 \quad (a \in \mathcal{A}_K). \tag{11}$$

For fixed K, U, these are linear equations in the entries of H. Conservativeness and convexity are separate properties. Under (7), both questions have a complete answer.

Let $P_c = I - \mathbf{1}\mathbf{1}^\top/P$ be the row-centering matrix, and put $A = P_c U$, $B = P_c K$. For distinct keys, F is constant if and only if $A = 0$: $F \equiv b$ implies $\sum_i e^{\beta k_i^\top q}(u_i - b) = 0$ for all q, so independence of distinct real exponentials gives $u_i = b$ for every i.

Theorem 4 (Generic convex-metric certificate)**.** *Let K satisfy (7), let $U \in \mathbb{R}^{P \times d}$ with $P_c U \neq 0$, and let $\beta > 0$. The following statements are equivalent:*

1. *$F = H^{-1}\nabla\Phi$ for some $H = H^\top \succ 0$ and convex $\Phi \in C^2(\mathbb{R}^d)$.*
2. *There is an $H = H^\top \succ 0$ with $AH = B$.*
3. *$\ker A^\top \subseteq \ker B^\top$, and $AB^\top$ is symmetric PSD with $\mathrm{rank}(AB^\top) = \mathrm{rank}\, A$.*

For any H in item 2 and $b = P^{-1}(UH - K)^\top \mathbf{1}$, the potential and energy

$$\Phi(q) = \beta^{-1} \log Z(q) + b^\top q, \qquad E_H(q) = \tfrac{1}{2} q^\top H q - \Phi(q) \tag{12}$$

satisfy

$$E_H(F(q)) - E_H(q) \leq -\tfrac{1}{2}\|F(q) - q\|_H^2 \qquad (q \in \mathbb{R}^d). \tag{13}$$

Corollary 2 reduces the question to $AH = \alpha B$ with $\alpha > 0$; rescaling H gives item 2. Positive matrix interpolation proves the remaining equivalence (Appendix A.6).

Constructing a metric or certifying infeasibility. Choose independent columns $X \in \mathbb{R}^{d \times r}$ of $A^\top$, and let $Y \in \mathbb{R}^{d \times r}$ be the matching columns of $B^\top$. First require every relation $A^\top w = 0$ also to satisfy $B^\top w = 0$. Then require $S = X^\top Y$ to be symmetric positive definite. Let N complete $W = [X\ N]$ to a basis, and put $R = N^\top Y$. For any $T = T^\top \succ 0$ on the complementary coordinates,

$$H = W^{-\top} \begin{pmatrix} S & R^\top \\ R & RS^{-1}R^\top + T \end{pmatrix} W^{-1} \tag{14}$$

is a certificate. When $r = d$, the empty complementary block is omitted and the normalized solution of $AH = B$ is unique; the metric in item 1 is then unique up to a positive scalar. An infeasible check returns, respectively, a vector w with $A^\top w = 0 \neq B^\top w$, a nonsymmetric entry of $X^\top Y$, or a nonzero vector c with $c^\top X^\top Y c \leq 0$. Appendix A.6 proves completeness and gives the witness construction.

The linear matrix problem and its symmetric/positive solutions are classical (Khatri and Mitra, 1976; Pinkus, 2004). Our use of it follows from the softmax necessity theorem. Likewise, the sufficient SPD outer return is already in Santos et al. (2025). The certificate identifies membership in this class. A constant map, treated separately from the theorem, has a linear potential in every positive-definite metric. Other energies, state-dependent metrics, and stable nonconservative dynamics are governed by different criteria.

A minimal Euclidean repair. For generic K, the closest convex-gradient values to any U in Frobenius norm have a closed form. With $K_c = P_c K$, $U_c = P_c U$, and row means $\bar{k}, \bar{u}$,

$$\alpha_* = \max\left\{0, \frac{\langle K_c, U_c\rangle_F}{\|K_c\|_F^2}\right\}, \quad b_* = \bar{u} - \alpha_* \bar{k}, \quad U_* = \alpha_* K + \mathbf{1} b_*^\top. \tag{15}$$

This is a one-dimensional constrained least-squares projection. For arbitrary keys, stacking the independent entries of (3) forms a linear operator $\mathcal{L}_K$; the nearest *conservative* values are the orthogonal projection onto $\ker \mathcal{L}_K$. Convexity must then be checked independently.

This exact repair raises a quantitative question when compatibility is only approximately observed: does a nearly symmetric attention Jacobian imply that only a small same-key Euclidean repair is required? We address this question next.

5 Approximate Euclidean compatibility and conditioning

Section 4 gives the exact same-key Euclidean repair by projection onto the conservative value space $\mathcal{C}_K = \ker \mathcal{L}_K$. We now ask whether such a repair is necessarily small when the attention Jacobian is only weakly asymmetric.

This question is nontrivial because the exact characterization in Section 3 depends on exact equalities among key-pair sums. A small perturbation of the keys can split a pair-sum collision and change $\mathcal{C}_K$, even though the resulting Jacobian remains nearly symmetric. Thus small Jacobian asymmetry need not automatically imply proximity to the new exact conservative class.

Put $u = \mathrm{vec}_{\mathrm{row}}(U)$ and let Π_K be the orthogonal projection onto $\mathcal{C}_K^\perp$. Then

$$\mathrm{dist}_F(U, \mathcal{C}_K) = \|\Pi_K u\|_2, \tag{16}$$

which is precisely the size of the smallest same-key conservative repair. To relate this repair distance to Jacobian asymmetry, use the upper-triangular curl entries:

$$\gamma_U(q) = \big((DF(q) - DF(q)^\top)_{ab}\big)_{a<b} = M_K(q)u, \qquad A_N = N^{-1/2}\begin{pmatrix} M_K(q_1) \\ \vdots \\ M_K(q_N) \end{pmatrix}. \tag{17}$$

Here $M_K(q)$ is explicit: with $T(q) = \beta J_{s(q)} K$, its row (a, b) has coefficient $T_{ib}(q)$ on u_{ia} and $-T_{ia}(q)$ on u_{ib}. The norm $\|A_N u\|_2$ is the RMS upper-triangular curl at the probes; the full skew-matrix Frobenius convention differs by $\sqrt{2}$.

Theorem 5 (Finite-query certification). *Let $K \in \mathbb{R}^{P\times d}$, $\beta > 0$, and $r = \mathrm{rank}\, \mathcal{L}_K > 0$. Let $\Omega \subseteq \mathbb{R}^d$ be a nonempty open set. Let $N \geq r$ and let $q_1, \ldots, q_N$ be independent random vectors taking values in Ω almost surely, each with a law absolutely continuous with respect to d-dimensional Lebesgue measure. Then $\ker A_N = \mathcal{C}_K$ almost surely. For any probes with this kernel,*

$$\mathrm{dist}_F(U, \mathcal{C}_K) \leq \frac{\|A_N u\|_2}{\sigma_r(A_N)}, \tag{18}$$

where σ_r is the smallest positive singular value. For arbitrary probes, let J be a matrix of compatible size and let $\rho = \|\Pi_K - J A_N\|_2 < 1$. Let $\eta \geq 0$ and $\|\widehat{y} - A_N u\|_2 \leq \eta$. Then $\ker A_N = \mathcal{C}_K$ and

$$\frac{(\|J\widehat{y}\|_2 - \|J\|_2\eta)_+}{1+\rho} \leq \mathrm{dist}_F(U, \mathcal{C}_K) \leq \frac{\|J\widehat{y}\|_2 + \|J\|_2\eta}{1-\rho}, \tag{19}$$

where $(x)_+ = \max\{x, 0\}$. *Moreover,* $\sigma_r(A_N) \geq (1-\rho)/\|J\|_2$.

Theorem 1 identifies the common kernel of the analytic query operators. This ensures a full-rank finite probe set; analyticity makes deficient random sets have measure zero. The bounds then follow from elementary inverse estimates (Appendix B.1). When $r = 0$, every value map is conservative and the distance is zero. The bound concerns conservativeness; convexity is a separate requirement. The sufficient probe count r is not a minimal count.

What is certified. Equation (19) admits a direct arithmetic check: propose J numerically and enclose $\Pi_K - JA_N$ using exact or interval arithmetic. An operator enclosure $\|A_N - \widehat{A}_N\|_2 \leq \xi$ also gives $\rho \leq \|\Pi_K - J\widehat{A}_N\|_2 + \|J\|_2\xi$. Thus an SVD rank threshold is unnecessary. A small upper bound certifies proximity, and a positive lower bound certifies separation; a large upper bound alone is inconclusive. The construction assumes known keys and permits errors in measured Jacobians. Uncertainty in the keys can change $\mathcal{C}_K$ itself and needs a different statement. If all values are available, projecting directly onto $\mathcal{C}_K$ already gives the exact distance; finite queries are useful when the available observations are Jacobians or their skew parts.

5.1 Near collisions prevent a uniform guarantee

The preceding bound shows that small Jacobian asymmetry implies a small repair only when the finite-query operator is well conditioned. Pair-sum collisions provide a natural place where this conditioning can fail: an exact collision can enlarge the conservative value space, while an arbitrarily small key perturbation removes that extra freedom.

Consider a perturbed binary square and anisotropic values:

$$K_\epsilon = \begin{pmatrix} -1+\epsilon & -1 \\ -1 & 1 \\ 1 & -1 \\ 1 & 1 \end{pmatrix}, \qquad U_\epsilon = K_\epsilon \operatorname{diag}(1,2), \qquad 0 < \epsilon < 1. \tag{20}$$

Write $F_\epsilon(q) = U_\epsilon^\top \operatorname{softmax}(\beta K_\epsilon q)$ and let μ_R be uniform probability measure on $[-R, R]^2$, $R > 0$.

Theorem 6 (Loss of uniform conditioning)**.** *Let* $\beta, R > 0$ *and let* (K_ϵ, U_ϵ) *be as in* (20)*. Then* $\sup_{q\in[-R,R]^2} \|DF_\epsilon(q) - DF_\epsilon(q)^\top\|_F = O(\epsilon)$*, while*

$$\operatorname{dist}_F(U_\epsilon, \mathcal{C}_{K_\epsilon})^2 = \frac{4(3\epsilon^2 - 8\epsilon + 16)}{3\epsilon^2 - 8\epsilon + 32} \longrightarrow 2. \tag{21}$$

Moreover, with $F_{K,V}(q) = V^\top \operatorname{softmax}(\beta K q)$*,*

$$\lim_{\epsilon\downarrow 0} \inf_{V\in\mathcal{C}_{K_\epsilon}} \|F_\epsilon - F_{K_\epsilon,V}\|_{L^2(\mu_R)} = \left[\tfrac{1}{2}\left(1 - \frac{\tanh(\beta R)}{\beta R}\right)\right]^{1/2} > 0. \tag{22}$$

For every $\epsilon > 0$ the pair sums are distinct, so the admissible repairs are $V = \alpha K_\epsilon + \mathbf{1}b^\top$. At $\epsilon = 0$ the space gains an extra dimension and already contains U_0. The proof computes the projection and the limiting field least-squares problem (Appendix B.2). Thus no bound with a key-independent constant can convert a small curl on a fixed domain into a small conservative repair with the same keys. This is also a field discrepancy, not merely a change in an unidentifiable parameterization.

The scope of this obstruction matters. The limit $F_0 = (\tanh(\beta q_1), 2\tanh(\beta q_2))$ is a convex gradient, and $F_\epsilon \to F_0$ in C^1 on compact sets. Every F_ϵ also has a constant-metric convex representation with $H = \operatorname{diag}(1, 1/2)$. The obstruction concerns Euclidean repairs with fixed keys; it does not exclude nearby arbitrary gradient fields or a change of metric.

From representation error to dynamical guarantees. For a convex-metric reference $F_0 = H^{-1}\nabla\Phi$ and a value perturbation $U = U_0 + R$, put $e = R^\top s$ and $d = F(q) - q$. The energy $E_H(q) = q^\top Hq/2 - \Phi(q)$ satisfies

$$E_H(F(q)) - E_H(q) \leq -\tfrac{1}{2}\|d\|_H^2 + \langle e(q), d\rangle_H. \tag{23}$$

Appendix A.7 gives RMS periodic-increment and spectral bounds. In particular, if the reference has contraction margin $m > 0$ on a closed convex invariant domain, a derivative defect $\delta < m$ guarantees convergence to a unique fixed point (Corollary 11). Without a positive margin, even an arbitrarily small defect can permit oscillation.

