MLP LAYERS REFINE AND EXPAND CLUSTERS IN CAUSAL ATTENTION DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern transformers exhibit rich dynamical phenomena such as clustering, alignment, and dimensional collapse. These effects are often explained through attention-only dynamics, where spectral properties of the value matrix drive tokens toward dominant eigendirections. However, in practical transformer blocks, attention is coupled with a nonlinear multilayer perceptron (MLP), whose dynamical effect remains poorly understood. In this work, we study a minimal model of causal attention interacting with a nonlinear MLP. We show that the MLP can fundamentally reshape attention dynamics: even when attention alone selects dominant eigendirections, the coupled system can develop new attracting states that are not aligned with any eigenvector of the value matrix. These non-eigendirection attractors have basins of attraction of positive Lebesgue measure and therefore produce genuinely new clustering behavior. We identify the underlying mechanism: the MLP nonlinearly deforms the attention vector field, creates new equilibria, stabilizes them locally, and, through causal propagation, synchronizes tokens around these new states. We construct explicit two-dimensional examples and extend the analysis to multi-token causal dynamics. Our results demonstrate that MLP blocks are not merely auxiliary nonlinearities but can qualitatively change the long-time behavior of transformer dynamics, generating clustering patterns beyond those predicted by attention alone.

1 INTRODUCTION

Transformers Vaswani et al. (2017) have become a leading architecture in modern sequence modeling, with broad impact across natural language processing Dehghani et al. (2019); Dai et al. (2019); Al-Rfou et al. (2019); Baevski & Auli (2019); Raffel et al. (2020), computer vision Dosovitskiy et al. (2021); Ramesh et al. (2021); Radford et al. (2021); Liu et al. (2021), general machine learning Tay et al. (2022); Khan et al. (2022); Lin et al. (2022), and reinforcement learning Janner et al. (2021); Chen et al. (2021). Much of their effectiveness comes from large-scale pretraining, which provides strong representations that transfer well to downstream tasks Radford et al. (2019b); Devlin et al. (2019a); Radford et al. (2018). By replacing recurrence and convolution with self-attention, Transformers directly model global interactions among tokens and capture long-range dependencies, producing rich contextual representations Lin et al. (2017); Vig & Belinkov (2019); Cho et al. (2014); Tenney et al. (2019); Nguyen et al. (2022b); Parikh et al. (2016).

Causal attention plays a key role in modeling sequential data (Devlin et al., 2019b; Radford et al., 2019b; Brown et al., 2020a; Touvron et al., 2023; Bai et al., 2023). A growing line of work has analyzed attention through dynamical, mean-field, and mechanistic perspectives (Geshkovski et al., 2023; 2025; Karagodin et al., 2024; Rigollet, 2025), revealing that attention alone can induce structured collective behavior, including alignment, dimensional collapse, synchronization, attention sinks, and clustering (Sander et al., 2022; Dong et al., 2021; Chen et al., 2026; Elhage et al., 2021; Olsson et al., 2022). In many simplified attention-only models, these phenomena are driven by the spectral properties of the value matrix: token representations tend to collapse toward dominant eigendirections or low-dimensional spectral subspaces (Geshkovski et al., 2023; Karagodin et al., 2024; Rigollet, 2025).

However, in practical Transformer architectures, attention does not act in isolation. Each attention layer is coupled with a multilayer perceptron (MLP) which applies a nonlinear transformation

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.

independently to each token (Vaswani et al., 2017; Devlin et al., 2019b; Brown et al., 2020a; Touvron et al., 2023). While attention introduces nonlocal interactions across tokens, the MLP acts locally, modifying each representation without explicit mixing across positions. Prior work has shown that feedforward layers play important roles in prediction construction, memory, representation editing, and model expressivity (Geva et al., 2021; 2022; Dai et al., 2022; Meng et al., 2022; Liu et al., 2022). Nevertheless, compared with attention, the dynamical role of the MLP, and in particular its interaction with attention, remains much less understood.

Our main contributions. In this work, we isolate the interaction between causal attention and the MLP in Transformer dynamics through a minimal analytically tractable model. We introduce a continuous-time dynamical system that couples single-head causal attention with a one-hidden-layer MLP, formulated as an ordinary differential equation on the unit sphere. This follows the continuous-depth viewpoint on deep networks and Transformers (Chen et al., 2018; Haber & Ruthotto, 2017; Lu et al., 2018; Dupont et al., 2019; Weinan et al., 2017; Ruthotto & Haber, 2020). Within this framework, we establish robust equilibrium-and-Jacobian criterion and establish robust non-spectral synchronized sinks in every finite dimension $d \geq 2$. Our results show that the feed-forward map, together with the tangent projections, can change the geometry of the attracting states and the locations at which tokens synchronize. Our main contributions are as follows:

- **Creation of robust non-spectral first-token attractors.** For a diagonal reference value matrix with a simple largest eigenvalue, the attention-only first-token dynamics converge to the dominant eigendirections for almost every initial state. We construct feed-forward matrices for which the composed flow admits additional hyperbolic sinks. These non-eigendirection attractors persist under small joint perturbations of V, W_1, W_2 and have open basins of positive surface measure (Theorem 3.1).
- **Local synchronization of finite causal prefixes.** Under the additional condition $\lambda_1+\lambda_2 > 0$ when $d \geq 3$, these first-token attractors give rise to stable synchronized equilibria of every finite causal prefix, for arbitrary fixed query and key matrices. At synchronization, derivatives of the attention weights cancel from the linearized value average. We establish stability of all diagonal tangent blocks and use the block triangular structure of the full causal Jacobian to prove local exponential convergence from open sets of initial configurations, allowing initially distinct tokens. One parameter neighborhood works for all Q, K and every finite prefix length, although the attracting neighborhoods may depend on them (Theorem 4.1).
- **Robust clustering at prescribed non-spectral directions.** We extend the construction to arbitrary real reference value matrices. Given any $s \in \mathbb{S}^{d-1}$ with $\mathsf{P}_s(V^0 s) \neq 0$, we construct feed-forward matrices for which the synchronized configurations at $\pm s$ are hyperbolic sinks for every finite causal prefix. These sinks persist as nearby non-spectral synchronized states under small joint perturbations of all three matrices, with a parameter neighborhood independent of Q, K and the prefix length. Such a construction is possible for every value matrix that is not a scalar multiple of the identity, without symmetry or spectral-sign assumptions (Theorem 4.2).

We further support our theoretical findings with experiments on several Transformer-based models such as GPT2-XL, Microsoft Phi-2, LLaMA-2-7B, and Mistral-7B.

Organization. Section 2 formulates the dynamical system governing the continuous-time Transformer dynamics. Section 3 shows that the MLP component can generate new clusters from the first-token dynamics, while Section 4 proves that these MLP-induced clusters persist across subsequent tokens. Section 6 validates the theoretical findings in practical Transformer models.

2 Background: Projected Causal-Attention–MLP Dynamics

We work in a reduced but analytically tractable setting that retains the key mechanisms of Transformer token dynamics. In contrast to analyses that isolate the attention block in the previous works, our model includes single-head causal attention, a normalization effect, a residual connection, and a nonlinear MLP block.

Let $X = [x_1, \dots, x_L]^\top \in \mathbb{R}^{L\times d}$ denote a sequence of L token embeddings, where d is the embedding dimension. We consider a single Transformer layer that maps an input sequence X to an output sequence $\mathrm{Trans}(X) \in \mathbb{R}^{L\times d}$ through

$$\mathrm{Trans}(X) = X + \mathrm{RMSNorm}\big(\mathrm{MLP}(\widehat{X})\big), \tag{1}$$

$$\widehat{X} = \mathrm{RMSNorm}\big(\mathrm{CauAttn}(X)\big), \tag{2}$$

Root mean square layer normalization (RMSNorm) Zhang & Sennrich (2019) controls the scale of token embeddings. In the continuous dynamical model below, this motivates a spherical formulation in which token directions evolve on $\mathbb{S}^{d-1}$. Spherical formulation is an idealized normalization device used in several dynamical analyses of attention (Geshkovski et al., 2025; Wu et al., 2024a; Burger et al., 2025; Karagodin et al., 2025).

Causal attention. Let $Q, K, V \in \mathbb{R}^{d\times d}$ denote the query, key, and value matrices. The causal attention output at position ℓ is

$$\mathrm{CauAttn}(X)_\ell = \sum_{j=1}^{\ell} a_{\ell j}(X)\, V x_j, \qquad \ell = 1, \dots, L, \tag{3}$$

where

$$a_{\ell j}(X) = \frac{\exp\big(\beta\langle Qx_\ell, Kx_j\rangle\big)}{\sum_{k=1}^{\ell}\exp\big(\beta\langle Qx_\ell, Kx_k\rangle\big)}, \qquad j \le \ell. \tag{4}$$

Here $\beta > 0$ is an inverse-temperature parameter; in standard scaled dot-product attention one takes $\beta = d^{-1/2}$. The causal mask restricts token ℓ to attend only to the prefix $x_1, \dots, x_\ell$.

MLP block. The MLP block is applied independently to each token. We take a two-layer MLP of the form

$$\mathrm{MLP}(\widehat{x}_\ell) = W_2\sigma(W_1\widehat{x}_\ell + b_1) + b_2, \tag{5}$$

where $W_1, W_2 \in \mathbb{R}^{d\times d}$, $b_1, b_2 \in \mathbb{R}^d$, and $\sigma : \mathbb{R} \to \mathbb{R}$ is applied componentwise. For simplicity, and to focus on the interaction between attention and the nonlinear activation, we set $b_1 = b_2 = 0$ throughout the theoretical analysis.