6 OSCILLATION NEAR THE COMPATIBILITY BOUNDARY

The preceding bounds provide dynamical guarantees when a positive stability margin is available, but do not exclude oscillation near criticality where that margin vanishes. Bifurcations of iterated softmax and the spectra of recurrent self-attention have been studied previously (Tiňo, 2009; Tomihari and Karakida, 2025). Here we give a fixed-memory softmax construction that undergoes an oscillatory bifurcation arbitrarily close to a convex-gradient reference. Let R_θ denote planar rotation and choose

$$K = \begin{pmatrix} 1 & 0 \\ -1 & 0 \\ 0 & 1 \\ 0 & -1 \end{pmatrix}, \qquad U = KR_\theta^\top, \tag{24}$$
$$F_{\beta,\theta}(x,y) = \frac{R_\theta}{\cosh(\beta x) + \cosh(\beta y)} \begin{pmatrix} \sinh(\beta x) \\ \sinh(\beta y) \end{pmatrix}.$$

At $\theta = 0$, this is the tied convex-gradient map. Varying θ rotates the effective values while leaving the keys fixed.

Proposition 7 (Oscillatory boundary). *Let $0 < \theta < \pi$ with $e^{ik\theta} \neq 1$ for $k = 1, 2, 3, 4$. For (24), the origin is fixed for every $\beta > 0$, with multipliers $(\beta/2)e^{\pm i\theta}$. At $\beta = 2$, it undergoes a supercritical Neimark–Sacker bifurcation. In the coordinate $w = x + iy$, the resonant radial cubic coefficient at the crossing is $-1/2$. Moreover, for every $\beta > 0$,*

$$\sup_{q \in \mathbb{R}^2} \|F_{\beta,\theta}(q) - F_{\beta,0}(q)\|_2 \leq 2\sin(\theta/2). \tag{25}$$

Proof outline. The derivative at the origin is $DF_{\beta,\theta}(0) = (\beta/2)R_\theta$, so the multipliers cross the unit circle transversely at $\beta = 2$. At the crossing, the Taylor expansion is

$$w^+ = e^{i\theta}\left(w - \frac{1}{2}w|w|^2 + \frac{1}{6}\overline{w}^3\right) + O(|w|^5). \tag{26}$$

All quadratic coefficients vanish. Removing the nonresonant cubic term leaves radial coefficient $-1/2$, giving the supercritical bifurcation (Kuznetsov, 2004). The bound (25) follows from $F_{\beta,\theta} = R_\theta F_{\beta,0}$, $\|F_{\beta,0}(q)\|_2 \leq 1$, and $\|R_\theta - I\|_2 = 2\sin(\theta/2)$. Appendix A.8 gives the normal-form calculation and a derivative bound. □

The nonreal multipliers rule out any constant-SPD potential representation: such a representation would make the Jacobian similar to a symmetric matrix. Nevertheless, taking admissible $\theta \downarrow 0$ makes the uniform value defect and, on bounded temperature intervals, the derivative defect arbitrarily small. Each fixed admissible θ retains the bifurcation, although the post-bifurcation parameter neighborhood may depend on θ. Thus oscillatory dynamics can occur arbitrarily close to a convex-gradient reference near its nonhyperbolic boundary. The resulting invariant curves may carry rational or irrational rotation, depending on the parameter.

7 EXACT CHECKS AND NUMERICAL ILLUSTRATIONS

The algebraic checks use exact rational keys and values. Table 1 verifies the geometry-dependent conservative spaces predicted in Section 3: generic memories recover the rigid common-scale family, whereas full binary Cartesian memories retain the additional coordinatewise degrees of freedom. Breaking the Cartesian pair-sum collision by an exact 10^{-12} perturbation collapses this extra freedom back to the generic dimension. The generic memory K_{g} and its metric-recovery example are specified in Appendix J.

7.1 VALIDATED FINITE-QUERY DIAGNOSTICS

We next test the finite-query guarantees from Section 5 near a pair-sum collision. We compare the generic memory, the exact square, and five near-square perturbations, all with $\beta = 1$ and $U = K\,\mathrm{diag}(1, 2)$. This value map is exactly conservative for the Cartesian square, but falls outside

Table 1: Memory geometry and the dimension of the conservative value space. The final column is the exact nullity $Pd - \operatorname{rank} \mathcal{L}_K$. The Cartesian rows use the full binary memories. The perturbation 10^{-12} is represented exactly.

Memory	P	d	Distinct pair sums	Conservative dimension
Generic K_{g}	5	2	10	3
Binary Cartesian	4	2	5	4
Binary Cartesian	8	3	19	6
Binary Cartesian	16	4	65	8
Cartesian with 10^{-12} perturbation	4	2	6	3

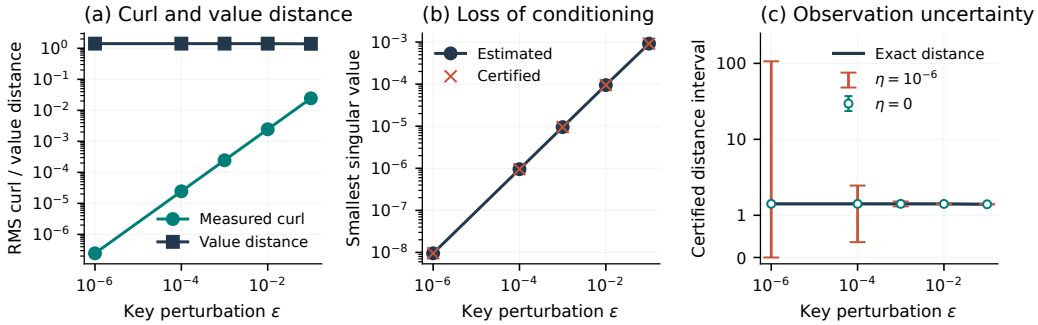

Figure 1: **Near-collision conditioning.** For (20), (a) measured curl shrinks while the distance to the same-key conservative class stays positive; (b) the smallest singular value deteriorates; (c) certified distance intervals widen as the allowed RMS observation error increases from zero to 10^{-6}. Panel (c) is linear below one and logarithmic above. All points are deterministic.

the generic conservative family once the exact collision is broken, making the near-square family a controlled test of the conditioning predicted by the theory.

All cases use the same 25 queries in $[-1, 1]^2$. Exact ranks and rational interval arithmetic verify the hypotheses of the distance certificate; Appendix B.3 gives the full protocol. Figure 1 and Table 2 summarize the results.

The three panels expose the mechanism behind the near-collision result. The vanishing measured curl in panel (a) does not reflect vanishing Euclidean distance: as predicted by the conditioning analysis, the finite-query operator simultaneously loses sensitivity in panel (b). Panel (c) shows the consequence for certification—the same small observation uncertainty produces increasingly weak distance bounds as the collision is approached.

With exact observations, the positive lower endpoint certifies separation despite small curl. At $\epsilon = 10^{-6}$, allowing $\eta = 10^{-6}$ gives $[0, 106.454588]$, which is inconclusive. Indeed the measured RMS curl is below η, so the same observations permit zero true curl within this error model. Exact values, when available, still determine the distance by projection. The best same-key conservative field discrepancy tends to 0.345258 on this domain, as predicted by Theorem 6. Meanwhile the metric $H = \operatorname{diag}(1, 1/2)$ certifies the original map without changing its output. Hence Euclidean projection and metric certification answer different practical questions.

An additional oscillatory illustration at $\theta = 0.6$ is provided in Appendix I: the exact multipliers cross the unit circle at $\beta = 2$, and the trajectory radii follow the local normal-form prediction.

8 Discussion and Conclusion

Memory geometry determines the exact conservative value space, and its conditioning determines whether small measured asymmetry is informative. The near-collision theorem shows that a same-

Table 2: Certified distance intervals for selected cases. Distances are rounded to six decimals; interval endpoints are rounded outward. The last column allows $\|\widehat{y} - A_N u\|_2 \le 10^{-6}$, an uncertainty allowance rather than a realized noise sample. All seven cases are released.

Memory	Distance	Interval, $\eta = 0$	Interval, $\eta = 10^{-6}$
Generic K_{g}	3.142411	[3.142411, 3.142412]	[3.142360, 3.142462]
Exact square	0	$[0, 10^{-25}]$	[0.000000, 0.000012]
Near square, $\epsilon = 10^{-2}$	1.412447	[1.412446, 1.412447]	[1.401906, 1.422987]
Near square, $\epsilon = 10^{-6}$	1.414213	[1.414213, 1.414214]	[0.000000, 106.454588]

key Euclidean repair may change the output field by a nonvanishing amount even as curl tends to zero. A constant metric can sometimes certify the original map instead. These statements concern representation classes; stable nonconservative dynamics and other Lyapunov functions remain possible.

The finite-query bounds require known keys and validated arithmetic errors. Uncertain keys, evolving memories, nonlinear feedback, and interacting heads require further analysis. Existing rank-collapse and clustering results concern evolving token representations (Dong et al., 2021; Geshkovski et al., 2023). Our normalization results and learned-LayerNorm counterexample (Appendices C–D) show why analyzing the complete feedback path matters. The imposed pretrained recurrence retains nonreal spectra under value tying when learned normalization and other projections remain (Appendix E); it is not an analysis of ordinary autoregressive generation.

The retrieval study provides a complementary task-level check: all arms converged, while mean long-horizon accuracy did not exceed the initial readout. Under this generator, an ideal semantic dot-product readout is Bayes optimal (Proposition 14), although training need not attain it. We therefore interpret this experiment as a scope check rather than evidence for recurrent accuracy gains.

Our results separate three questions that are often conflated in energy-based interpretations of attention: whether a fixed-memory softmax map admits a global potential, whether approximate compatibility reliably implies proximity to such a map, and whether that structure controls the resulting dynamics. The exact characterization identifies the first boundary, while the conditioning analysis determines when the second can be certified. The bifurcation construction shows that the third additionally depends on stability margins, especially near criticality. Together, these results clarify both the scope and the limitations of potential-based interpretations of fixed-memory softmax attention.

AI DISCLOSURE STATEMENT

Generative AI tools were used only to assist with language polishing and improving the readability of the manuscript. The authors reviewed all AI-assisted edits and take full responsibility for the paper's text, claims, and code.

REPRODUCIBILITY STATEMENT

The assumptions and complete proofs for the theoretical results are provided in the appendix. The appendix also specifies the numerical validation procedures, fixed experimental configurations, and evaluation protocols used for the reported results. The accompanying supplementary code provides reference implementations for the numerical checks and scripts for rerunning the empirical protocols and analyses.

REFERENCES

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. arXiv:1607.06450, 2016. https://arxiv.org/abs/1607.06450.

Rajendra Bhatia and Chandler Davis. A better bound on the variance. *The American Mathematical Monthly*, 107(4):353–357, 2000. https://doi.org/10.1080/00029890.2000.12005203.

Mete Demircigil, Judith Heusel, Matthias Löwe, Sven Upgang, and Franck Vermet. On a model of associative memory with huge storage capacity. *Journal of Statistical Physics*, 168(2):288–299, 2017. https://doi.org/10.1007/s10955-017-1806-y.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2793–2803, 2021. https://proceedings.mlr.press/v139/dong21a.html.

Bolin Gao and Lacra Pavel. On the properties of the softmax function with application in game theory and reinforcement learning. arXiv:1704.00805, 2017. https://arxiv.org/abs/1704.00805.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. The emergence of clusters in self-attention dynamics. *Advances in Neural Information Processing Systems*, 36:57026–57037, 2023. https://arxiv.org/abs/2305.05465.

Benjamin Hoover, Yuchen Liang, Bao Pham, Rameswar Panda, Hendrik Strobelt, Duen Horng Chau, Mohammed J. Zaki, and Dmitry Krotov. Energy Transformer. *Advances in Neural Information Processing Systems*, 36, 2023. https://arxiv.org/abs/2302.07253.

John J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982. https://doi.org/10.1073/pnas.79.8.2554.

Julie Huang, Maggie Chlon, and Leon Chlon. RoPE attention is an exact forward-pass gradient step with softmax intact. arXiv:2609.06685v1, 2026. https://arxiv.org/abs/2609.06685v1.

Daniel Jiwoong Im, Mohamed Ishmael Belghazi, and Roland Memisevic. Conservativeness of untied auto-encoders. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1):1694–1700, 2016. https://ojs.aaai.org/index.php/AAAI/article/view/10268.

Ali Kayyam, Anusha Madan Gopal, and M Anthony Lewis. Do transformers need three projections? Systematic study of QKV variants. arXiv:2606.04032v3, 2026. https://arxiv.org/abs/2606.04032v3.

C. G. Khatri and Sujit Kumar Mitra. Hermitian and nonnegative definite solutions of linear matrix equations. *SIAM Journal on Applied Mathematics*, 31(4):579–585, 1976. https://doi.org/10.1137/0131050.