Activation. We assume that σ is a smooth activation function, or at least C^1 with bounded derivatives. A canonical example is the Gaussian Error Linear Unit (Hendrycks & Gimpel, 2016),

$$\mathrm{GELU}(x) = x\Phi(x),$$

where Φ is the cumulative distribution function of the standard normal distribution. Near the origin, GELU satisfies

$$\mathrm{GELU}(x) = \frac{1}{2}x + \frac{1}{\sqrt{2\pi}}x^2 + O(x^3), \qquad x \to 0.$$

Consequently, $\mathrm{GELU}'(0) = \frac{1}{2}$, so the activation behaves linearly to first order near the origin, while its higher-order terms provide a nonlinear deformation of the vector field.

Remark 2.1 (LLM activations)**.** As shown in Table 1 in Appendix B, besides GELU, the Sigmoid Linear Unit (SiLU) (Elfwing et al., 2018) is also widely used as an activation function in modern large language models, often serving as a smooth alternative to ReLU. Since SiLU and GELU have closely related mathematical profiles, we focus on GELU as the representative activation function in our theoretical analysis in order to keep the presentation focused and rigorous.

Continuous-time Transformer dynamics. Following the dynamical-systems viewpoint on deep residual networks and Transformers, we regard layer depth as a continuous-time variable. Let $X(t) = [x_1(t), \dots, x_L(t)]^\top$ denote the token representations at depth t. The normalization in equations (1) and (2) motivates constraining the token directions to the sphere $x_\ell(t) \in \mathbb{S}^{d-1}$. For $x \in \mathbb{S}^{d-1}$, define the tangent projection

$$\mathsf{P}_x(z) := z - \langle x, z\rangle x, \qquad z \in \mathbb{R}^d.$$

This is the orthogonal projection onto $T_x\mathbb{S}^{d-1} = \{h \in \mathbb{R}^d : \langle x, h\rangle = 0\}$.

The causal attention–MLP dynamics considered in this paper is obtained as the continuous-time limit of equations (1) and (2), is given by:

$$\partial_t x_\ell(t) = \mathsf{P}_{x_\ell(t)}\Big(W_2(t)\,\sigma\big(W_1(t)\widehat{x}_\ell(t)\big)\Big), \tag{6}$$

$$\widehat{x}_\ell(t) = \mathsf{P}_{x_\ell(t)}\left(\sum_{j=1}^{\ell} a_{\ell j}(t)\,V(t)x_j(t)\right), \tag{7}$$

for $\ell = 1, \dots, L$. The causal attention weights are

$$a_{\ell j}(t) = \frac{\exp\big(\beta s_{\ell j}(t)\big)}{\sum_{k=1}^{\ell}\exp\big(\beta s_{\ell k}(t)\big)}, \qquad j \le \ell, \tag{8}$$

where

$$s_{\ell j}(t) = x_\ell(t)^\top T(t)x_j(t), \qquad T(t) = Q(t)^\top K(t). \tag{9}$$

The projection $\mathsf{P}_{x_\ell(t)}$ ensures that $\partial_t x_\ell(t) \in T_{x_\ell(t)}\mathbb{S}^{d-1}$. Therefore, if $x_\ell(0) \in \mathbb{S}^{d-1}$, then the continuous dynamics preserve the spherical constraint:

$$x_\ell(t) \in \mathbb{S}^{d-1} \qquad \text{for all } t \ge 0.$$

This model captures the two mechanisms studied in the sequel: causal attention produces prefix-dependent nonlocal interactions through the value matrix V, while the MLP applies a local nonlinear transformation to the attention output. Our goal is to understand how this local nonlinearity can modify, and in some cases qualitatively overturn, the clustering behavior predicted by attention-only dynamics.

3 First-Token Analysis: MLPs Can Induce New Clusters

We first show that the effect of the MLP is already visible in the dynamics of the first token. Since this token can attend only to itself, its causal attention weight is one. This removes interactions with other tokens and isolates the geometric effect of the MLP on a single representation.

In the attention-only case, the first-token dynamics is

$$\partial_t x = \mathsf{P}_x(Vx), \qquad x \in \mathbb{S}^{d-1}. \tag{10}$$

For the diagonal reference value matrix considered below, the largest eigenvalue is simple, and almost every initial state converges to one of the two dominant eigendirections.

With the MLP, the first-token dynamics becomes

$$\partial_t x = \mathsf{P}_x\big(W_2\sigma(W_1\hat{x})\big), \qquad \hat{x} = \mathsf{P}_x(Vx). \tag{11}$$

The maps W_1, W_2 and the activation reshape the attention-induced tangent vector before the final projection. For the activations considered below, $\sigma(0) = 0$, so every eigendirection of V remains an equilibrium. Nevertheless, the MLP can change the stability of these equilibria and create additional attracting equilibria that are not eigendirections of V.

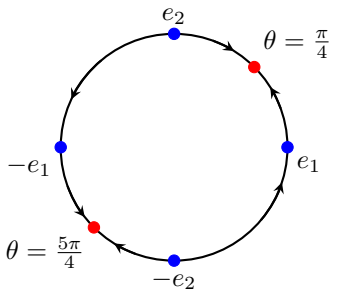


Figure 1: Reference phase portrait on the invariant circle $\mathbb{S}^{d-1} \cap \mathrm{span}\{e_1, e_2\}$. The blue points are eigendirections on this circle. The red points are the additional equilibria $\pm(e_1 + e_2)/\sqrt{2}$.

Throughout this section, vector norms are Euclidean. For a matrix $A = (a_{ij}) \in \mathbb{R}^{m\times n}$, define its norm by

$$\|A\| := \left(\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2\right)^{1/2} = \big(\mathrm{tr}(A^\top A)\big)^{1/2}.$$

Theorem 3.1 (MLPs create robust non-spectral attractors)**.** *Let $d \geq 2$ and let $\lambda_1, \ldots, \lambda_d \in \mathbb{R}$ satisfy*

$$\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_d.$$

Set

$$V^0 = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d), \qquad u = \frac{e_1 + e_2}{\sqrt{2}}.$$

(i) ***Attention only.*** *For*

$$\dot{x} = \mathsf{P}_x(V^0 x), \qquad x \in \mathbb{S}^{d-1},$$

every solution with $\langle x(0), e_1 \rangle \neq 0$ satisfies

$$x(t) \longrightarrow \operatorname{sign}\big(\langle x(0), e_1 \rangle\big) e_1.$$

Hence, attention-only dynamics converge to the dominant eigendirections $\pm e_1$ for almost every initial state with respect to surface measure on $\mathbb{S}^{d-1}$.

(ii) ***Attention with an MLP.*** *Let $\sigma = \mathrm{GELU}$, applied componentwise, and set*

$$W_1^0 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \oplus I_{d-2}, \qquad W_2^0 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \oplus I_{d-2},$$

with the last block row and column omitted when $d = 2$. Write

$$p = (V, W_1, W_2), \qquad p^0 = (V^0, W_1^0, W_2^0).$$

We use the product norm

$$\|p - p^0\|^2 := \|V - V^0\|^2 + \|W_1 - W_1^0\|^2 + \|W_2 - W_2^0\|^2.$$

There exist $\varepsilon > 0$ and C^1 equilibrium branches

$$x_\pm : \left\{ p \in \left(\mathbb{R}^{d \times d}\right)^3 : \|p - p^0\| < \varepsilon \right\} \longrightarrow \mathbb{S}^{d-1}$$

such that

$$x_\pm(p^0) = \pm u,$$

and, for every $\|p - p^0\| < \varepsilon$, the points $x_\pm(p)$ are hyperbolic sinks of

$$\dot{x} = \mathsf{P}_x W_2 \sigma(W_1 \mathsf{P}_x(Vx)), \qquad x \in \mathbb{S}^{d-1}.$$

Moreover,

$$\left\| \mathsf{P}_{x_\pm(p)}(V x_\pm(p)) \right\| \geq \frac{\lambda_1 - \lambda_2}{4} > 0,$$

so these attracting states are not eigendirections of V.

Each sink has an open basin of positive surface measure, and convergence is locally exponential. The perturbed matrices need not preserve symmetry or the reference block structure.

4 Subsequent-Token Analysis: Persistence of New MLP-Induced Clusters

Theorem 4.1. *Under the assumptions and notation of Theorem 3.1(ii), assume additionally that $\lambda_1 + \lambda_2 > 0$ when $d \geq 3$. There exists $\varepsilon > 0$ such that, for every p with $\|p - p^0\| < \varepsilon$, every $Q, K \in \mathbb{R}^{d \times d}$, and every finite $L \geq 1$, both synchronized configurations*

$$\big(x_\pm(p), \ldots, x_\pm(p)\big) \in (\mathbb{S}^{d-1})^L$$

are hyperbolic sinks of equations (6) and (7). Each has an open neighborhood $U_{\pm,L} \subset (\mathbb{S}^{d-1})^L$ of positive product surface measure such that every initial configuration in $U_{\pm,L}$ satisfies

$$x_j(t) \longrightarrow x_\pm(p) \qquad \text{as } t \to \infty, \qquad 1 \leq j \leq L.$$

Convergence is locally exponential, and the common limiting state remains non-spectral.

The parameter radius ε is independent of Q, K, L; the attracting neighborhoods may depend on the parameters and on L.

Theorem 4.2 (Clustering at prescribed non-spectral directions)**.** *Consider the causal attention–MLP dynamics equations* (6) *and* (7) *with componentwise GELU and $d \geq 2$. Let $V^0 \in \mathbb{R}^{d\times d}$ and $s \in \mathbb{S}^{d-1}$ satisfy $H_s := \mathsf{P}_s(V^0 s) \neq 0$. Choose $W_1^0 \in \mathbb{R}^{d\times d}$ with $z := W_1^0 H_s \neq 0$, and set*

$$W_2^0 := \frac{sz^\top}{\|z\|^2}. \tag{12}$$

Write $p = (V, W_1, W_2)$ and $p^0 = (V^0, W_1^0, W_2^0)$, with the parameter norm of Theorem 3.1.