Hyunjik Kim, George Papamakarios, and Andriy Mnih. The Lipschitz constant of self-attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5562–5571, 2021. https://proceedings.mlr.press/v139/kim21i.html.

Dmitry Krotov and John J. Hopfield. Dense associative memory for pattern recognition. *Advances in Neural Information Processing Systems*, 29, 2016. https://papers.nips.cc/paper_files/paper/2016/hash/eaae339c4d89fc102edd9dbdb6a28915-Abstract.html.

Dmitry Krotov and John Hopfield. Large associative memory problem in neurobiology and machine learning. *International Conference on Learning Representations*, 2021. https://arxiv.org/abs/2008.06996.

Yuri A. Kuznetsov. *Elements of Applied Bifurcation Theory*. Springer, third edition, 2004. https://doi.org/10.1007/978-1-4757-3978-7.

Asher Labovich. Stability and generalization in looped transformers. arXiv:2604.15259v2, 2026. https://arxiv.org/abs/2604.15259v2.

Allan Pinkus. Interpolation by matrices. *Electronic Journal of Linear Algebra*, 11:281–291, 2004. https://doi.org/10.13001/1081-3810.1138.

Ángel Poc-López and Miguel Aguilera. Dynamical mean-field theory of self-attention neural networks. arXiv:2406.07247, 2024. https://arxiv.org/abs/2406.07247.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. OpenAI technical report, 2019. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, Victor Greiff, David Kreil, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield networks is all you need. arXiv:2008.02217v3, 2021 revision. https://arxiv.org/abs/2008.02217v3.

Michael E. Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. Sinkformers: Transformers with doubly stochastic attention. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 3515–3530, 2022. https://proceedings.mlr.press/v151/sander22a.html.

Saul Santos, Vlad Niculae, Daniel McNamee, and André F. T. Martins. Hopfield–Fenchel–Young networks: A unified framework for associative memory retrieval. *Journal of Machine Learning Research*, 26(265):1–51, 2025. https://jmlr.org/papers/v26/24-1961.html.

Matthew Smart, Alberto Bietti, and Anirvan M. Sengupta. In-context denoising with one-layer transformers: Connections between attention and associative memory retrieval. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 55950–55971, 2025. https://proceedings.mlr.press/v267/smart25a.html.

Peter Tiňo. Bifurcation structure of equilibria of iterated softmax. *Chaos, Solitons & Fractals*, 41(4):1804–1816, 2009. https://doi.org/10.1016/j.chaos.2008.07.026.

Akiyoshi Tomihari and Ryo Karakida. Recurrent self-attention dynamics: An energy-agnostic perspective from Jacobians. *Advances in Neural Information Processing Systems*, 38, 2025. https://arxiv.org/abs/2505.19458.

Lloyd N. Trefethen. Pseudospectra of linear operators. *SIAM Review*, 39(3):383–406, 1997. https://doi.org/10.1137/S0036144595295284.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008. https://doi.org/10.1561/2200000001.

Yongyi Yang, Zengfeng Huang, and David P. Wipf. Transformers from an optimization perspective. *Advances in Neural Information Processing Systems*, 35, 2022. https://doi.org/10.52202/068431-2678.

A. L. Yuille and Anand Rangarajan. The concave-convex procedure. *Neural Computation*, 15(4):915–936, 2003. https://doi.org/10.1162/08997660360581958.

A Proof details and scope of the characterization

A.1 Covariance expansion and coefficient operator

The scalar identity-feedback example uses $K = (1, -1)^\top$, $V = K$, and $\beta = 2$, giving $F(q) = \tanh(2q)$. Changing V to $-V$ with a new identity return gives $-\tanh(2q)$. If $a > 0$ solves $a = \tanh(2a)$, the latter has the two-cycle $\{a, -a\}$. A compensated return leaves the original recurrence unchanged.

For any probability vector s, expanding $\sum_{i<j} s_i s_j (u_i - u_j)(k_i - k_j)^\top$ gives $\sum_i s_i u_i k_i^\top - (\sum_i s_i u_i)(\sum_j s_j k_j)^\top = U^\top J_s K$. The diagonal pairs contribute zero. This identity holds for arbitrary key multiplicities. For each $a \in \mathcal{A}_K$ and $1 \le r < c \le d$, put one row in $\mathcal{L}_K$ for $(B_a)_{rc}$. An individual pair adds $(k_{i,c} - k_{j,c})$ to the coefficient of $u_{i,r}$, its negative to that of $u_{j,r}$, $-(k_{i,r} - k_{j,r})$ to that of $u_{i,c}$, and its negative to that of $u_{j,c}$. Thus $\mathcal{L}_K \operatorname{vec}_{\mathrm{row}}(U) = 0$ is exactly Theorem 1. It has Pd columns and $|\mathcal{A}_K| d(d-1)/2$ rows, possibly dependent or zero.

Enumerating pair sums costs $O(P^2 d)$ arithmetic operations; constructing the sparse coefficient entries costs $O(P^2 d^2)$. Exact rank and nullspace computation can be expensive for large memories and large rational denominators. The implementation is a reference certificate for exact inputs. Floating-point scale-up requires certified arithmetic error bounds. The implementation keeps exact pair sums separate. For generic keys one may use the affine criterion directly once its hypotheses have been established. Orthogonal projection onto $\ker \mathcal{L}_K$ gives the closest conservative values in Frobenius norm. This elementary linear least-squares step addresses conservativeness; convexity must be checked independently.

A.2 Noncollinear anchor argument

For a nonzero vector v, $uv^\top = vu^\top$ implies $u \in \operatorname{span}(v)$: apply the equality to any vector orthogonal to v. Choose noncollinear keys k_a, k_b, k_c. Adding $u_a - u_b = \alpha_{ab}(k_a - k_b)$ and $u_b - u_c = \alpha_{bc}(k_b - k_c)$ and comparing with $u_a - u_c$ yields $\alpha_{ab} = \alpha_{bc} = \alpha_{ac}$ by linear independence. For any additional key k_i off the line through k_a, k_b, use triangle (i, a, b). If it lies on that line, distinctness and the third anchor make (i, a, c) noncollinear. In either case $u_i - u_a = \alpha(k_i - k_a)$. Hence $u_i = \alpha k_i + b$ for one b.

For interior s, $\operatorname{tr}(K^\top J_s K) = \frac{1}{2}\sum_{i,j} s_i s_j \|k_i - k_j\|^2 > 0$ when two keys differ. This proves the sign part of Corollary 2, including rank-deficient key spans. For $U = \alpha K + \mathbf{1} b^\top$, the potential is $\Phi(q) = \alpha \log Z(q)/\beta + b^\top q$. The generic measure-zero assertion assumes a density on all Pd key coordinates. Normalized or otherwise constrained sampling distributions require their own ambient measure, while the finite criterion applies directly to their realized keys.

A.3 The stronger all-simplex quantifier

Proposition 8. *Let $K, U \in \mathbb{R}^{P \times d}$ have distinct key rows of affine-span dimension at least two. Then $U^\top J_s K$ is symmetric for every $s \in \operatorname{int} \Delta_P$ if and only if $U = \alpha K + \mathbf{1} b^\top$ for some $\alpha \in \mathbb{R}, b \in \mathbb{R}^d$. It is PSD for every such s if and only if $\alpha \ge 0$ in that representation.*

Proof. Continuity extends symmetry to the boundary of the simplex. A distribution supported on a pair gives $t(1-t)(u_i - u_j)(k_i - k_j)^\top$, $0 < t < 1$. The anchor argument applies for every pair-sum geometry. The converse and sign follow from the covariance calculation above. □

Theorem 3 with unequal coordinate scales has a symmetric actual-query Jacobian and violates the all-simplex condition. This makes the distinction substantive. If the centered keys have rank $P - 1$, the map $q \mapsto \beta((k_1 - k_P)^\top q, \ldots, (k_{P-1} - k_P)^\top q)$ is surjective. These are all log odds against row P, so the simplex interior is attainable. This includes generic memories with $P \le d + 1$.

A.4 Proof of the Cartesian classification

Proof. Write $t_j = \tanh(\beta q_j)$. For each sign vertex σ,

$$s_\sigma(q) = 2^{-d} \prod_j (1 + \sigma_j t_j), \qquad f_\ell(t) = 2^{-d} \sum_{\sigma \in \{-1,1\}^d} u_{\sigma,\ell} \prod_j (1 + \sigma_j t_j).$$

Then $F_\ell(q) = f_\ell(t)$ with f_ℓ multi-affine. Its polynomial extension satisfies $f_\ell(\eta) = u_{\eta,\ell}$ at every sign vertex η. Symmetry requires, for $j \neq \ell$,

$$(1 - t_j^2)\, \partial_j f_\ell(t) = (1 - t_\ell^2)\, \partial_\ell f_j(t).$$

This polynomial identity holds on $(-1, 1)^d$. Its right side has degree at most one in t_j, whereas the coefficient of t_j^2 on the left is $-\partial_j f_\ell$, independent of t_j. Thus $\partial_j f_\ell = 0$. Consequently $f_\ell(t) = b_\ell + a_\ell t_\ell$, and evaluation at the vertices gives (9). Conversely, these values give $F_j(q) = b_j + a_j \tanh(\beta q_j)$ and (10). The Hessian is diagonal with entries $\beta a_j \operatorname{sech}^2(\beta q_j)$. □

A.5 Invertible images of Cartesian memories

Let $A \in \mathbb{R}^{d \times d}$ be invertible, let $k_0 \in \mathbb{R}^d$, and let $k_s = k_0 + As$ for every $s \in \{-1, 1\}^d$. The common key shift cancels from softmax. The field is conservative if and only if

$$u_s = b + A \operatorname{diag}(a) s, \qquad a, b \in \mathbb{R}^d.$$

Indeed set $x = A^\top q$. In these coordinates the gradient transforms as $\nabla_x \Phi(A^{-\top} x) = A^{-1} F(A^{-\top} x)$, whose keys are the sign vectors and whose values are $A^{-1} u_s$. Apply Theorem 3. The potential is $\Phi(q) = b^\top q + \sum_j a_j \log\cosh(\beta (A^\top q)_j)/\beta$. Its Hessian is $A \operatorname{diag}(\beta a_j \operatorname{sech}^2(\beta (A^\top q)_j)) A^\top$, so convexity is equivalent to all $a_j \geq 0$. This coordinate argument covers the covector transformation of a gradient. Arbitrary state conjugacies need their own Euclidean-gradient calculation. There are $3^d - 2^d$ distinct unordered pair sums: each coordinate is in $\{-2, 0, 2\}$, and at least one coordinate must be zero for two different sign vertices. A sum with $z \geq 1$ zero coordinates has 2^{z-1} unordered pair representations.

A.6 Complete metric construction

Proof. Apply Corollary 2 to HF. Convexity gives $UH = \alpha K + \mathbf{1} b_0^\top$ with $\alpha \geq 0$. Centering gives $AH = \alpha B$. Since $A \neq 0$ and H is invertible, $\alpha > 0$. Thus $\widehat{H} = H/\alpha$ is SPD and satisfies $A\widehat{H} = B$, giving item 2. Conversely, centering removes only a common row shift, so item 2 gives the displayed Φ.

If $AH = B$, then $AB^\top = AHA^\top$, proving necessity of item 3. For sufficiency, the kernel condition makes $L(A^\top w) = B^\top w$ a well-defined linear prescription on $\operatorname{range} A^\top$. The remaining conditions make $x^\top L y$ a symmetric positive-definite bilinear form on that subspace. The Schur completion below extends the prescription to an SPD operator on $\mathbb{R}^d$.

Finally, convexity gives

$$\Phi(F(q)) - \Phi(q) \geq F(q)^\top H (F(q) - q),$$

which yields (13). □

We prove the interpolation step for any $A, B \in \mathbb{R}^{P \times d}$. If $AH = B$ with $H = H^\top \succ 0$, then $HA^\top = B^\top$, $\ker A^\top \subseteq \ker B^\top$, and $AB^\top = AHA^\top$ is PSD of rank $\operatorname{rank} A$. Conversely let these conditions hold and let $r = \operatorname{rank} A > 0$. Choose independent columns X of $A^\top$ and the paired columns Y of $B^\top$. Write $A^\top = XD$ with $D \in \mathbb{R}^{r \times P}$ of full row rank. The kernel condition implies $B^\top = YD$: for each column its relation to the selected columns must also hold in $B^\top$. Hence $AB^\top = D^\top X^\top Y D$. Symmetry, PSD, and rank r of this matrix imply that $S = X^\top Y$ is symmetric positive definite, because D is surjective onto $\mathbb{R}^r$.

Let $W = [X\ N]$ be any invertible completion. The matrix in (14) is positive definite by its Schur complement $T \succ 0$; congruence gives $H \succ 0$. $W^\top H X = (S^\top, R^\top)^\top = W^\top Y$, so $HX = Y$ and $HA^\top = HXD = YD = B^\top$. This proves existence. Every SPD extension has the same first block

column, and its remaining block has a positive-definite Schur complement; thus (14) describes all extensions for this basis. For $r = 0$, an extension exists precisely when $A = B = 0$.

A kernel obstruction supplies a vector w with $A^\top w = 0 \neq B^\top w$, incompatible with any linear H. An asymmetric S contradicts symmetry of $X^\top HX$. For symmetric nonpositive-definite S, the reference code uses the first nonpositive leading principal Schur pivot: if $S_{1:k-1,1:k-1} \succ 0$, take $c_{1:k-1} = -S^{-1}_{1:k-1,1:k-1} S_{1:k-1,k}$, $c_k = 1$, and subsequent components zero. Then $c^\top Sc \leq 0$ while $Xc \neq 0$, contradicting positivity of H. These exact witnesses apply to the normalized convex-metric equation. Conservative metric representations with a concave potential follow a different criterion. For example, $U = -K_{\mathrm{g}}$ lies beyond the convex SPD-metric class, while its map is a global contraction when $\beta < 4/41$. Indeed the squared key diameter is 41, so the covariance range bound gives $\sup_q \|DF(q)\|_2 \leq 41\beta/4 < 1$. Certificate exclusion and dynamical instability are logically independent.

Descent consequences. Every iterate after the first lies in $\mathrm{conv}\{u_i\}$, a compact invariant set. For a certified convex metric, the continuous E_H is bounded below there, and (13) makes $\sum_t \|q_{t+1} - q_t\|_H^2$ finite. Every accumulation point is fixed by continuity of F, and every periodic orbit is constant. This argument establishes asymptotic regularity and fixed accumulation points. Convergence to a single limit requires an additional argument.

Nearest generic convex values. The squared Frobenius objective decomposes as

$$\|U - \alpha K - \mathbf{1}b^\top\|_F^2 = \|U_{\mathrm{c}} - \alpha K_{\mathrm{c}}\|_F^2 + P\|\bar{u} - \alpha\bar{k} - b\|^2.$$

Minimization over b eliminates the second term. The first is a strictly convex quadratic in α since $K_{\mathrm{c}} \neq 0$; constraining $\alpha \geq 0$ gives (15).

A.7 Dynamical bounds from representation errors

We now consider perturbations of a constant-metric convex-gradient map. The following standard convexity estimate gives precise consequences of an approximate representation. Write $\langle x, y\rangle_H = x^\top Hy$ and $\|x\|_H^2 = x^\top Hx$.

Proposition 9 (Descent and periodic-increment bounds)**.** *Let $H = H^\top \succ 0$, let $\Phi \in C^1(\mathbb{R}^d)$ be convex, and let $F_0 = H^{-1}\nabla\Phi$. Let $\widetilde{F} = F_0 + e$, $E_H(q) = q^\top Hq/2 - \Phi(q)$, and $d(q) = \widetilde{F}(q) - q$. Then, for every q,*

$$E_H(\widetilde{F}(q)) - E_H(q) \leq -\tfrac{1}{2}\|d(q)\|_H^2 + \langle e(q), d(q)\rangle_H. \tag{27}$$

Let $\Omega \subseteq \mathbb{R}^d$ satisfy $\widetilde{F}(\Omega) \subseteq \Omega$. If $\|e(q)\|_H \leq \eta\|d(q)\|_H$ on Ω for a constant $0 \leq \eta < 1/2$, then the energy decreases strictly at every nonfixed point in Ω.

If instead $\|e(q)\|_H \leq \varepsilon$ on Ω, every period-m orbit there satisfies

$$\left(\frac{1}{m}\sum_{t=0}^{m-1}\|q_{t+1} - q_t\|_H^2\right)^{1/2} \leq 2\varepsilon. \tag{28}$$

Proof. Convexity gives $\Phi(q+d) - \Phi(q) \geq \langle F_0(q), d\rangle_H = \langle q + d - e(q), d\rangle_H$, yielding (27). Cauchy–Schwarz gives the relative-defect conclusion. Summing (27) around a cycle yields $\frac{1}{2}\sum_t \|d_t\|_H^2 \leq \varepsilon\sum_t \|d_t\|_H$; another Cauchy–Schwarz inequality proves (28). □

The bound controls the root-mean-square *increment*. Orbit diameter, period, and prediction error need their own bounds. The constant 2 is sharp for the stated general class (the example below). Relative descent requires the defect to shrink with the increment near approximate fixed points.

For softmax values $U = U_0 + R$, the error is $e(q) = R^\top s(q)$. Let $r_i^\top$ be the rows of R. A global value-error bound is $\varepsilon = \max_i \|r_i\|_H$. The following derivative bound holds uniformly over all queries with the memory fixed. This differs from varying the full token input, as in global Lipschitz analyses of self-attention (Kim et al., 2021).

Proposition 10 (A spectral bound from value residuals). *Let* $F_0 = U_0^\top \mathrm{softmax}(\beta Kq) = H^{-1}\nabla\Phi(q)$, *where* $H = H^\top \succ 0$, $\beta > 0$, *and* $\Phi \in C^2(\mathbb{R}^d)$ *is convex. Let* $\widetilde{F} = (U_0 + R)^\top \mathrm{softmax}(\beta Kq)$, *and define*

$$D_R = \max_{i,j} \|r_i - r_j\|_H, \qquad D_K = \max_{i,j} \|k_i - k_j\|_{H^{-1}}, \qquad \delta = \beta D_R D_K/4.$$

For every q, $\|H^{1/2} De(q) H^{-1/2}\|_2 \le \delta$, *and every* $\lambda \in \mathrm{spec}(D\widetilde{F}(q))$ *has distance at most* δ *from* $\mathrm{spec}(DF_0(q)) \subseteq [0,\infty)$. *In particular,* $|\operatorname{Im}\lambda| \le \delta$ *and* $\operatorname{Re}\lambda \ge -\delta$.

The proof combines the covariance formula with a scalar range–variance bound (Bhatia and Davis, 2000) and the resolvent bound for a symmetric matrix (Trefethen, 1997); as detailed below. A small δ does not rule out a complex unit-circle crossing near $+1$. Section 6 gives an explicit oscillatory construction arbitrarily close to a convex-gradient reference.

For the sharpness of (28), set $H = 1$, $\Phi = 0$, $\Omega = [-\varepsilon, \varepsilon]$, and $e(q) = -q$. Then $\widetilde{F}(q) = -q$, and the two-cycle $\{\varepsilon, -\varepsilon\}$ has RMS increment 2ε. The relative defect ratio there is exactly $1/2$, where the energy remains constant along the two-cycle. This example establishes sharpness for the stated perturbation proposition.

For Proposition 10, $De = \beta R^\top J_s K$. For any Euclidean unit vectors x, y, the bilinear form $x^\top H^{1/2} De H^{-1/2} y$ is β times the covariance, under s, of the scalars $x^\top H^{1/2} r_i$ and $y^\top H^{-1/2} k_i$. A scalar supported on an interval of length D has variance at most $D^2/4$: centering at the midpoint bounds its second moment about that midpoint, and variance is the minimum second moment over centers. Cauchy–Schwarz for covariance now gives the bound $\beta D_R D_K/4$.

The matrix $A_0 = H^{1/2} DF_0 H^{-1/2} = H^{-1/2}\nabla^2\Phi H^{-1/2}$ is symmetric PSD. For a complex λ outside its spectrum, $\|(\lambda I - A_0)^{-1}\|_2 = 1/\operatorname{dist}(\lambda, \operatorname{spec} A_0)$. If that distance exceeds $\|H^{1/2} De H^{-1/2}\|_2$, the Neumann-series test makes $\lambda I - A_0 - H^{1/2} De H^{-1/2}$ invertible. Thus each perturbed eigenvalue lies within the stated distance. Similarity preserves both spectra. The conclusion concerns spectral location; eigenvalue pairing, eigenvectors, and task accuracy lie beyond this bound.

Corollary 11 (Contraction with a margin). *Let* $H = H^\top \succ 0$ *and let* $\Omega \subseteq \mathbb{R}^d$ *be nonempty, closed, and convex. Let* $F_0 = H^{-1}\nabla\Phi$ *with convex* $\Phi \in C^2$ *on a neighborhood of* Ω, *and let* $\widetilde{F} = F_0 + e$ *be continuously differentiable there with* $\widetilde{F}(\Omega) \subseteq \Omega$. *Let* $0 < m \le 1$ *and* $0 \le \delta < m$ *satisfy*

$$\sup_{q\in\Omega} \|H^{1/2} DF_0(q) H^{-1/2}\|_2 \le 1 - m, \qquad \sup_{q\in\Omega} \|H^{1/2} De(q) H^{-1/2}\|_2 \le \delta.$$

Then $\widetilde{F}$ *has a unique fixed point in* Ω, *and its iterates from every point of* Ω *converge to that point with contraction factor at most* $1 - m + \delta < 1$ *in* $\|\cdot\|_H$.

Proof. Integrating the derivative along the segment between two points of the convex set bounds their image distance by $(1 - m + \delta)$ times their distance. The set is complete in the H norm, so the contraction theorem applies. □

For the residuals in Proposition 10, $\delta = \beta D_R D_K/4$ is sufficient. This is a standard contraction consequence, included to separate the regime with a positive margin from the critical construction below.

A.8 Normal form and the small-defect boundary

Before rotation, expanding (24) at $\beta = 2$ gives

$$f_x = x - \tfrac{1}{3}x^3 - xy^2 + O(\|(x,y)\|^5), \qquad f_y = y - \tfrac{1}{3}y^3 - x^2 y + O(\|(x,y)\|^5).$$

Writing $w = x + iy$ makes the cubic $-w|w|^2/2 + \overline{w}^3/6$. Let $\lambda = e^{i\theta}$ and $w = v + h\overline{v}^3$, where $h = -(\lambda/6)/(\lambda - \overline{\lambda}^3)$. The transformed coefficient of $\overline{v}^3$ is $\lambda/6 + (\lambda - \overline{\lambda}^3)h = 0$; the resonant coefficient is $-\lambda/2$. The radial coefficient in this coordinate convention is therefore $-1/2$. The remaining low-order nonresonance conditions exclude the strong resonances in the Neimark–Sacker theorem (Kuznetsov, 2004). Writing $\mu = \beta/2 - 1$ gives $r^+ = r[1 + \mu - r^2/2 + O(\mu^2 + |\mu| r^2 + r^4)]$,

hence $r \sim \sqrt{\beta - 2}$ on the attracting invariant curve for sufficiently small positive $\beta - 2$ at fixed admissible θ.

The tied covariance has $\|DF_{\beta,0}(q)\|_2 \le \beta$ because all four keys have unit norm. Therefore

$$\sup_q \|DF_{\beta,\theta}(q) - DF_{\beta,0}(q)\|_2 \le 2\beta \sin(\theta/2).$$

Both this bound and (25) tend to zero for $\theta \downarrow 0$ uniformly on bounded temperature intervals. At the crossing, $DF_{2,\theta}(0) = R_\theta$ has imaginary parts $\pm \sin\theta$. SPD self-adjoint matrices are similar to real symmetric matrices and therefore have real spectra; the complex pair excludes that representation. This local spectral obstruction is independent of the generic pair-sum hypothesis, and these four keys have pair-sum collisions.

A.9 Additional certificate and consistency checks

Additional implementation checks complement the examples in Section 7. A three-dimensional embedding with rank-two centered keys checks the nonunique metric completion case. Separate tests return each type of obstruction: inconsistent prescribed images, nonsymmetry, and a nonpositive quadratic form. Duplicate keys, collinear keys, and constant maps test the general criterion beyond the generic classification.

A separate set of 100 fixed-seed numerical checks tests the covariance identity, the Cartesian formula, directional derivatives, and the energy and spectral bounds. The maximum covariance-identity discrepancy is 2.08×10^{-15}, and the maximum directional-derivative discrepancy is 2.48×10^{-9}. These checks use fixed inputs without model training or confirmation-data selection.

A.10 Additional related settings

Iterated-softmax bifurcations were studied by Tiňo (2009), and an asymmetric self-attention mean-field model by Poc-López and Aguilera (2024). Full looped-transformer stability and task behavior (Labovich, 2026) and the empirical consequences of projection sharing (Kayyam et al., 2026) address different questions from an exact fixed-memory certificate. These settings define the present scope and motivate targeted extensions for tying and stability.