There exist $\varepsilon > 0$ and C^1 equilibrium branches $x_\pm(p) \in \mathbb{S}^{d-1}$, defined for $\|p - p^0\| < \varepsilon$, such that

$$x_\pm(p^0) = \pm s, \qquad \left\|\mathsf{P}_{x_\pm(p)}(V x_\pm(p))\right\| \geq \frac{1}{2}\|H_s\| > 0. \tag{13}$$

For every such p, arbitrary fixed Q, K, and every finite $L \geq 1$, both synchronized configurations

$$\big(x_\pm(p), \ldots, x_\pm(p)\big) \in (\mathbb{S}^{d-1})^L$$

are hyperbolic sinks. Each attracts an open neighborhood of positive product surface measure, with locally exponential convergence to the corresponding non-spectral state.

The parameter radius ε is independent of Q, K, L and allows unrestricted joint perturbations of V, W_1, W_2; the attracting neighborhoods and convergence constants may depend on Q, K, L and p.

Together, Theorems 3.1 and 4.1 show that a local feed-forward transformation can change the asymptotic clustering of an entire causal prefix, not just the first token. The detailed proof is given in the appendix.

Remark 4.3 (Beyond attention-only equilibria)**.** Here "non-spectral" means that $x_\pm(p)$ are not eigendirections of V and hence are not equilibria of the attention-only dynamics equation 10. The MLP therefore creates attracting states that attention alone cannot sustain. It changes where tokens cluster, rather than necessarily preventing collapse.

Remark 4.4 (Query–key independence)**.** At synchronization, derivatives of the normalized attention weights cancel because all value vectors coincide. For fixed (V, W_1, W_2), the equilibrium branches and their full synchronized-state Jacobians are therefore independent of Q, K, although nonlinear trajectories and attraction basins may still depend on them.

Remark 4.5 (Robust local clustering)**.** Neither identical initial tokens nor the exact reference matrix structure is required: the basins are open in $(\mathbb{S}^{d-1})^L$, and the result persists under small joint perturbations of V, W_1, W_2. One parameter neighborhood works for every finite L, but basin sizes may depend on the parameters and L.

5 Ideas behind the proof

To illustrate the proof idea without entering the technical details, consider the first token in dimension $d = 2$, take $V^0 = \mathrm{diag}(\lambda_1, \lambda_2)$ with $\lambda_1 > \lambda_2$, and write $x(\theta) = (\cos\theta, \sin\theta)$. The attention-only equation becomes

$$\dot\theta = -(\lambda_1 - \lambda_2)\sin\theta\cos\theta.$$

Thus, whenever $\cos\theta(0) \neq 0$, we have $x(t) \to \mathrm{sign}(\cos\theta(0))e_1$, a dominant eigendirection of V^0. This is the main mechanism behind the works of Karagodin et al. (2024); Nguyen et al. (2026b;a). Now let $\sigma = \mathrm{GELU}$ and choose

$$W_1^0 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \qquad W_2^0 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

The coupled first-token dynamics then reduces to

$$\dot\theta = F_0(\theta) := -\sin\theta\,\sigma((\lambda_1 - \lambda_2)\sin\theta\cos\theta(\sin\theta - \cos\theta)).$$

Since GELU preserves the sign of its argument, the phase portrait shows that, for almost every initial angle, $\theta(t)$ converges modulo 2π to one of 0, $\pi/4$, π, or $5\pi/4$. The MLP therefore introduces two new attracting limits, $\pm u$, where $u = (e_1 + e_2)/\sqrt{2}$. Their stability and non-spectral nature follow from

$$F_0'(\pi/4) = F_0'(5\pi/4) = -\frac{\lambda_1 - \lambda_2}{4} < 0, \qquad \|\mathsf{P}_{\pm u}(V^0(\pm u))\| = \frac{\lambda_1 - \lambda_2}{2} > 0.$$

Indeed, the negative angular derivatives imply local exponential attraction by linearized stability theory, while the positive attention residual shows that these states are not equilibria of the attention-only flow. To make the perturbation argument explicit, write $p = (V, W_1, W_2)$, $p^0 = (V^0, W_1^0, W_2^0)$, $\tau(\theta) = (-\sin\theta, \cos\theta)$, and define

$$F(\theta; p) := \big\langle \tau(\theta), W_2\sigma\big(W_1 \mathsf{P}_{x(\theta)}(Vx(\theta))\big)\big\rangle .$$

This is the angular vector field for nearby parameters, with $F(\theta; p^0) = F_0(\theta)$. At $\theta_+^0 = \pi/4$ and $\theta_-^0 = 5\pi/4$, we have $F(\theta_\pm^0; p^0) = 0$ and

$$\partial_\theta F(\theta_\pm^0; p^0) = -\frac{\lambda_1 - \lambda_2}{4} \neq 0.$$

The implicit function theorem therefore yields C^1 equilibrium branches $\theta_\pm(p)$ satisfying

$$F(\theta_\pm(p); p) = 0, \qquad \theta_\pm(p^0) = \theta_\pm^0.$$

Differentiating this identity at p^0 gives

$$D_p\theta_\pm(p^0) = -\frac{D_pF(\theta_\pm^0; p^0)}{\partial_\theta F(\theta_\pm^0; p^0)} = \frac{4}{\lambda_1 - \lambda_2} D_pF(\theta_\pm^0; p^0),$$

and hence $\theta_\pm(p) = \theta_\pm^0 + O(\|p - p^0\|)$. By continuity, after shrinking the parameter neighborhood, the continued equilibria $x_\pm(p) := x(\theta_\pm(p))$ satisfy

$$\partial_\theta F(\theta_\pm(p); p) \leq -\frac{\lambda_1 - \lambda_2}{8}, \qquad \|\mathsf{P}_{x_\pm(p)}(Vx_\pm(p))\| \geq \frac{\lambda_1 - \lambda_2}{4}.$$

These bounds preserve both exponential attraction and the non-spectral property under arbitrary sufficiently small joint perturbations of V, W_1, W_2. In higher dimensions, the same argument applies in local sphere coordinates once stability in every transverse tangent direction has been established through the full tangent linearization equation 25.

For later tokens, we exploit the causal structure. Each first-token equilibrium determines a synchronized equilibrium of every finite prefix. At synchronization, all value vectors coincide, so the contributions from derivatives of the normalized attention weights cancel. The resulting linearized system equation 36 has a block lower triangular Jacobian, whose spectrum is the union of the spectra of its diagonal blocks. The linearized equations can therefore also be analyzed successively by induction on the token index. For the two-dimensional reference construction above, these equations even decouple as $\dot\eta_\ell = -(\lambda_1 - \lambda_2)\eta_\ell/4$. Under perturbations, preceding-token couplings may appear, but triangularity persists. Extending the dependence on $1/\ell$ to the compact interval $[0, 1]$ allows continuity to preserve a common negative stability margin for all diagonal blocks on one parameter neighborhood, as quantified in equation 44. Linearized stability theory then gives local exponential attraction of the synchronized configuration for every finite causal prefix.

6 Experimental Results

In this section, we empirically validate our theoretical framework at two levels: tokens motion of equations (6) and (7) near the constructed sinks (Section 6.1); and trained blocks of pretrained language models inserted into the same dynamics (Section 6.2). Every reported number is computed in the original dimension; PCA is used only as a display camera (one basis fitted once per figure or column). Pretrained checkpoints are never modified. Protocols and complete results are in Appendix D; the supplementary code is used to reproduce every results.

Certificates. For a synchronized configuration $X^* = (r, \dots, r)$ we report the equilibrium residual $res_{eq} := \big(\sum_{\ell \leq L} \|F_\ell(X^*)\|^2\big)^{1/2}$, where $F_\ell(X)$ is the right-hand side of equation (6) for token ℓ; the attention residual $\|H(r; V)\|$ with $H(x; V) := \mathsf{P}_x(Vx)$, which vanishes exactly at eigendirections of V; and the spectral abscissa $\mu_L := \mu(J_L)$ of the full tangent Jacobian J_L, computed by central differences, where $\mu(M) := \max \mathrm{Re}\,\mathrm{spec}(M)$. As in the proofs of Theorems 4.1 and 4.2, J_L is block lower triangular with diagonal blocks $A_\ell(p) = A(p, 1/\ell)$, the tangent linearizations at r of $x \mapsto \mathsf{P}_x W_2\sigma(W_1\mathsf{P}_x V(\alpha x + (1-\alpha)r))$ at $\alpha = 1/\ell$; the closed form of $A(p, \alpha)$ gives the predicted value of μ_L, and $\sup_{\alpha\in[0,1]} \mu(A(p, \alpha)) < 0$ certifies every L and every Q, K (Lemma C.5).

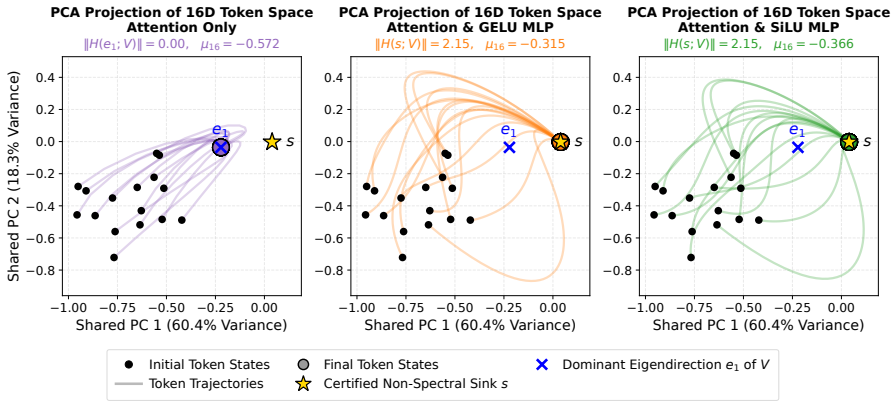


Figure 2: Comparative PCA projections of $d = 16$ causal token dynamics ($L = 16$ tokens) with Attention Only (**Left**), Attention & GELU MLP (**Middle**) and Attention & SiLU MLP (**Right**): the same non-symmetric V, the same Q, K and the same initial tokens in every panel, and one PCA basis shared by the three panels. Attention alone drives every token to the dominant eigendirection e_1 (**blue X**); with the MLP the whole prefix synchronizes at the non-spectral direction s (**gold star**). Colored line under each title: full-dimensional attention residual and prefix abscissa μ_{16} at the limit.