B Proofs and validation of quantitative certification

B.1 Query operators and finite probes

For $a < b$, the (a, b) curl entry is $\beta \sum_i (u_{ia}(J_s K)_{ib} - u_{ib}(J_s K)_{ia})$. This proves the formula for $M_K(q)$ in (17). Theorem 1 implies

$$\bigcap_{q \in \Omega} \ker M_K(q) = \mathcal{C}_K$$

for every nonempty open Ω. The rows of these operators therefore span the r-dimensional space $\mathcal{C}_K^\perp$. Choose queries successively, each supplying at least one new row direction, until this span is reached. At most r queries suffice. Duplicate queries can pad the list to r points. Fix a nonzero $r \times r$ minor of the resulting stacked matrix. As a function of the r query vectors, this minor is real analytic on $\mathbb{R}^{dr}$ and is not identically zero. Its zero set has Lebesgue measure zero. Independence and absolute continuity of the individual laws give an absolutely continuous joint law of the first r queries, which therefore avoids this set almost surely; further queries cannot lower the rank. The laws need not be identical or uniform, and their densities need no positive lower bound. The assertion does not cover discrete sampling or probes restricted to a lower-dimensional set. It uses no quantitative separation assumption.

Write $u = u_\parallel + u_\perp$ with $u_\parallel \in \mathcal{C}_K$ and $u_\perp = \Pi_K u$. Every query annihilates $u_\parallel$. If $\ker A_N = \mathcal{C}_K$, the singular-value bound on $u_\perp$ gives (18). For the validated reconstruction, set $\zeta = \widehat{y} - A_N u$. The identity

$$u_\perp = J A_N u + (\Pi_K - J A_N) u_\perp$$

gives $(1-\rho)\|u_\perp\|_2 \le \|J\widehat{y}\|_2 + \|J\|_2\eta$. Conversely,

$$J\widehat{y} = u_\perp - (\Pi_K - JA_N)u_\perp + J\zeta,$$
$$\|J\widehat{y}\|_2 \le (1+\rho)\|u_\perp\|_2 + \|J\|_2\eta.$$

Together with nonnegativity of distance, this proves both sides of (19). The first identity also gives $(1-\rho)\|x\|_2 \le \|J\|_2\|A_N x\|_2$ for $x \in \mathcal{C}_K^\perp$, establishing the lower singular bound and kernel equality. For verified enclosures $t_- \le \|J\widehat{y}\|_2 \le t_+$, $\rho \le \bar{\rho} < 1$, and $\|J\|_2 \le \bar{g}$, a certified distance interval is

$$\left[\frac{(t_- - \bar{g}\eta)_+}{1+\bar{\rho}},\ \frac{t_+ + \bar{g}\eta}{1-\bar{\rho}}\right]. \tag{29}$$

The lower endpoint requires a lower enclosure t_-; an upper norm bound cannot replace it. The implementation bounds ρ and $\|J\|_2$ from above by Frobenius norms and encloses both endpoints outward.

An integrated version. Let μ be a probability measure with a density positive almost everywhere on a nonempty open set. The entries of $M_K(q)$ are bounded for fixed finite K, because s is a probability vector. Hence $G = \int M_K(q)^\top M_K(q)\, d\mu(q)$ exists. If $u^\top G u = 0$, then $M_K(q)u = 0$ almost everywhere on the open set. Continuity makes it zero throughout that set, and Theorem 1 implies $u \in \mathcal{C}_K$. Consequently $\ker G = \mathcal{C}_K$ and, for $r > 0$,

$$\operatorname{dist}_F(U, \mathcal{C}_K) \le \lambda_{\min}^+(G)^{-1/2}\|\gamma_U\|_{L^2(\mu)}.$$

This is a fixed-memory inverse estimate. Neither positivity of the constant nor almost-sure full probe rank supplies a useful uniform numerical bound.

B.2 The near-collision obstruction

For $0 < \epsilon < 1$, the six unordered pair sums of K_ϵ are $(-2+\epsilon, 0)$, $(\epsilon, -2)$, $(\epsilon, 0)$, $(0,0)$, $(0,2)$, and $(2,0)$. They are distinct, and the keys are noncollinear. Thus Corollary 2 applies. Let $v = 4 - 2\epsilon + 3\epsilon^2/4$ be the centered sum of squares of the first key coordinate. The second has sum of squares 4. Projection onto the common-scale family gives

$$\alpha_\epsilon = \frac{v+8}{v+4} = \frac{3\epsilon^2 - 8\epsilon + 48}{3\epsilon^2 - 8\epsilon + 32} > 0, \qquad b_\epsilon = \bar{u}_\epsilon - \alpha_\epsilon \bar{k}_\epsilon.$$

The squared residual is $v(1-\alpha_\epsilon)^2 + 4(2-\alpha_\epsilon)^2 = 4v/(v+4)$, which is (21). The Euclidean convex repair has the same distance because the minimizing scale is positive.

For $q = (x, y)$, direct use of (5) yields

$$(DF_\epsilon - DF_\epsilon^\top)_{12} = \frac{\beta}{Z_\epsilon(q)^2}\left[2\epsilon e^{\beta\epsilon x} + 2\epsilon e^{\beta(\epsilon-2)x} - 4e^{\beta\epsilon x} + 4\right]. \tag{30}$$

For $|x| \le R$ and $0 < \epsilon \le 1$, the bracket has magnitude at most $\epsilon e^{\beta R}(2 + 2e^{2\beta R} + 4\beta R)$. The arithmetic–geometric mean inequality gives $Z_\epsilon(q) \ge 4e^{\beta\epsilon x/4} \ge 4e^{-\beta R/4}$. Therefore the full curl Frobenius norm is at most $C_{\beta,R}\epsilon$, where the explicit finite constant

$$C_{\beta,R} = \frac{\sqrt{2}\beta}{16} e^{3\beta R/2}(2 + 2e^{2\beta R} + 4\beta R)$$

suffices. This constant is not asserted to be sharp.

For the field claim, put $T_\epsilon(q) = K_\epsilon^\top \operatorname{softmax}(\beta K_\epsilon q)$. Every admissible repaired field is $\alpha T_\epsilon + b$. Both T_ϵ and F_ϵ converge uniformly on the square to $T_0 = (\tanh(\beta x), \tanh(\beta y))$ and $F_0 = (\tanh(\beta x), 2\tanh(\beta y))$. The Gram matrix of the three vector-valued regressors T_0, e_1, e_2 is positive definite: a relation $\alpha T_0 + b = 0$ on an open square implies $\alpha = b = 0$. Thus the least-squares Gram matrices remain invertible for small ϵ and their minimizers and minimum values converge. Symmetry at the limit gives $b = 0$, $\alpha = 3/2$, and

$$\inf_{\alpha,b}\|F_0 - \alpha T_0 - b\|_{L^2(\mu_R)}^2 = \tfrac{1}{2}\frac{1}{2R}\int_{-R}^{R}\tanh^2(\beta x)\, dx = \tfrac{1}{2}\left(1 - \frac{\tanh(\beta R)}{\beta R}\right).$$

This proves (22). The limit of the generic repair families is smaller than the conservative family at the exact collision. Finally $U_\epsilon \operatorname{diag}(1, 1/2) = K_\epsilon$, proving the metric assertion.

Why this does not contradict approximation by arbitrary gradients. On a ball about zero define $\Psi(q) = \int_0^1 F(tq)^\top q\, dt$. Differentiation gives

$$\nabla\Psi(q) - F(q) = \int_0^1 t\big(DF(tq)^\top - DF(tq)\big)q\, dt.$$

Thus a uniform operator-norm curl bound c gives $\|\nabla\Psi(q) - F(q)\|_2 \le c\|q\|_2/2$. The same segment argument holds on a convex square. Such gradients need not be expressible as softmax maps with the original keys. Theorem 6 concerns that restricted class.

B.3 Deterministic diagnostic and arithmetic validation

The new diagnostic uses $\beta = 1$, the 25 points $\{-1, -1/2, 0, 1/2, 1\}^2$, and $U = K\,\mathrm{diag}(1, 2)$. It evaluates K_{g}, the exact binary square, and the five perturbations $\epsilon \in \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-6}\}$. This deterministic illustration evaluates the conditioning mechanism using a fixed query grid and exact arithmetic validation. The complete design is saved in `protocol/quantitative_diagnostic.json`. No training or model selection is involved. The fixed grid uses the deterministic certificate in Theorem 5; the almost-sure sampling statement is not invoked.

The exact rational operator $\mathcal{L}_K$ gives its rank and Π_K. An SVD proposes J using this exact rank, without dropping small directions. To enclose A_N, each rational exponential argument is reduced to $|z| \le 1/2$. For $z \ge 0$, the degree-32 Taylor sum has positive remainder at most $z^{33}/(33!\,(1 - z/34))$; interval squaring restores the scale and inversion handles negative arguments. Every interval operation rounds outward to a 96-bit dyadic grid using integer arithmetic. Stored rational bounds verify $\|\Pi_K - JA_N\|_F < 1$ and both sides of (29). For each component interval $[a, b]$ of $J\widehat{y}$, its minimum absolute value is zero if $0 \in [a, b]$ and $\min(|a|, |b|)$ otherwise. Summing the squared component minima and rounding the square root downward gives t_-; component maxima and upward rounding give t_+. The interval calculation includes error in A_N, but assumes its exact rational keys and probes. A separate $\eta = 10^{-6}$ bounds the Euclidean error of the RMS-normalized observation vector. It represents an allowed uncertainty, not a realized noise sample.

The seven cases all validate. The largest verified reconstruction residual ρ is below 6×10^{-9}. At $\epsilon = 10^{-6}$, the measured RMS curl is 2.44792×10^{-7} and the value distance is 1.414213386. The certified lower singular bound is 9.52014×10^{-9} after downward rounding. The exact-observation distance interval is $[1.414213376, 1.414213393]$ after outward rounding; its positive lower endpoint certifies separation. Allowing $\eta = 10^{-6}$ gives $[0, 106.454588]$, which certifies neither proximity nor separation. Since $\|\widehat{y}\|_2 < \eta$, the error model permits both the generating nonconservative values with zero observation error and any conservative values with observation error $\widehat{y}$. This ambiguity concerns query observations; exact known values can still be projected directly. A tied-value control with these same keys has exact distance zero but a noise-allowance upper endpoint above 105, confirming why a large upper endpoint alone cannot diagnose separation. A residual-only bound $\|A_N u\|/\sigma_r(A_N)$ is valid but looser for this particular direction; reconstruction uses the measured direction as well.

The best same-key conservative field fit is also evaluated using a fixed 41-point Gauss–Legendre rule in each coordinate. These field integrals are numerical diagnostics, not interval certificates; the positive limiting value is proved above. Symbolic checks independently verify the projection formula and curl numerator. A deliberately repeated single probe fails validation, and a rank-zero one-dimensional case returns distance zero. Direct Jacobian evaluation agrees with the query operator to 8.4×10^{-17} in the recorded implementation check. The supplied scripts generate the rational reconstruction matrices and bounds. A separate checker validates both endpoints of all fourteen generated witnesses (seven memories, two error allowances) without SVD or floating-point arithmetic. An independent interval implementation checks the same witnesses using pairwise covariance coefficients and a conservative-space basis. These computations supplement the written proofs.

C Supporting normalization and descent results

Consider the tied normalized map

$$\Psi(z) = az + cK^\top \mathrm{softmax}\big(\beta K\nabla G(z)\big), \qquad a, c \in \mathbb{R}, \tag{31}$$

where $G : \mathbb{R}^d \to \mathbb{R}$ is convex and twice continuously differentiable. Arbitrary key norms are allowed. Both the identity query map, $G(z) = \|z\|^2/2$, and affine-free LayerNorm fit this form.

Proposition 12 (Spectral restriction)**.** *Let $K \in \mathbb{R}^{P\times d}$, $\beta > 0$, $a, c \in \mathbb{R}$, and let $G \in C^2(\mathbb{R}^d)$ be convex. At every z, let $s = \mathrm{softmax}(\beta K \nabla G(z))$, $A = K^\top J_s K$, and $B = \nabla^2 G(z)$. Then*

$$\mathrm{spec}(D\Psi(z)) = \{a + c\beta\mu : \mu \in \mathrm{spec}(B^{1/2}AB^{1/2})\}, \qquad \mu \geq 0. \tag{32}$$

In particular, the spectrum is real for all a, c, and nonnegative when $a, c \geq 0$.

Proof. The derivative is $aI + c\beta AB$. Both A and B are positive semidefinite, and AB and $B^{1/2}AB^{1/2}$ have the same characteristic polynomial by the XY, YX determinant identity, with $X = AB^{1/2}$ and $Y = B^{1/2}$. This also handles singular B. □

A local Neimark–Sacker crossing requires complex multipliers, whereas the spectrum here is real. A flip requires a negative multiplier. If $c < 0$, even a tied scalar example $\Psi(z) = -\tanh(\beta z)$ crosses -1 at $\beta = 1$. If $a = c = 1$, every fixed point has all multipliers at least one. For example, $z^+ = z + \tanh(2z)$ diverges from any positive initial state. Real multipliers alone give local spectral information. Global convergence requires additional conditions, and a multiplier at $+1$ permits several codimension-one bifurcations.

Normalization scope. Let $P_0 = I - \mathbf{1}\mathbf{1}^\top/d$, $\varepsilon > 0$, and $\sigma(z) = (\|P_0 z\|^2/d + \varepsilon)^{1/2}$. The convex potential

$$G_{\mathrm{LN}}(z) = d\,\sigma(z), \qquad \nabla G_{\mathrm{LN}}(z) = P_0 z/\sigma(z) \tag{33}$$

gives affine-free LayerNorm (Ba et al., 2016). For $d \geq 2$ and $P_0 z \neq 0$, its Hessian has eigenvalues 0 in the constant direction, ε/σ^3 in the centered radial direction, and $1/\sigma$ on the remaining $d-2$ directions. At $P_0 z = 0$, all centered directions have eigenvalue $1/\sqrt{\varepsilon}$. A nonnegative scalar gain and a constant shift also give a convex gradient. Independent positive coordinate gains may break Hessian symmetry; Appendix D supplies an exact example at a tied fixed point. An analysis of standard learned LayerNorm must include its affine gains.

A global statement. Write $\Delta_P = \{p \geq 0 : \mathbf{1}^\top p = 1\}$ and $h(p) = \sum_i p_i \log p_i$, with $0 \log 0 = 0$. Define

$$T(p) = \mathrm{softmax}\left(\beta K \nabla G(K^\top p)\right), \qquad \mathcal{E}_\beta(p) = \beta^{-1} h(p) - G(K^\top p). \tag{34}$$

Theorem 13 (Descent under relaxed tied updates)**.** *Let $K \in \mathbb{R}^{P\times d}$, $\beta > 0$, $G \in C^1(\mathbb{R}^d)$ be convex, and $0 < \eta \leq 1$. For $p \in \Delta_P$, let $q = T(p)$ and $p^+ = (1-\eta)p + \eta q$. Then*

$$\mathcal{E}_\beta(p^+) - \mathcal{E}_\beta(p) \leq -\frac{\eta}{\beta} D_{\mathrm{KL}}(p\|q). \tag{35}$$

For any $p_0 \in \Delta_P$, the resulting sequence satisfies $\|p_{t+1} - p_t\|_1 \to 0$, and every accumulation point is fixed by T. The state sequence $z_t = K^\top p_t$ satisfies $z_{t+1} = (1-\eta)z_t + \eta K^\top \mathrm{softmax}(\beta K \nabla G(z_t))$ and has no nonconstant periodic orbit.

Proof. Convexity of h bounds $h(p^+) - h(p)$ by $\eta(h(q) - h(p))$. Convexity of G bounds its increment from below by $\eta \nabla G(K^\top p)^\top K^\top (q - p)$. Using $\log q_i = \beta(K \nabla G(K^\top p))_i - \log Z$ gives

$$\mathcal{E}_\beta(p^+) - \mathcal{E}_\beta(p) \leq \frac{\eta}{\beta}\{h(q) - h(p) - (q-p)^\top \log q\} = -\frac{\eta}{\beta} D_{\mathrm{KL}}(p\|q).$$

The continuous energy is bounded below on the compact simplex, so summing implies $D_{\mathrm{KL}}(p_t\|T(p_t)) \to 0$. Pinsker's inequality and continuity of T give asymptotic regularity and the accumulation-point claim. Since $z_{t+1} - z_t = K^\top(p_{t+1} - p_t) \to 0$, asymptotic regularity forces every periodic state orbit to be constant. □

For the bare update $\eta = 1$, every initial state enters $\mathrm{conv}\{k_i\}$ after one step. For relaxed updates the stated lift assumes $z_0 = K^\top p_0$. Unique-limit convergence requires an additional argument. The undamped residual with $a = c = 1$ belongs to another case. The energy construction is a specialization of familiar convex-concave reasoning; the explicit inequality and assumptions clarify its use here. The normalized Hopfield framework of Santos et al. (2025) already supplies related descent results. In particular, setting $q_t = \mathrm{LN}(z_t)$ turns our bare pre-normalized state update into a post-normalized recurrence in q_t. This relation becomes an invertible coordinate change only under additional conditions on LayerNorm.

D LEARNED LAYERNORM CAN INVALIDATE THE SPECTRAL PREMISE

For $u = P_0 z/\sigma$, direct differentiation yields $D\,\mathrm{LN}(z) = \sigma^{-1}(P_0 - uu^\top/d)$. The constant, centered radial, and remaining centered eigenspaces give the eigenvalues in Section C. With learned gain vector γ, the derivative becomes $\mathrm{diag}(\gamma)D\,\mathrm{LN}(z)$, which is generally asymmetric. The spectral argument above relies on two positive-semidefinite factors and therefore requires a different premise in this setting.

A tied example with four unit keys in three dimensions has an exact complex spectrum. Let $v_1 = (1,0,0)^\top$, $v_2 = (0,3/5,-4/5)^\top$, and

$$K = V = \begin{pmatrix} v_1^\top \\ -v_1^\top \\ v_2^\top \\ -v_2^\top \end{pmatrix}, \qquad \Gamma = \mathrm{diag}(67/25, 3, 1), \qquad L(z) = \Gamma\,\mathrm{LN}(z).$$

Choose any $\varepsilon > 0$ and $\beta = \sqrt{\varepsilon}$. At $z_* = 0$, $L(z_*) = 0$, uniform attention returns zero, and $DL(0) = \Gamma P_0/\sqrt{\varepsilon}$. Let $Q = (v_1, v_2)$, whose columns are orthonormal. The Jacobian is $D\Psi(0) = \frac{1}{2}QQ^\top\Gamma P_0$, has range in $\mathrm{span}(Q)$, and restricts there to

$$Q^\top D\Psi(0)Q = \begin{pmatrix} 67/75 & 67/750 \\ -1/6 & 67/75 \end{pmatrix}.$$

Consequently the full spectrum is

$$\mathrm{spec}(D\Psi(0)) = \left\{0,\ \frac{67}{75} \pm i\frac{\sqrt{335}}{150}\right\}.$$

All gains are strictly positive, all key and value rows have unit norm, and the fixed point is exact. Thus positive learned coordinate gains can produce a complex spectrum with literal tying and zero residual coefficient. The construction establishes possibility; prevalence must be measured across trained models.

E A PRETRAINED FEEDBACK EXPERIMENT

A defined intervention. We use the public GPT-2 small checkpoint (Radford et al., 2019) with frozen weights and memories. For one head, let $x_0 \in \mathbb{R}^D$ be the last token's residual input to its layer, $W_Q \in \mathbb{R}^{D\times d_h}$ and b_Q its query projection, and $O \in \mathbb{R}^{d_h\times D}$ its slice of the attention output projection. Let L be the layer's learned LayerNorm, including its affine parameters. Define

$$q_0 = W_Q^\top L(x_0) + b_Q, \qquad y_0 = V^\top \mathrm{softmax}(Kq_0/\sqrt{d_h}), \tag{36}$$

$$q(y) = W_Q^\top L\big(x_0 + O^\top(y - y_0)\big) + b_Q, \qquad F_g(y) = V^\top \mathrm{softmax}\left(gKq(y)/\sqrt{d_h}\right). \tag{37}$$

The state is the selected head's value output. Its change is reinjected into the anchored residual input; keys, values, and the anchor remain fixed. Thus y_0 is fixed at $g = 1$ by construction, while its stability is measured. The protocol studies this imposed recurrence. Ordinary forward passes, autoregressive generation, and complete-block iteration are separate settings.

Under $V' = VS$, $O' = S^{-1}O$, the state transformation $y' = S^\top y$ conjugates (37). The full derivative is

$$DF_g(y) = \frac{g}{\sqrt{d_h}} V^\top J_s K\, W_Q^\top DL(x(y))O^\top. \tag{38}$$

Replacing V by K retains the final feedback factors. Our tied control makes that replacement and recomputes y_0, retaining K, O, W_Q, L, x_0. A second, bare tied control is the covariance $K^\top J_s K/\sqrt{d_h}$ at q_0; it is a structural reference evaluated at the actual query. Fixed-point status for the bare map must be computed independently.

Fixed design and numerical validity. Before examining the new measurements, we fixed layers $0, 3, 6, 9$ and heads $0, 4, 8$, one development passage, and six confirmation contexts. Every selected head appears in every context. The implementation gates compare full-model logits against the

Table 3: Implementation checks. Errors are maxima. The reference comparison uses three separate prompts; Jacobian and basis-change checks cover every case in each stage.

Check	Development	Confirmation
Cases	12	72
Directional Jacobian error	6.25×10^{-10}	1.70×10^{-9}
Compensated value-basis error	8.28×10^{-16}	6.13×10^{-16}
Independent autograd Jacobian error	5.00×10^{-16}	8.33×10^{-16}
Reference logit error (three prompts)	1.80×10^{-7}	
Reference top-token agreement	3/3	

Table 4: All continuation outcomes, with development and confirmation kept separate. The validity-failure row records the post-run rejection in its own category.

Quantity	Development	Confirmation
Baseline spectral-radius range	$[0.0643, 0.6962]$	$[0.0371, 0.9725]$
Resolved crossings: negative real	2	11
Resolved crossings: complex pair	0	5
No crossing detected in range	10	53
Newton continuation failure	0	1
Refinement failure	0	1
Crossing validity failure	0	1
Resolved crossing gain range	$[2.4171, 3.9343]$	$[1.4216, 5.8237]$

reference GPT-2 implementation, test analytic Jacobians and compensated value-basis changes, and distinguish a slowly converging contraction from a periodic orbit. The confirmation sample contains 72 cases formed by repeating 12 heads across six contexts and defines a bounded sample. Table 3 reports validation; Appendix F gives the checkpoint, contexts, tolerances, and audit procedure.

Tying retains complex feedback spectra. Every case has nonreal eigenvalues in both the original and tied feedback maps, whereas the bare covariance control has a real spectrum in every case (Figure 2). The imaginary parts are well above numerical tolerance. All original feedback anchors are locally stable at $g = 1$; Table 4 gives their spectral-radius range. These observations coexist: a fixed point can attract while its perturbations rotate. Component-level attribution requires further intervention. The observations show that changing V to K alone still permits complex spectra in this recurrence.

Continuation and censoring. We increase g from one to ten on a fixed grid, continue fixed points with damped Newton iteration, and refine the first detected bracket in spectral radius across one. This tracks a numerical branch on the sampled grid; crossings between grid points remain possible. Solver exceptions are censored. A post-run audit also rejects one nominal confirmation crossing: its returned multiplier was 1.027501, inconsistent with a radius-one root despite a narrow gain bracket. This exposed branch switching during refinement. The raw candidate is retained, and a uniform validity check is applied to every candidate, as specified in the appendix. Every censored case retained its original grid.

Local consistency and finite-horizon outcomes. At each admitted crossing, we continue to $0.98g_c$ and $1.02g_c$ and compare the Jacobian prediction with a central finite-difference return matrix on the real critical eigenspace. Agreement validates the derivative implementation; these are two evaluations of the same local map. Predictive utility requires an independent test. We then iterate from a small critical-mode perturbation and from the baseline anchor. Table 5 reports all confirmation trajectories. Every admitted negative-real crossing is followed by a numerically resolved period-two orbit in both initializations. Four complex crossings are followed by unresolved nonstationarity through the computation cap. In the remaining complex case, both trajectories converge to a different fixed point. Replay places that endpoint at distance 1.296664 from the continued unstable point. Local instability and sustained nonstationarity therefore diverge for this initialization.