6.1 Tokens Dynamic Simulation

Setup. $V \in \mathbb{R}^{16\times 16}$ is non-symmetric Gaussian with a simple real dominant eigenvalue, Q, K are Gaussian, $\beta = 1$, and $\|H_s\| = 2.15$ with $H_s = \mathsf{P}_s(Vs)$. We start from the reference point of Theorem 4.2 ($W_1^0 = I_d$, $z = W_1^0 H_s$, $W_2^0 = sz^\top/\|z\|^2$) and add one rank-one term,

$$W_2 = W_2^0 + \rho\, w n^\top, \qquad n \perp \sigma(\pm z)\ \ (\sigma \in \{\mathrm{GELU}, \mathrm{SiLU}\}), \qquad w \in T_s\mathbb{S}^{d-1}, \qquad \rho = 3. \tag{14}$$

Then $\pm s$ remain exact equilibria, and $A(p, \alpha) = -c_+ I + \rho\, w\, b_\alpha^\top$ with $c_+ = z^\top \sigma(z)/\|z\|^2$ (Remark C.4) and b_α affine in α; choosing w with $b_\alpha^\top w \le 0$ at $\alpha \in \{0, 1\}$ gives $\mu_L = -c_+$ for every L and all Q, K (Appendix D.2). The proof of Theorem 4.2 therefore applies verbatim at this p; since ρ is not small, we certify rather than invoke the perturbation radius. The extra term matters for the picture: at the reference point itself W_2^0 has range $\mathrm{span}\{s\}$, every token moves on a great circle through s, and GELU and SiLU trace *identical* arcs at different speeds (Figure 6).

Results. Figure 2 shows the central contrast of the paper beyond two dimensions. From identical initial tokens and attention parameters, attention alone selects e_1 ($\mu_{16} = -0.572$ at $(e_1, \ldots, e_1)$), whereas one coupled with MLP synchronizes the whole prefix at the non-spectral s. The measured abscissae $\mu_{16} = -0.315$ (GELU) and -0.366 (SiLU) equal the closed form $-c_+$ to 2×10^{-10}; $res_{eq} \le 3 \times 10^{-13}$, and halving the RK4 step moves the saved states by less than 10^{-6} degrees. The two activations reach the same certified state along different routes. The activation shapes the transient geometry, while the location and the linear stability of the limit are fixed by the construction, as the theory predicts.

6.2 Learned blocks of pretrained models

Setup. We selected four specific architectures to execute a balanced, ablation study across the modern LLMs: Meta's LLaMA-2-7B and Mistral-7B-v0.1 (utilizing SiLU activations), alongside OpenAI's GPT-2-XL and Microsoft's Phi-2 (utilizing GELU activations). Each model reads the first WikiText-2 validation paragraph (Merity et al., 2017), fixed in advance, and is run twice: unmodified, and with every MLP output replaced by zero; weights, biases, normalizations and positional encodings are unchanged. Reading the layer index as time, we record the residual stream $X^{(k)} = [x_1^{(k)}, \ldots, x_L^{(k)}]^\top$ entering each layer ($X^{(0)}$: embeddings; $X^{(K)}$: output of the last layer), normalize every token, and draw the first 32 tokens after the attention sink (Xiao et al., 2024) in one PCA basis per model. We also report the mean pairwise cosine $\bar{c}(X^{(K)})$, which equals 1 exactly when the prefix has synchronized (Appendix D.4).

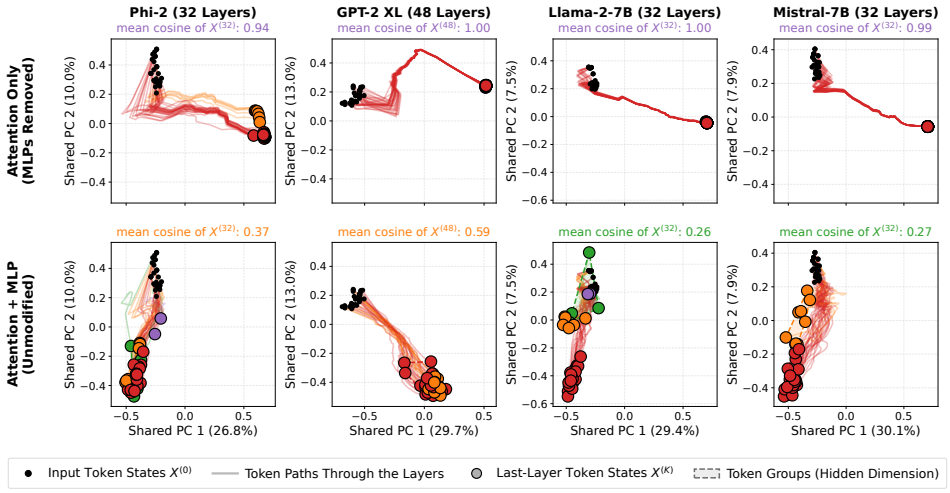


Figure 3: Comparative PCA projections of token trajectories for one prompt through all layers, normalized to the unit sphere (PCA, one basis per model). **Top:** every MLP removed (the last-layer dots overlap); **bottom:** unmodified. Black dots: $X^{(0)}$; large dots: $X^{(K)}$, colored by group in the hidden dimension (dashed hulls); subtitles: mean pairwise cosine $\bar{c}(X^{(K)})$.

Results. Removing the MLPs collapses the tokens of every model (Figure 3, top): the paths merge and their last-layer dots overlap in one group (two adjacent groups for Phi-2), and $\bar{c}(X^{(K)})$ is 0.94, 1.00, 1.00 and 0.99 (Phi-2, GPT-2 XL, Llama-2-7B, Mistral-7B). With the MLPs (bottom), the same tokens stay apart in several groups, with $\bar{c} = 0.37, 0.59, 0.26$ and 0.27. The two rows differ only by the MLP blocks, yet they reproduce the two regimes of our analysis. Without MLPs every network synchronizes its tokens, as attention-only theory predicts (Theorem 3.1(i); Dong et al., 2021; Geshkovski et al., 2025; Karagodin et al., 2024), although residual connections and normalizations are kept. With the MLPs the streams stay far from synchronization and settle in several groups, as Theorems 3.1(ii)–4.2 establish: the MLP creates attracting states that attention alone cannot sustain, and causality lets them capture whole prefixes. The contrast holds across activations, normalizations, positional encodings and block types (Table 5), and with biases and learned positions zeroed.

7 Conclusion and Limitations

We studied a minimal dynamical model of causal attention coupled with a nonlinear MLP. Our results show that the MLP can qualitatively reshape the clustering behavior induced by attention. In the attention-only setting, token dynamics are governed by the spectral structure of the value matrix and converge toward dominant eigendirections. In contrast, adding a nonlinear MLP can create new attracting equilibria that are not aligned with any eigenvector of the value matrix. We constructed an explicit example where these non-eigendirection attractors have basins of positive Lebesgue measure. We further showed that, through the triangular structure of causal attention, such MLP-induced clusters can propagate across subsequent tokens. This identifies a concrete mechanism by which local nonlinear feedforward blocks can generate global clustering patterns beyond attention-only predictions.

Limitations. Our theoretical analysis focuses on a simplified continuous-time, single-head model with spherical normalization and explicit constructions. We do not yet treat multi-head attention, general high-dimensional value matrices or complete comprehensive theory for gated MLP of GLU variants. Extending the mechanism to realistic Transformer architectures remains an important direction for future work.

AI USE STATEMENT

In this work, we used generative AI tools for language editing and polishing. We have not used generative AI tools for generating complete proofs or generating entire sections of the manuscript. We have reviewed all AI-assisted work. All AI-assisted text was carefully reviewed and revised by the authors to ensure accuracy, clarity, and consistency with the intended mathematical content. All mathematical arguments, proofs, and claims remain the responsibility of the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Considering the scope and focus of this work, we do not anticipate any adverse societal or ethical consequences.

REPRODUCIBILITY STATEMENT

The source code for all experiments is included in the paper's supplementary materials. Detailed descriptions of our experimental setup can be found in Section 6 and Appendix D. All datasets used are publicly accessible via an anonymous link provided in the README of the supplementary materials.

REFERENCES

Marah I Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat S Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Álvaro Rodríguez Abella, João Pedro Silvestre, and Paulo Tabuada. Consensus is all you get: The role of attention in transformers. In *Forty-second International Conference on Machine Learning*, 2025.

Juan A Acebrón, Luis L Bonilla, Conrad J Pérez Vicente, Félix Ritort, and Renato Spigler. The kuramoto model: A simple paradigm for synchronization phenomena. *Reviews of modern physics*, 77(1):137–185, 2005.

Rami Al-Rfou, Dokook Choe, Noah Constant, Mandy Guo, and Llion Jones. Character-level language modeling with deeper self-attention. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 3159–3166, 2019.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022.

Albert Alcalde, Giovanni Fantuzzi, and Enrique Zuazua. Clustering in pure-attention hardmax transformers and its role in sentiment analysis. *arXiv preprint arXiv:2407.01602*, 2024.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*, 2023.