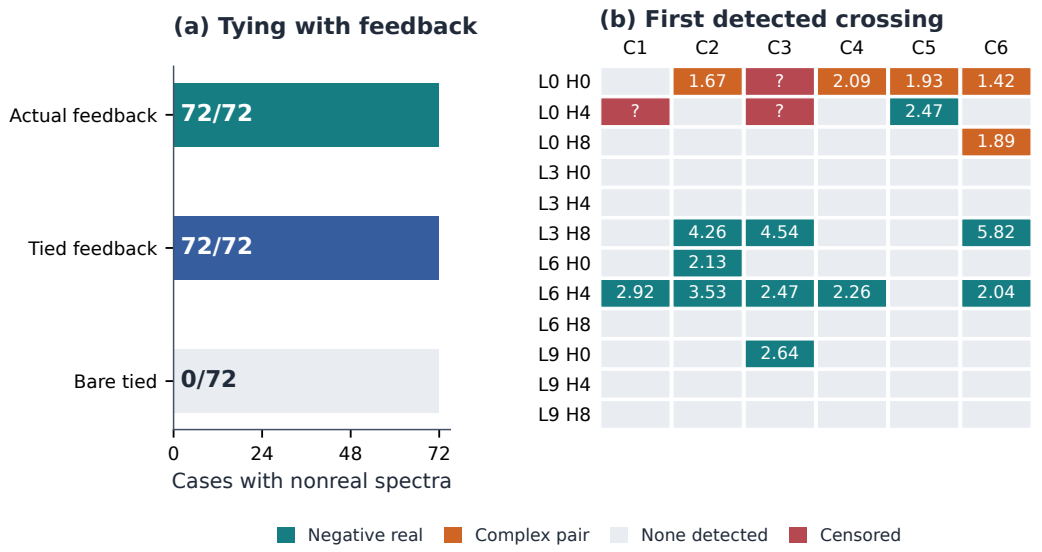

Figure 2: **The complete feedback changes the tying conclusion.** (a) Nonreal spectra occur in every confirmation case under both original and tied feedback; the bare tied covariance has a real spectrum in every case. (b) All 72 continuation outcomes. Colored cells give the gain at a numerically resolved crossing, gray cells mark grids with zero detected crossings through $g = 10$, and question marks retain the three censored cases. C1–C6 are the six fixed contexts. A complex crossing identifies a numerical candidate; nonlinear Neimark–Sacker certification requires normal-form or invariant-circle analysis.

Table 5: Confirmation follow-up at the 16 admitted crossings. Each gain uses two initializations. Unresolved trajectories reach 32,768 iterations and retain an unresolved label; quasiperiodicity or chaos classification requires a dedicated diagnostic.

Outcome	$0.98g_c$	$1.02g_c$
Local finite-difference sign agreement	16/16	16/16
Trajectories converging to a fixed point	32	2
Numerically resolved period-two trajectories	0	22
Unresolved nonstationary trajectories	0	8
Maximum local spectral-radius discrepancy	6.49×10^{-8}	

F Experimental protocol and validation

Checkpoint and contexts. The checkpoint is `openai-community/gpt2`, pinned to revision `607a30d783dfa663caf39e06633721c8d4cfcd7e`. The weight file SHA-256 is `248dfc3911869ec493c76e65bf2fcf7f615828b0254c12b473182f0f81d3a707`. Tokenization uses its pinned tokenizer. The original development passage has 132 tokens. Confirmation passages were written and frozen before pretrained measurements and concern a pendulum explanation, a station narrative, a numerical procedure, a transport argument, workshop arithmetic, and a prism explanation. They contain 91–104 tokens. Layer and head indices are zero-based. The sample is deliberately bounded; language-model benchmarking and random-head prevalence require different designs.

Derivative and extraction checks. For $x = x_0 + O^\top(y - y_0)$, $u = (x - \bar{x})/\sigma$, and learned gain γ,

$$Dq(y) = \frac{1}{\sigma} W_Q^\top \operatorname{diag}(\gamma) \left(I - \frac{\mathbf{1}\mathbf{1}^\top}{D} - \frac{uu^\top}{D} \right) O^\top.$$

Substitution into (38) yields the analytic derivative. Two fixed-seed directional finite differences use step $10^{-6}\max(1, \|y\|)$ and a relative error bound 10^{-5} at a perturbed anchor and $g = 1.7$. Value-basis tests use a random orthogonal factor and singular values evenly spaced from 0.7 to 1.3, with error bound 10^{-8}. A post-run independent PyTorch automatic-differentiation check covers every baseline Jacobian and requires maximum entrywise discrepancy at most 10^{-8}. The NumPy full forward pass matches the reference GPT-2 implementation in float64 on three separate prompts. A float32 embedding addition accounts for small extraction discrepancies. All saved numerical arrays are checked for finiteness.

Continuation. The grid is $\{1, 1.25, \ldots, 10\}$. Damped Newton iteration uses at most 70 steps, a relative residual target 10^{-10}, and at most 22 backtracking reductions. The first detected bracket with spectral radius crossing one is bisected to relative gain width 10^{-5}. Newton failure retains its own label. The leading multiplier is labeled complex if its imaginary part exceeds 10^{-6}; baseline complex-spectrum counts use threshold 10^{-8}. Newton iteration may switch numerical branches, and finite grid spacing permits crossings between sampled gains.

Crossing validation. The raw confirmation run reported 17 candidates. In case C3, layer 0, head 4, the refined gain was 1.0000038147, while the multiplier was 1.0275014485; the baseline spectral radius was 0.9724754036. This candidate violates the root condition and is censored as a branch/refinement validity failure. A uniform post-run check requires $|\rho - 1| \leq 10^{-4}$, remaining spectral radius below one, and normalized fixed-point residual at most 10^{-8} for every candidate. Sixteen pass. These added validity criteria have explicit post-run status. Continuation and sample selection retain the original execution. The remaining two censored cases record an initial Newton continuation failure and a refinement failure.

Local perturbation measurement. At $0.98g_c$ and $1.02g_c$, the fixed point is recomputed from the saved crossing state. For a real critical multiplier, we use its normalized real eigenvector. For a complex pair, we orthonormalize the real and imaginary parts into $Q \in \mathbb{R}^{d_h \times 2}$. Central differences with step $10^{-7}\max(1, \|y_*\|)$ estimate $DF_g(y_*)Q$. We compare the radius of the projected matrix $Q^\top DF_g(y_*)Q$ with the analytic spectral radius. This avoids interpreting a phase-dependent Euclidean norm ratio as a complex multiplier's modulus. Agreement is a local implementation check only.

Trajectory classification. At each paired gain, the two initializations are y_0 and $y_* + 10^{-3}\max(1, \|y_*\|)Q_{:,1}$. Iteration stops as converged after twelve consecutive steps with increment below $10^{-11}(1 + \|y\|)$. Otherwise, every 1,024 iterations the final 512 states are tested for periods 2 through 64, with discrepancy bound $10^{-8}(1 + \max\|y\|)$, nonzero spread above that threshold, and agreement across two successive windows. A detected period is a finite-tolerance numerical label; an exact mathematical orbit requires analytic proof. Tail spread, drift, residual, and candidate period are retained at every checkpoint. The cap is 32,768 iterations. All unclassified cases retain the unresolved label; quasiperiodicity requires a dedicated diagnostic.

A control map $z^+ = \tanh(0.9999z)$ is a global contraction, even though a short trajectory can retain a visible spread. The classifier keeps this control unresolved over its finite horizon. A fast contraction and the known $z^+ = -\tanh(2z)$ period-two map provide positive controls for the two resolved categories.

F.1 Validation of the trajectory classifier

A lag-two discrepancy alone is not sufficient to distinguish a period-two orbit from a slowly contracting trajectory: a monotone contraction can have small lag-two error while retaining nonnegligible drift across the full tail. For example, $x_{t+1} = 0.99999x_t$ with $x_0 = 0.02$ and $x_{t+1} = 0.99999999x_t$ with $x_0 = 0.1$ provide explicit contraction controls with this behavior.

To exclude such false positives, the final classifier fits one phase center for each candidate period over the entire trajectory tail. It requires nonconstant phase amplitude, every phase sample to lie within the prescribed tolerance of its phase center, the lag condition, and at least four observations per phase. For pretrained trajectories, the classification must also agree at two consecutive checkpoints.

We validate the criterion on synthetic controls. It rejects slow contraction and linear drift while correctly retaining exact periods 2, 3, 7, 31, and 64. Applying the same criterion to the 26 reported numerical period-two trajectories—22 confirmation and four development cases—retains all 26 classifications. One development trajectory and three confirmation trajectories first satisfy the criterion at step 3072 rather than 2048. All trajectories are evaluated with the same gains, contexts, heads, initializations, and 32,768-step cap used in the corresponding experiments. These classifications are finite-tolerance numerical observations.

As an independent numerical audit performed during development, a separate NumPy implementation reproduced all 2048 updates in every evaluation condition for the 15 trained checkpoints. It matched every recorded prediction, convergence label, and convergence iteration. The largest saved-tail discrepancy is 2.15×10^{-15}, the largest confidence discrepancy is 2.51×10^{-13}, and the largest compatible energy-inequality excess is 4.45×10^{-16}.

G TRAINED ASSOCIATIVE RETRIEVAL: CONVERGENCE AND ACCURACY

We compare three trained attention retrievers on the same episodic task to examine how potential compatibility relates to convergence and accuracy. The prospectively fixed protocol separates development from confirmation. An episode contains sixteen memory items with semantic and nuisance coordinates. Its query contains the target item's corrupted semantic vector and independently sampled nuisance. The network must identify the target memory index. Every confirmation episode has a fresh memory; Appendix H specifies the generator. With $P = d = 16$, this study does not test the overcomplete setting $P > d + 1$ emphasized by the actual-query characterization.

A compatible learned return and two controls. All models learn a shared query/key encoder and unit-normalize the keys and initial query once. They then apply the actual recurrence $q^+ = CV^\top \mathrm{softmax}(\beta K q)$, with fixed memory within an episode, and predict the target from $\beta K q_T$. The compatible model uses $V = K$ and $C = M = LL^\top \succ 0$, with a learned triangular factor. The tied control uses $V = K$ with a free return matrix C. The free control learns a separate value encoder and a free C. Both controls therefore test more than removal of a formal energy certificate: they also change the architecture's degrees of freedom.

For the compatible model, the metric transformation in Section 4 gives the energy and descent bound

$$E_M(q) = \tfrac{1}{2} q^\top M^{-1} q - \beta^{-1} \log \sum_i e^{\beta k_i^\top q}, \quad E_M(q^+) - E_M(q) \le -\tfrac{1}{2} \|q^+ - q\|^2_{M^{-1}}. \tag{39}$$

This is an architectural guarantee for the learned retriever once its episode memory is fixed. Retrieval accuracy remains an empirical quantity.

Comparison budget and decision rule. All three arms start from the same attention function and receive identical ordered training episodes within a seed. Each is trained at depth four with 512 AdamW steps and batch size 64; the final checkpoint is used. The registered trainable parameter counts are 648, 768, and 1280, respectively. We execute the same number of key, value, and query projections and attention updates in every arm; estimated forward arithmetic differs by less than 0.3%. This comparison equalizes data and optimizer steps and approximately matches forward arithmetic. Functional capacity and wall time differ across arms.

Confirmation uses five fixed training seeds, 256 unseen episodes per seed and noise condition, and 2048 updates for every episode. The primary gate requires accuracy noninferiority within two percentage points and a convergence-failure reduction exceeding five percentage points against *both* controls. The four one-sided paired-seed t bounds use Bonferroni adjustment. Higher-noise evaluation, prediction changes, and intermediate depths are secondary.

Convergence and answer accuracy diverge. All models converge on every tested episode, so the failure-reduction criterion is not met. The accuracy noninferiority criterion is satisfied against both controls. The compatible model retains the highest mean accuracy at the final horizon, and every paired seed favors it in the primary condition. The adjusted accuracy-difference lower bound includes zero against the tied control (Table 7). The evidence establishes a bounded endpoint advantage on this task; a general causal advantage requires matched parameterization.

Table 6: Held-out retrieval accuracy and convergence after training at depth four. Accuracy is the mean across five paired seeds, each with 256 episodes per condition. $T = 0$ is the query readout of the same trained model before recurrent updates. Failures count episodes whose final increments remain above the convergence threshold at $T = 2048$.

Architecture	$T = 0$ (%)	$T = 4$ (%)	$T = 2048$ (%)	Failures
Matched query noise, $\sigma = 1$				
Compatible metric	92.27	91.56	88.67	0/1280
Tied values, free return	91.48	91.09	86.33	0/1280
Free values and return	90.94	88.52	82.81	0/1280
Higher query noise, $\sigma = 1.5$				
Compatible metric	68.20	67.81	65.94	0/1280
Tied values, free return	68.36	67.81	64.14	0/1280
Free values and return	66.09	64.38	60.47	0/1280

Table 7: Predeclared primary comparisons, in percentage points. Positive failure reduction would favor compatibility. Both accuracy noninferiority requirements are satisfied, whereas neither failure-reduction requirement is met.

Control	Quantity	Mean difference	Lower bound	Required bound
Tied/free return	Accuracy	$+2.34$	-0.49	> -2.00
Free values/return	Accuracy	$+5.86$	$+3.34$	> -2.00
Tied/free return	Failure reduction	0.00	0.00	$> +5.00$
Free values/return	Failure reduction	0.00	0.00	$> +5.00$

More substantially, each architecture loses accuracy with extended unrolling, despite converging. The zero-iteration readout of each trained model also exceeds its final-horizon accuracy in both conditions (Table 6). Thus this task favors the simpler query readout over long recurrence. The compatible model's better retention is a bounded secondary result, and its different parameterization limits causal attribution. Figure 3 shows every seed trajectory alongside the endpoint average.

Post hoc decomposition of the endpoint difference. Under matched query noise, the final accuracy difference combines a gap already present at training depth four with a difference in subsequent accuracy change. Relative to the tied-value control, these components are 0.47 and 1.88 percentage points, respectively. Relative to the free-value control, they are 3.05 and 2.81 percentage points. This descriptive decomposition does not isolate a causal effect of compatibility.

H Prospective recurrent-retrieval study

Episode generation and models. Each episode draws sixteen independent unit semantic vectors in $\mathbb{R}^{16}$ and sixteen independent nuisance vectors distributed as $N(0, I_{16}/16)$. Concatenated vectors are multiplied by a fixed orthogonal matrix in $\mathbb{R}^{32\times32}$. The target index is uniform. Its query combines the target semantic vector plus $N(0, \sigma^2 I_{16}/16)$ with fresh nuisance. Training uses $\sigma = 1$; evaluation uses $\sigma = 1$ and 1.5 on matched base episodes and noise directions. A shared learned encoder $W \in \mathbb{R}^{32\times16}$ produces unit-normalized keys and initial query. In the free-value arm, a separate W_V produces unit-normalized values. Otherwise the value projection shares W with the keys; both projection calls execute in the common implementation. Normalization occurs once before the recurrence. Temperature is fixed at $\beta = 12$ throughout.

The compatible return is $M = LL^\top$, with 120 free strictly lower-triangular entries and 16 diagonal entries parameterized by $\mathrm{softplus}(a_i) + 0.001$. The other arms use a free 16×16 return matrix. Their parameter counts, including the shared encoder, are given in Section G. All encoders have the same initial query/key matrix within a seed; the separate value encoder is initialized to that matrix. The computed identity scale of the SPD parameterization is used for the free-return arms so the initial functions agree exactly in training precision.

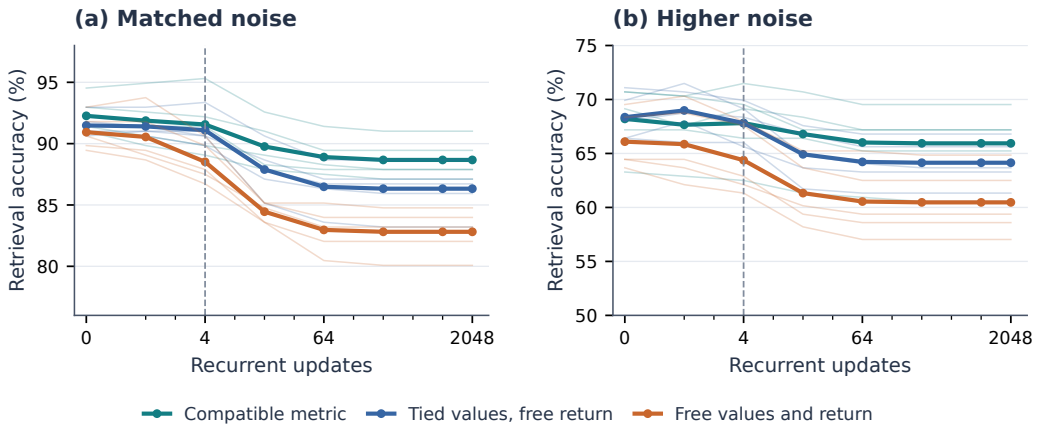

Figure 3: **Stable iteration can reduce retrieval accuracy.** Thin traces show each of the five paired seeds; thick traces show their means. The dashed line marks training depth four. Every trajectory meets the convergence criterion; accuracy decreases under longer unrolling. The zero-update readout uses the same learned encoder as the recurrent model. The horizontal axis uses a symmetric logarithmic scale to include zero.

The information available before recurrence.

Proposition 14 (Bayes readout for the frozen episode generator)**.** *Let $s_1, \dots, s_P \in \mathbb{R}^h$ be unit vectors, let Y be uniform on $\{1, \dots, P\}$ independently of them, and let $c = s_Y + \xi$, where $\xi \sim N(0, \sigma^2 I_h/h)$ and $\sigma > 0$. Let n contain observations conditionally independent of (Y, c) given the semantic memory. Then*

$$\Pr(Y = i \mid c, s_1, \dots, s_P, n) = \frac{\exp(h\, s_i^\top c/\sigma^2)}{\sum_j \exp(h\, s_j^\top c/\sigma^2)}.$$

The Bayes classifier for zero–one loss maximizes $s_i^\top c$.

Proof. The Gaussian likelihood is proportional to $\exp[-h\|c - s_i\|^2/(2\sigma^2)]$. The common $\|c\|^2$ and unit-norm terms cancel in the posterior, and n does not change it by conditional independence. □

The fixed orthogonal mixing in this generator can be inverted by an encoder that selects the semantic coordinates. Unit normalization of its query leaves the dot-product ordering unchanged almost surely. Thus the shared encoder class contains a depth-zero Bayes classifier. The experiment's fixed $\beta = 12$ need not equal h/σ^2; this changes calibration but not that ordering. This proposition concerns the ideal encoder and the specified generator. It does not prove that a trained encoder attains the Bayes rule or that recurrence cannot improve a particular learned representation. The proposition provides a task-level explanation for why recurrent updates need not improve the depth-zero readout under this generator.

Training and confirmation. Train with AdamW, learning rate 0.003, weight decay 10^{-4}, batch size 64, gradient norm clipping at one, and 512 updates. The objective consists solely of cross entropy from the final depth-four logits $\beta K q_4$, and the final fixed-step checkpoint is used. Development uses seed 1001. Confirmation uses 1101–1105, with disjoint training and evaluation generator seeds. The same ordered training batches and evaluation episodes are shared across variants within a seed. Every consumed training stream is hashed and independently regenerated in the audit. All fifteen scheduled confirmation trainings completed exactly 32,768 episodes each.

Training uses float32 and evaluation casts the frozen checkpoint to float64. CPU execution uses one thread. Estimated training forward multiply-accumulate counts per episode are 23,360 for the compatible model and 23,296 for each control. The difference is the small SPD matrix construction. The estimate covers forward matrix multiply-accumulates, while the recorded wall times include nonlinear operations, backward passes, and optimizer work. The setup approximately matches executed

forward arithmetic and exactly matches data and optimizer steps, while the arms retain different functional capacities.

Every confirmation episode receives all 2048 recurrent updates, even after apparent convergence. An episode is declared converged only if its final eight normalized increments $\|q_t - q_{t-1}\|/(1 + \|q_t\|)$ are below 10^{-7}. The recorded convergence time ends the first eight-step window in the final uninterrupted below-tolerance run. The mean of the five seed-level median times is 19.0, 18.6, and 20.2 iterations in the matched-noise condition, respectively. The compatible model ranks between the two controls in median convergence time. All final states, the last 256 states, predictions at eight depths, and case-level metrics are retained. A repeated-window period test and an unresolved label were available under the fixed protocol. Every confirmation episode resolved under the convergence criterion.

The fixed practical-benefit criterion. The primary outcome uses matched noise at the final horizon. For each control, define paired seed differences in accuracy (compatible minus control) and reductions in convergence failure (control minus compatible). The prespecified joint criterion requires the one-sided lower t bound across five seed pairs to exceed -0.02 for accuracy and $+0.05$ for failure reduction, for both controls. Four bounds use $\alpha = 0.05/4$, giving $t_{4,0.9875} = 3.495406$. A compatible mean accuracy of at least one half is also required. Table 7 reports every primary comparison. These small-sample bounds characterize this synthetic design. Cross-task robustness requires new data, and zero observed failure difference is compatible with multiple population rates. The failure-reduction requirement is not satisfied for either control, so the joint practical-benefit criterion is not met.

Independent numerical verification. The audit checks all fifteen scheduled checkpoints, all thirty evaluation arrays, and all training-stream hashes. It recomputes accuracy and final convergence classifications, replays every depth-four prediction from the saved checkpoint and regenerated inputs, and independently checks saved transitions with NumPy. The maximum transition discrepancy is 2.78×10^{-16}. The largest metric descent-inequality violation is 4.44×10^{-16}. All saved arrays are finite. The confirmation contains 15,728,640 executed episode updates. Every model ran the full schedule, and all scheduled models remain in the analysis.

I Numerical illustration of the oscillatory boundary

We examine the four-key construction at fixed $\theta = 0.6$ while varying the inverse temperature β across the critical value $\beta = 2$. Figure 4(a) shows the complex-conjugate multipliers crossing the unit circle, with modulus $\beta/2$, so that the origin is locally attracting for $\beta < 2$ and loses stability at $\beta = 2$. Just above the threshold, at $\beta = 2.04$, the trajectory approaches a small closed curve around the origin (Figure 4(b)), consistent with the attracting invariant curve established locally by Proposition 7.

Figure 4(c) quantifies the onset of this oscillation using the mean radius $|x + iy|$ over the trajectory tail. The subcritical runs have negligible tail radius, whereas a nonzero radius emerges above $\beta = 2$ and grows with β. Near the threshold, the numerical radii closely follow the leading normal-form scaling $r \sim \sqrt{\beta - 2}$ derived in Appendix A.8; deviations farther from the threshold are expected because this is a local leading-order approximation. Together, the three panels connect the complex unit-circle crossing to the emergence and growth of a small-amplitude oscillatory attractor in the specified softmax recurrence.

J Exact illustration memories

The generic two-dimensional memory is

$$K_{\mathrm{g}}^{\top} = \begin{pmatrix} 0 & 1 & 0 & 3 & 5 \\ 0 & 0 & 2 & 4 & -2 \end{pmatrix}. \tag{40}$$

The near-Cartesian example changes the first coordinate of the $(-1, -1)$ key by 10^{-12}. This breaks the sole pair-sum collision and reduces the conservative value dimension from four to three. Setting

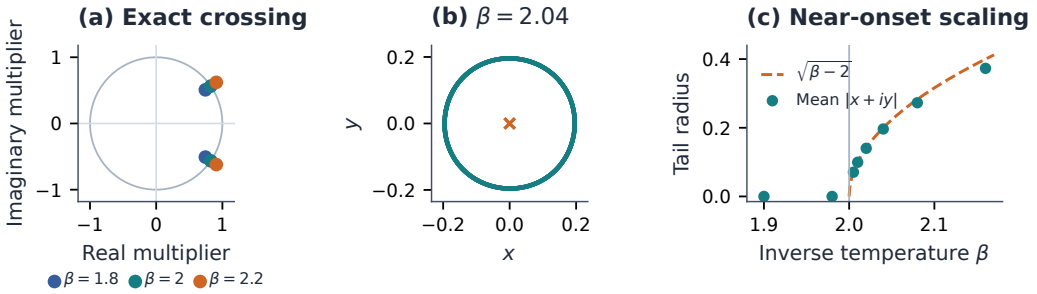

Figure 4: **Numerical illustration of the oscillatory boundary.** For (24) at $\theta = 0.6$, (a) the exact multipliers cross the unit circle at $\beta = 2$; (b) a trajectory at $\beta = 2.04$ approaches a small closed curve; and (c) the physical orbit radius is compared with the leading normal-form scaling $r \sim \sqrt{\beta - 2}$. The trajectory uses a fixed initialization and 20,000 iterations, with the final 2,000 states shown.

$U = K\,\mathrm{diag}(1, 2)$ for the perturbed memory then violates the Euclidean gradient criterion. Thus an arbitrarily small geometric perturbation can remove the extra Cartesian solutions; rounding away the perturbation would give a false certificate.

Metric recovery is illustrated by

$$U = K_{\mathrm{g}} M + \mathbf{1} b^{\top}, \qquad M = \begin{pmatrix} 2 & 1/3 \\ 1/3 & 1 \end{pmatrix}, \qquad b = \begin{pmatrix} 1 \\ -2 \end{pmatrix}. \tag{41}$$

These values fail the Euclidean gradient criterion. The metric construction instead returns

$$H = M^{-1} = \frac{1}{17} \begin{pmatrix} 9 & -3 \\ -3 & 18 \end{pmatrix}, \tag{42}$$

which gives a constant-metric convex representation. Additional certificate tests and numerical consistency checks appear in Appendix A.9.