Alexei Baevski and Michael Auli. Adaptive input representations for neural language modeling. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=ByxZX20qFQ.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Han Bao, Ryuichiro Hataya, and Ryo Karakida. Self-attention networks localize when qk-eigenspectrum concentrates. *arXiv preprint arXiv:2402.02098*, 2024.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. *arXiv preprint arXiv:2304.01373*, 2023.

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. Gpt-neox-20b: An open-source autoregressive language model. *arXiv preprint arXiv:2204.06745*, 2022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020a.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020b.

Martin Burger, Samira Kabri, Yury Korolev, Tim Roith, and Lukas Weigand. Analysis of mean-field models arising from self-attention dynamics in transformer architectures with layer normalization, 2025. URL https://arxiv.org/abs/2501.03096.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Shi Chen, Zhengjiang Lin, Yury Polyanskiy, and Philippe Rigollet. Critical attention scaling in long-context transformers. *International Conference on Learning Representations*, 2026.

Tianlong Chen, Zhenyu Zhang, Yu Cheng, Ahmed Awadallah, and Zhangyang Wang. The principle of diversity: Training stronger vision transformers calls for reducing all levels of redundancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12020–12030, 2022.

Jingpu Cheng, Qianxiao Li, Ting Lin, and Zuowei Shen. Interpolation, approximation and controllability of deep neural networks. *arXiv preprint arXiv:2309.06015*, 2023.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans (eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1179. URL https://aclanthology.org/D14-1179.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8493–8502, 2022.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2978–2988, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1285. URL https://aclanthology.org/P19-1285.

DeepSeek-AI et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. Universal transformers. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HyzdRiR9Y7.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019a. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019b.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International Conference on Machine Learning*, pp. 2793–2803. PMLR, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Gbetondji JS Dovonon, Michael M Bronstein, and Matt J Kusner. Setting the record straight on transformer oversmoothing. *arXiv preprint arXiv:2401.04301*, 2024.

Emilien Dupont, Arnaud Doucet, and Yee Whye Teh. Augmented neural odes. *Advances in neural information processing systems*, 32, 2019.

Subhabrata Dutta, Tanya Gautam, Soumen Chakrabarti, and Tanmoy Chakraborty. Redesigning the transformer architecture with insights from multi-particle dynamical systems. *Advances in Neural Information Processing Systems*, 34:5531–5544, 2021.

Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 107:3–11, 2018. doi: 10.1016/j.neunet.2017.12.012.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

Ruili Feng, Kecheng Zheng, Yukun Huang, Deli Zhao, Michael Jordan, and Zheng-Jun Zha. Rank diminishing in deep neural networks. *Advances in Neural Information Processing Systems*, 35: 33054–33065, 2022.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. The emergence of clusters in self-attention dynamics. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=aMjaEkkXJx.

Borjan Geshkovski, Hugo Koubbi, Yury Polyanskiy, and Philippe Rigollet. Dynamic metastability in the self-attention model. *arXiv preprint arXiv:2410.06833*, 2024a.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. The emergence of clusters in self-attention dynamics. *Advances in Neural Information Processing Systems*, 36, 2024b.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on transformers. *Bulletin of the American Mathematical Society*, 62(3):427–479, 2025. doi: 10.1090/bull/1863.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, 2021.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pp. 30–45, 2022.

Xiaojun Guo, Yifei Wang, Tianqi Du, and Yisen Wang. Contranorm: A contrastive learning perspective on oversmoothing and beyond. *arXiv preprint arXiv:2303.06562*, 2023.

Seung-Yeal Ha, Jeongho Kim, Jinyeong Park, and Xiongtao Zhang. Complete cluster predictability of the cucker–smale flocking model on the real line. *Archive for Rational Mechanics and Analysis*, 231:319–365, 2019.

Eldad Haber and Lars Ruthotto. Stable architectures for deep neural networks. *Inverse problems*, 34 (1):014004, 2017.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs), 2016. URL https://arxiv.org/abs/1606.08415.

Michael Janner, Qiyang Li, and Sergey Levine. Offline reinforcement learning as one big sequence modeling problem. *Advances in neural information processing systems*, 34:1273–1286, 2021.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Amir Joudaki, Hadi Daneshmand, and Francis Bach. On the impact of activation and normalization in obtaining isometric embeddings at initialization. *Advances in Neural Information Processing Systems*, 36:39855–39875, 2023.

Nikita Karagodin, Yury Polyanskiy, and Philippe Rigollet. Clustering in causal attention masking. *Advances in Neural Information Processing Systems*, 37:115652–115681, 2024.

Nikita Karagodin, Shu Ge, Yury Polyanskiy, and Philippe Rigollet. Normalization in attention dynamics, 2025. URL https://arxiv.org/abs/2510.22026.

Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022.

Samory Kpotufe and Abdeslam Boularias. Gradient weights help nonparametric regressors. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

Ulrich Krause et al. A discrete nonlinear and non-autonomous model of consensus formation. *Communications in difference equations*, 2000:227–236, 2000.

Yoshiki Kuramoto. Self-entrainment of a population of coupled non-linear oscillators. In *International Symposium on Mathematical Problems in Theoretical Physics: January 23–29, 1975, Kyoto University, Kyoto/Japan*, pp. 420–422. Springer, 1975.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

Qianxiao Li, Ting Lin, and Zuowei Shen. Deep learning via dynamical systems: An approximation perspective. *Journal of the European Mathematical Society*, 25(5):1671–1709, 2022.

Hongzhou Lin and Stefanie Jegelka. Resnet with one-neuron hidden layers is a universal approximator. *Advances in neural information processing systems*, 31, 2018.

Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. A survey of transformers. *AI open*, 3: 111–132, 2022.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=BJC_jUqxe.

Bingbin Liu, Jordan T Ash, Surbhi Goel, Akshay Krishnamurthy, and Cyril Zhang. Transformers learn shortcuts to automata. *arXiv preprint arXiv:2210.10749*, 2022.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Yiping Lu, Aoxiao Zhong, Quanzheng Li, and Bin Dong. Beyond finite layer neural networks: Bridging deep architectures and numerical differential equations. In *International Conference on Machine Learning*, pp. 3276–3285. PMLR, 2018.

Yiping Lu, Zhuohan Li, Di He, Zhiqing Sun, Bin Dong, Tao Qin, Liwei Wang, and Tie-Yan Liu. Understanding and improving transformer from a multi-particle dynamic system point of view. *arXiv preprint arXiv:1906.02762*, 2019.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Byj72udxe.

Microsoft. Phi-1. https://huggingface.co/microsoft/phi-1, 2023a. Hugging Face model card, accessed 2026-05-05.

Microsoft. Phi-2. https://huggingface.co/microsoft/phi-2, 2023b. Hugging Face model card, accessed 2026-05-05.

Sebastien Motsch and Eitan Tadmor. Heterophilious dynamics enhances consensus. *SIAM review*, 56 (4):577–621, 2014.

Timur Mudarisov, Mikhail Burtsev, and Radu State. Feed-forward steering in transformer residual dynamics, 2026. URL https://arxiv.org/abs/2608.02071.

Tan Nguyen, Minh Pham, Tam Nguyen, Khai Nguyen, Stanley Osher, and Nhat Ho. Fourierformer: Transformer meets generalized fourier integral theorem. *Advances in Neural Information Processing Systems*, 35:29319–29335, 2022a.

Tan Minh Nguyen, Tam Minh Nguyen, Hai Ngoc Do, Khai Nguyen, Vishwanath Saragadam, Minh Pham, Nguyen Duy Khuong, Nhat Ho, and Stanley Osher. Improving transformer with an admixture of attention heads. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022b. URL https://openreview.net/forum?id=0VFQhPGF1M3.

Trinh Tien Nguyen, Duy-Tung Pham, Hoang-Son Do, Tan Minh Nguyen, and Thieu N. Vo. Two-clustering regime of token dynamics in causal attention. In *Advances in Neural Information Processing Systems*, 2026a.

Trinh Tien Nguyen, Duy-Tung Pham, Minh-Khoi Nguyen-Nhat, Hoang-Son Do, Tan Minh Nguyen, and Thieu N. Vo. First-token attraction in mamba dynamics. In *Advances in Neural Information Processing Systems*, 2026b.

Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. Codegen: An open large language model for code with multi-turn program synthesis. *arXiv preprint arXiv:2203.13474*, 2022.

Lorenzo Noci, Sotiris Anagnostidis, Luca Biggio, Antonio Orvieto, Sidak Pal Singh, and Aurelien Lucchi. Signal propagation in transformers: Theoretical perspectives and the role of rank collapse. *Advances in Neural Information Processing Systems*, 35:27198–27211, 2022.

Lorenzo Noci, Chuning Li, Mufan Li, Bobby He, Thomas Hofmann, Chris J Maddison, and Dan Roy. The shaped transformer: Attention models in the infinite depth-and-width limit. *Advances in Neural Information Processing Systems*, 36, 2024.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. In Jian Su, Kevin Duh, and Xavier Carreras (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2249–2255, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1244. URL https://aclanthology.org/D16-1244.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019b.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pp. 8821–8831. Pmlr, 2021.

Philippe Rigollet. The mean-field dynamics of transformers. *arXiv preprint arXiv:2512.01868*, 2025.

Adam Roberts, Hyung Won Chung, Anselm Levskaya, Gaurav Mishra, James Bradbury, Daniel Andor, Sharan Narang, Brian Lester, Colin Gaffney, Afroz Mohiuddin, et al. Scaling up models and data with t5x and seqio. *arXiv preprint arXiv:2203.17189*, 2022.

Domenec Ruiz-Balet and Enrique Zuazua. Neural ode control for classification, approximation, and transport. *SIAM Review*, 65(3):735–773, 2023.

Lars Ruthotto and Eldad Haber. Deep neural networks motivated by partial differential equations. *Journal of Mathematical Imaging and Vision*, 62(3):352–364, 2020.

Michael E Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. Sinkformers: Transformers with doubly stochastic attention. In *International Conference on Artificial Intelligence and Statistics*, pp. 3515–3530. PMLR, 2022.

Michael Scholkemper, Xinyi Wu, Ali Jadbabaie, and Michael T Schaub. Residual connections and normalization can provably prevent oversmoothing in gnns. *arXiv preprint arXiv:2406.02997*, 2024.

Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Paulo Tabuada and Bahman Gharesifard. Universal approximation power of deep residual neural networks through the lens of control. *IEEE Transactions on Automatic Control*, 68(5):2715–2728, 2022.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey. *ACM Computing Surveys*, 55(6):1–28, 2022.

Gemma Team. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1452. URL https://aclanthology.org/P19-1452.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Jesse Vig and Yonatan Belinkov. Analyzing the structure of attention in a transformer language model. In Tal Linzen, Grzegorz Chrupala, Yonatan Belinkov, and Dieuwke Hupkes (eds.), *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 63–76, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4808. URL https://aclanthology.org/W19-4808.

E Weinan, Yali Duan, Linghua Kong, and Min Guo. A proposal on machine learning via dynamical systems. *Links*, 2024:08–27, 2017.

Xinyi Wu, Amir Ajorlou, Yifei Wang, Stefanie Jegelka, and Ali Jadbabaie. On the role of attention masks and LayerNorm in transformers. In *Advances in Neural Information Processing Systems*, volume 37, 2024a. doi: 10.52202/079017-0472. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/1ac3030fc57850b0fb11dfe9d4880ad7-Abstract-Conference.html.

Xinyi Wu, Amir Ajorlou, Zihui Wu, and Ali Jadbabaie. Demystifying oversmoothing in attention-based graph neural networks. *Advances in Neural Information Processing Systems*, 36, 2024b.

xAI. Grok-1: Open release. https://github.com/xai-org/grok-1, 2024. Accessed: 2026-05-05.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*, 2024.

Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Joshua M Susskind. Stabilizing transformer training by preventing attention entropy collapse. In *International Conference on Machine Learning*, pp. 40770–40803. PMLR, 2023.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in neural information processing systems*, 32, 2019.

Han Zhang, Xi Gao, Jacob Unterman, and Tom Arodz. Approximation capabilities of neural odes and invertible residual networks. In *International Conference on Machine Learning*, pp. 11086–11095. PMLR, 2020.

Appendix of "MLP Layers Refine and Expand Clusters in Causal Attention Dynamics"

Table of Contents

A Related Works **17**

B LLMs Activations: GELU and SiLU **18**

C Main Theory **18**

C.1 Model and terminology 19

C.2 Comparison of the three theorems 20

C.3 Main ideas 20

C.4 Proofs of the main theorems 21

C.5 Technical Lemmas 35

D Further Experimental Details **36**

D.1 Protocol to the simulation experiments 36

D.2 The rank-one augmented point of Figure 2 36

D.3 Theorems reference points 36

D.4 Pretrained models 37

D.5 2D Minimal Dynamics 39

A Related Works

Continuous-depth views of deep networks. A useful way to analyze deep architectures is to interpret layer index as a time variable and hidden representations as states evolving along a trajectory. This perspective is implicit in classical multilayer networks (LeCun et al., 2015) and becomes especially natural for residual architectures, where each block updates a representation by adding an increment to the current state (Chen et al., 2018; Haber & Ruthotto, 2017; Kpotufe & Boularias, 2012). Neural ODE models make this analogy precise by replacing discrete residual updates with continuous flows, allowing tools from differential equations to be used in the analysis of learning dynamics (Lin & Jegelka, 2018; Zhang et al., 2020; Li et al., 2022; Nguyen et al., 2022a; Ruiz-Balet & Zuazua, 2023; Cheng et al., 2023). This viewpoint has led to results on expressivity, approximation, interpolation, and controllability of deep models (Lin & Jegelka, 2018; Zhang et al., 2020; Li et al., 2022; Tabuada & Gharesifard, 2022; Ruiz-Balet & Zuazua, 2023; Cheng et al., 2023). Our paper adopts a similar continuous-depth formulation, but applies it to the coupled evolution induced by causal attention and a nonlinear MLP block.

Degeneracies in Transformer representations. A second line of work studies how representations evolve, and sometimes degenerate, across Transformer layers. Two related phenomena are particularly relevant: oversmoothing, where token representations become increasingly similar, and rank collapse, where hidden states concentrate near a low-dimensional subspace. Several studies show that self-attention by itself can promote such degeneracies and may not be sufficient to maintain representation diversity (Wu et al., 2024b; Dovonon et al., 2024; Chen et al., 2022). This has motivated architectural mechanisms designed to stabilize deep representations, including contrastive normalization (Guo et al., 2023), residual-normalization designs (Scholkemper et al., 2024), and other approaches aimed at preventing low-rank collapse (Dong et al., 2021; Feng et al., 2022; Noci et al., 2022; Joudaki et al., 2023; Zhai et al., 2023; Noci et al., 2024; Bao et al., 2024). These works mainly ask how collapse can be delayed, avoided, or stabilized. By contrast, our analysis asks a different question: how can the MLP change the geometry of the limiting states themselves?

Attention as interacting token dynamics. Many recent analyses model attention layers as systems of interacting particles, where tokens influence one another through data-dependent weights. This connects Transformer dynamics to broader models of collective behavior, including consensus, synchronization, and flocking-type systems (Krause et al., 2000; Motsch & Tadmor, 2014; Kuramoto, 1975; Acebrón et al., 2005; Ha et al., 2019). Related perspectives interpret self-attention through continuum or particle-inspired mechanisms, including links to convection–diffusion equations and particle-based Transformer designs (Lu et al., 2019; Dutta et al., 2021). Within this framework, a central theoretical question is whether token trajectories align, cluster, or collapse as depth increases. Several works have shown that attention dynamics can produce limiting clustered structures, metastable multi-cluster states, and eventual merging behavior (Geshkovski et al., 2024b; 2025; 2024a). Other results establish convergence to clustered equilibria in hardmax or causal attention settings, and more recently for broader classes of attention mechanisms (Alcalde et al., 2024; Karagodin et al., 2024; Abella et al., 2025). Besides, Mudarisov et al. (2026) add a local feed-forward steering field to an attention-induced residual-direction flow. They characterize nonlinear projective equilibria and tangent Jacobians, analyze invariant regions and attention–feed-forward splitting errors, and evaluate the decomposition in pretrained GPT-2, Pythia, Mistral, and Llama models. Their basic continuous model is additive in the attention and local feed-forward fields.

B LLMs Activations: GELU and SiLU

In Table 1, modern Large Language Models have nearly replaced the traditional Rectified Linear Unit (ReLU) with smoother activation functions, mainly the Gaussian Error Linear Unit (GELU) and the Sigmoid Linear Unit (SiLU). This transition primarily comes from the properties of these functions: whereas ReLU is non-differentiable at zero and strictly clips negative inputs, both GELU and SiLU provide smooth, continuous, and non-monotonic activation curves. They satisfy the functional condition that $\sigma : \mathbb{R} \to \mathbb{R}$ is C^1 and sign-preserving, with

$$\sigma(0) = 0, \qquad t\sigma(t) > 0 \quad (t \neq 0), \qquad a := \sigma'(0) > 0$$

This includes the exact GELU: $\sigma(t) = t\Phi(t)$ and its standard tanh approximation (Hendrycks & Gimpel, 2016), SiLU: $\sigma(t) = t/(1 + e^{-t})$ (Elfwing et al., 2018), and both have $a = 1/2$. Given that SiLU and GELU are the two most widely adopted activations and share highly similar mathematical geometries, they serve an identical structural purpose in overcoming the limitations of ReLU. Therefore, to maintain a focused and rigorous scope, we will work on GELU as the primary representative activation function for our proposed theory and analysis.

In addition, while the activation functions themselves improved training stability, the structural mechanism of the Feed-Forward Network (FFN) was revolutionized with the introduction of Gated Linear Units (Shazeer, 2020). Also shown in Table 1, frontier models like Llama, Mistral, and Gemma do not just use standard SiLU or GELU; they use $SwiGLU$ and $GeGLU$. In a standard FFN, the input is projected once, activated, and projected back. In a GLU variant, the model projects the input into two separate continuous matrices. The activation function (SiLU or GELU) is applied to one projection to act as a dynamic gate, which is then multiplied element-wise by the second projection:

$$\mathrm{FFN}_{\mathrm{GeGLU}}(x, W_1, V, W_2) = \big(\mathrm{GELU}(xW_1) \odot (xV)\big)W_2$$

$$\mathrm{FFN}_{\mathrm{SwiGLU}}(x, W_1, V, W_2) = \big(\mathrm{SiLU}(xW_1) \odot (xV)\big)W_2$$

C Main Theory

This section collects the assumptions and mechanisms behind Theorems 3.1, 4.1, and 4.2. The first theorem gives an explicit comparison with spectral alignment at the first token. The second establishes local synchronization of finite causal prefixes for that construction. The third constructs attracting synchronized states at prescribed directions with nonzero attention residual, without spectral assumptions on the value matrix. Table 2 summarizes their respective assumptions and conclusions; the estimates below explain how the proofs fit together.

Table 1: LLMs Activations: Primary Survey

Model / Family	Structure	Base
Meta Llama (Touvron et al., 2023)	SwiGLU	SiLU
Mistral & Mixtral (Jiang et al., 2023)	SwiGLU	SiLU
Alibaba Qwen (Bai et al., 2023)	SwiGLU	SiLU
Google PaLM (Chowdhery et al., 2022)	SwiGLU	SiLU
DeepSeek (DeepSeek-AI et al., 2024)	SwiGLU	SiLU
Microsoft Phi-3 & 4 (Abdin et al., 2024)	SwiGLU	SiLU
Google Gemma (Team, 2024)	GeGLU	GELU
xAI Grok (xAI, 2024)	GeGLU	GELU
Google FLAN-T5 (Roberts et al., 2022)	GeGLU	GELU
DeepMind Flamingo (Alayrac et al., 2022)	GeGLU	GELU
TII Falcon (Almazrouei et al., 2023)	Standard	GELU
Microsoft Phi-1 & 2 (Microsoft, 2023a;b)	Standard	GELU
OpenAI GPT-1/2/3 (Radford et al., 2019a; Brown et al., 2020b)	Standard	GELU
Meta BART (Lewis et al., 2019)	Standard	GELU
Salesforce CodeGen (Nijkamp et al., 2022)	Standard	GELU
EleutherAI Pythia (Biderman et al., 2023)	Standard	GELU
GPT-NeoX & GPT-J (Black et al., 2022)	Standard	GELU

C.1 Model and terminology

We use the autonomous, single-head projected dynamics equations (6) and (7):

$$\dot{x}_\ell = \mathsf{P}_{x_\ell} W_2 \sigma\left(W_1 \mathsf{P}_{x_\ell} \sum_{j \le \ell} a_{\ell j}(X; Q, K) V x_j \right), \qquad 1 \le \ell \le L, \tag{15}$$

where $X = (x_1, \dots, x_L) \in (\mathbb{S}^{d-1})^L$ and

$$a_{\ell j}(X; Q, K) = \frac{\exp(\langle Q x_\ell, K x_j \rangle)}{\sum_{k \le \ell} \exp(\langle Q x_\ell, K x_k \rangle)}, \qquad \mathsf{P}_x y = y - \langle x, y \rangle x.$$

The matrices are real $d \times d$ matrices, fixed in time and shared across token positions. Throughout the three theorems, $d \ge 2$ and $\sigma(t) = t\Phi(t)$ is exact GELU, applied componentwise, where Φ is the standard Gaussian distribution function. In particular,

$$\sigma(0) = 0, \qquad \sigma'(0) = \frac{1}{2}, \qquad \sigma(t) = \frac{1}{2} t + O(t^2) \quad (t \to 0). \tag{16}$$

Write $p = (V, W_1, W_2)$ and let $p^0 = (V^0, W_1^0, W_2^0)$ denote the reference triple for the theorem under consideration. We use $\| \cdot \|$ for the Euclidean norm of vectors and the Frobenius norm of matrices, so that

$$\|v\|^2 = \sum_i |v_i|^2, \qquad \|A\|^2 = \sum_{i,j} |a_{ij}|^2 \quad \text{for } A = (a_{ij}).$$

For tuples, the same notation denotes the corresponding product norm; in particular,

$$\|p - p^0\|^2 := \|V - V^0\|^2 + \|W_1 - W_1^0\|^2 + \|W_2 - W_2^0\|^2.$$

Since the first attention row contains only one token, its field is

$$H(x; V) := \mathsf{P}_x(Vx), \qquad \mathcal{F}(x; p) := \mathsf{P}_x W_2 \sigma(W_1 H(x; V)). \tag{17}$$

An equilibrium of $\mathcal{F}$ is called *non-spectral* when $H(x; V) \neq 0$. This condition is equivalent to x not being an eigendirection of V, because

$$H(x; V) = 0 \quad \iff \quad Vx = \langle x, Vx \rangle x.$$

Thus a non-spectral equilibrium of the coupled flow is not an equilibrium of the attention-only first-token flow.

Table 2: Assumptions and conclusions of the three theorems. All results use $d \geq 2$ and exact GELU, with $u = (e_1 + e_2)/\sqrt{2}$. The assumptions specify reference parameters; robustness permits arbitrary sufficiently small joint perturbations of (V, W_1, W_2). For both causal-prefix results, the parameter neighborhood is independent of Q, K, L, while attracting neighborhoods and convergence constants may depend on them.

Result	**Reference assumptions and construction**	**Conclusion**
Theorem 3.1 First token	Diagonal V^0 with $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_d$, and W_1^0, W_2^0 as specified in Theorem 3.1(ii).	At the reference parameters, attention alone selects $\pm e_1$ almost everywhere, while the coupled flow has non-spectral hyperbolic sinks at $\pm u$. These persist as nearby sinks $x_\pm(p)$ with open basins and local exponential attraction.
Theorem 4.1 Causal prefixes	The same reference construction, with $\lambda_1 + \lambda_2 > 0$ when $d \geq 3$.	Both synchronized configurations $(x_\pm(p), \ldots, x_\pm(p))$ are hyperbolic sinks. Every initial configuration in a sufficiently small open neighborhood converges exponentially to the corresponding common non-spectral state.
Theorem 4.2 Prescribed directions	Any real V^0 and unit s with $H_s := \mathsf{P}_s(V^0 s) \neq 0$. Choose W_1^0 with $z = W_1^0 H_s \neq 0$ and $W_2^0 = sz^\top/\|z\|^2$. No symmetry or eigenvalue-sign condition on V^0.	For arbitrary fixed Q, K and every finite L, the reference synchronized states $(\pm s, \ldots, \pm s)$ are hyperbolic sinks. They persist as synchronized non-spectral sinks $(x_\pm(p), \ldots, x_\pm(p))$ with open basins and local exponential attraction.

C.2 Comparison of the three theorems

For both causal-prefix results, one parameter neighborhood works for all Q, K and every finite L. The attracting neighborhoods and convergence constants may depend on the parameters and on L. The basins are open in the full sphere or product of spheres and therefore have positive surface or product surface measure. The reference structure need not survive perturbation: in particular, V may become nonsymmetric and the output matrix in Theorem 4.2 need not remain rank one. Corollary C.2 shows that the prescribed-direction construction is available for at least one direction precisely when V^0 is not a scalar multiple of the identity.

C.3 Main ideas

The first idea is constructive: we use the freedom in the feed-forward matrices to create locally attracting directions that attention alone cannot sustain. We choose W_1^0, W_2^0 so that each target direction satisfies the nonlinear equilibrium equation while retaining a nonzero attention residual. We establish local attraction through a detailed linearized analysis, computing the full tangent Jacobian at each reference equilibrium and verifying that all its eigenvalues have strictly negative real parts.

The second idea is to extend these constructions to an open set in the full parameter space. The strictly stable tangent Jacobians at the reference equilibria are invertible, so the implicit function theorem continues the first-token equilibria as C^1 branches $x_\pm(p)$ under small, independent perturbations of all entries of V, W_1, W_2. Continuity preserves both their strict stability and their nonzero attention residuals. Together with the uniform stability estimates for the causal dynamics, this yields a single open parameter neighborhood valid for arbitrary fixed Q, K and every finite prefix length under the respective theorem hypotheses. Within this neighborhood, the matrices need not retain the reference block or rank-one structure: the attracting directions vary with p and remain non-spectral. These explicit constructions and their continuation establish the robust local clustering conclusions of Theorems 3.1, 4.1, and 4.2.

C.4 PROOFS OF THE MAIN THEOREMS

We first prove Theorem 3.1, including the full tangent estimate behind the invariant-circle illustration. We then prove Theorem 4.1 using the causal diagonal blocks. Finally, we prove Theorem 4.2 and record the characterization of admissible value matrices.

Proof of Theorem 3.1. **Attention-only dynamics.** For the reference matrix

$$V^0 = \mathrm{diag}(\lambda_1, \lambda_2, \dots, \lambda_d), \qquad \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_d,$$

the solution of the attention-only equation is

$$x(t) = \frac{e^{tV^0}x(0)}{\|e^{tV^0}x(0)\|}.$$

Indeed, differentiating this expression gives

$$\dot{x} = V^0x - \langle x, V^0x\rangle x = \mathsf{P}_x(V^0x).$$

Writing

$$x(0) = \sum_{j=1}^{d} a_j e_j, \qquad \sum_{j=1}^{d} a_j^2 = 1,$$

and factoring out $e^{\lambda_1 t}$, we obtain

$$x(t) = \frac{a_1e_1 + \sum_{j=2}^{d} a_j e^{-(\lambda_1-\lambda_j)t}e_j}{\left(a_1^2 + \sum_{j=2}^{d} a_j^2 e^{-2(\lambda_1-\lambda_j)t}\right)^{1/2}}.$$

Since $\lambda_1 - \lambda_j \geq \lambda_1 - \lambda_2 > 0$ for every $j \geq 2$, whenever $a_1 \neq 0$ we have

$$x(t) \longrightarrow \mathrm{sign}(a_1)e_1 \qquad \text{as } t \to \infty.$$

The exceptional set

$$\{x \in \mathbb{S}^{d-1} : \langle x, e_1\rangle = 0\}$$

has zero $(d-1)$-dimensional surface measure. Thus convergence to one of the dominant eigendirections $\pm e_1$ holds for almost every initial state. This proves part (i).

The MLP vector field and the reference equilibria. Write the exact GELU activation as

$$\sigma(z) = z\Phi(z),$$

where Φ is the standard Gaussian distribution function. Then

$$\sigma(0) = 0, \qquad \sigma'(0) = \frac{1}{2},$$

and $\sigma(z)$ has the same sign as z. The same letter denotes its componentwise extension to $\mathbb{R}^d$.

For $p = (V, W_1, W_2)$, define

$$H(x; V) := \mathsf{P}_x(Vx),$$

and

$$\mathcal{F}(x; p) := \mathsf{P}_x\big(W_2\sigma\big(W_1H(x; V)\big)\big). \tag{18}$$

Thus the first-token equation is $\dot{x} = \mathcal{F}(x; p)$. The vector field is smooth, tangent to the sphere, and depends smoothly on all $3d^2$ parameter entries. For each fixed p, its solutions exist for all $t \geq 0$ because the sphere is compact. Set

$$v = \frac{-e_1 + e_2}{\sqrt{2}}.$$

The vectors u, v are orthonormal. For either choice of sign,

$$\langle \pm u, V^0(\pm u)\rangle = \frac{\lambda_1 + \lambda_2}{2},$$

and

$$H(\pm u; V^0) = \mp \frac{\lambda_1 - \lambda_2}{2} v. \tag{19}$$

Since $W_1^0 v = 0$, we have

$$W_1^0 H(\pm u; V^0) = 0.$$

Consequently,

$$W_2^0 \sigma\big(W_1^0 H(\pm u; V^0)\big) = 0, \qquad \mathcal{F}(\pm u; p^0) = 0.$$

Moreover, equation 19 gives

$$\|\mathsf{P}_{\pm u}(V^0(\pm u))\| = \frac{\lambda_1 - \lambda_2}{2} > 0. \tag{20}$$

Thus both reference equilibria are non-eigendirections.

Higher-dimensional sphere dynamics on the circle. Define

$$\mathcal{C} := \mathbb{S}^{d-1} \cap \operatorname{span}\{e_1, e_2\} = \{\cos\theta\, e_1 + \sin\theta\, e_2 : \theta \in \mathbb{R}\},$$

and write

$$x(\theta) = \cos\theta\, e_1 + \sin\theta\, e_2, \qquad \tau(\theta) = -\sin\theta\, e_1 + \cos\theta\, e_2.$$

A direct computation gives

$$H(x(\theta); V^0) = -(\lambda_1 - \lambda_2)\sin\theta\cos\theta\,\tau(\theta).$$

Hence

$$W_1^0 H(x(\theta); V^0) = q(\theta)(e_1 + e_2),$$

where

$$q(\theta) := (\lambda_1 - \lambda_2)\sin\theta\cos\theta(\sin\theta - \cos\theta).$$

Since σ acts componentwise and $\sigma(0) = 0$,

$$W_2^0 \sigma\big(W_1^0 H(x(\theta); V^0)\big) = \sigma(q(\theta))e_1.$$

Moreover,

$$\mathsf{P}_{x(\theta)} e_1 = -\sin\theta\,\tau(\theta).$$

Therefore,

$$\mathcal{F}(x(\theta); p^0) = -\sin\theta\,\sigma(q(\theta))\,\tau(\theta).$$

Since $\tau(\theta) \in T_{x(\theta)}\mathcal{C}$, the vector field is tangent to $\mathcal{C}$, and hence $\mathcal{C}$ is forward invariant. Thus

$$x(0) \in \mathcal{C} \quad \Longrightarrow \quad x(t) \in \mathcal{C} \qquad \text{for all } t \geq 0.$$

At the same time, we obtain

$$\dot\theta = F_0(\theta), \qquad F_0(\theta) := \langle \tau(\theta), \mathcal{F}(x(\theta); p^0)\rangle, \tag{21}$$

where

$$F_0(\theta) = -\sin\theta\,\sigma\big((\lambda_1 - \lambda_2)\sin\theta\cos\theta(\sin\theta - \cos\theta)\big). \tag{22}$$

Its equilibria in $[0, 2\pi)$ are exactly

$$0, \quad \frac{\pi}{4}, \quad \frac{\pi}{2}, \quad \pi, \quad \frac{5\pi}{4}, \quad \frac{3\pi}{2}.$$

Because $\sigma(z) = z\Phi(z)$ and $\Phi(z) > 0$, the functions $\sigma(z)$ and z have the same sign. Away from the equilibria listed above,

$$\operatorname{sgn} F_0(\theta) = \operatorname{sgn}\big(-\sin^2\theta\cos\theta(\sin\theta - \cos\theta)\big) = \operatorname{sgn}(-\cos\theta(\sin\theta - \cos\theta)).$$

The resulting sign chart is recorded in Table 3.

Table 3: Sign chart for the reference angular field F_0 on the invariant circle $\mathcal{C}$. The intervals are the connected components left after removing the six reference equilibria.

θ interval	$\operatorname{sgn}(\cos\theta)$	$\operatorname{sgn}(\sin\theta - \cos\theta)$	$\operatorname{sgn} F_0(\theta)$
$\left(0, \frac{\pi}{4}\right)$	$+$	$-$	$+$
$\left(\frac{\pi}{4}, \frac{\pi}{2}\right)$	$+$	$+$	$-$
$\left(\frac{\pi}{2}, \pi\right)$	$-$	$+$	$+$
$\left(\pi, \frac{5\pi}{4}\right)$	$-$	$+$	$+$
$\left(\frac{5\pi}{4}, \frac{3\pi}{2}\right)$	$-$	$-$	$-$
$\left(\frac{3\pi}{2}, 2\pi\right)$	$+$	$-$	$+$

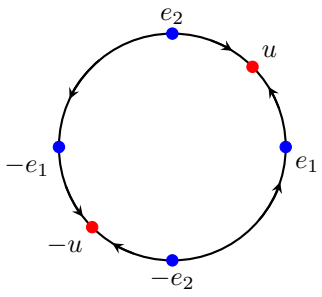


Figure 4: Reference phase portrait on the invariant circle $\mathcal{C} = \mathbb{S}^{d-1} \cap \operatorname{span}\{e_1, e_2\}$. The blue points are spectral equilibria and the red points are the additional sinks $\pm u$.

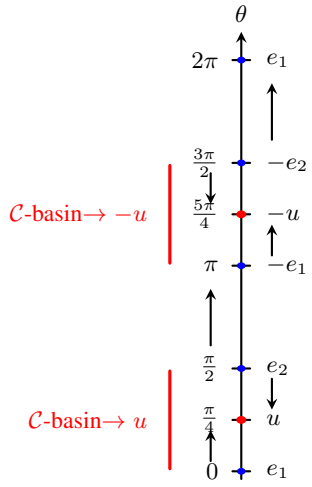


Figure 5: Vertical phase line for the reference equation equation 22 restricted to $\mathcal{C}$. Blue points are spectral equilibria, red points are the non-spectral sinks $\pm u$, and the red bars mark their exact open basins within $\mathcal{C}$. For $d > 2$, these one-dimensional basins are only a visualization; full-dimensional attraction is established by the tangent linearization below.

Thus the restricted flow points toward $\theta = \pi/4$, hence toward u, throughout the open arc $0 < \theta < \pi/2$. Likewise, it points toward $\theta = 5\pi/4$, hence toward $-u$, throughout the open arc $\pi < \theta < 3\pi/2$. Figure 4 shows this reference flow on the circle, while Figure 5 displays the equivalent vertical phase line.

Linearization in all tangent directions. We now show that the equilibria $\pm u$ attract in every tangent direction of the full sphere. The dynamics on the reference circle only involves λ_1, λ_2, while the transverse rates depend on $\lambda_3, \dots, \lambda_d$.

Fix $s \in \{u, -u\}$. The vectors $v, e_3, \dots, e_d$ form an orthonormal basis of $T_s\mathbb{S}^{d-1}$. Write

$$h = av + \sum_{k=3}^{d} b_k e_k \in T_s\mathbb{S}^{d-1},$$

with all sums over $k \geq 3$ omitted when $d = 2$. For ε small, consider the sphere-valued perturbation

$$x_\varepsilon := \frac{s + \varepsilon h}{\sqrt{1 + \varepsilon^2 \|h\|^2}} = s + \varepsilon h + O(\varepsilon^2).$$

Differentiating

$$H(x; V^0) = V^0 x - \langle x, V^0 x \rangle x$$

and using symmetry of the reference matrix gives

$$D_x H(s; V^0)h = V^0 h - \frac{\lambda_1 + \lambda_2}{2} h - 2\langle h, V^0 s \rangle s.$$

The identities

$$V^0 u = \frac{\lambda_1 + \lambda_2}{2} u - \frac{\lambda_1 - \lambda_2}{2} v,$$
$$V^0 v = -\frac{\lambda_1 - \lambda_2}{2} u + \frac{\lambda_1 + \lambda_2}{2} v$$

imply, at either $s = u$ or $s = -u$,

$$D_x H(s; V^0)v = \frac{\lambda_1 - \lambda_2}{2} u.$$

For each $k \geq 3$, we have $V^0 e_k = \lambda_k e_k$ and $\langle e_k, V^0 s \rangle = 0$, so

$$D_x H(s; V^0)e_k = -\frac{\lambda_1 + \lambda_2 - 2\lambda_k}{2} e_k.$$

Consequently,

$$H(x_\varepsilon; V^0) = H(s; V^0) + \varepsilon \left(\frac{\lambda_1 - \lambda_2}{2} au - \frac{1}{2} \sum_{k=3}^{d} (\lambda_1 + \lambda_2 - 2\lambda_k) b_k e_k \right) + O(\varepsilon^2). \tag{23}$$

Since

$$W_1^0 H(s; V^0) = 0, \qquad W_1^0 u = 2u, \qquad W_1^0 e_k = e_k \quad (k \geq 3),$$

we obtain

$$W_1^0 H(x_\varepsilon; V^0) = \varepsilon \left((\lambda_1 - \lambda_2) au - \frac{1}{2} \sum_{k=3}^{d} (\lambda_1 + \lambda_2 - 2\lambda_k) b_k e_k \right) + O(\varepsilon^2).$$

The componentwise activation satisfies

$$\sigma(\xi) = \frac{1}{2}\xi + O(\|\xi\|^2) \qquad \text{as } \xi \to 0.$$

Therefore,

$$\sigma\big(W_1^0 H(x_\varepsilon; V^0)\big) = \varepsilon \left(\frac{\lambda_1 - \lambda_2}{2} au - \frac{1}{4} \sum_{k=3}^{d} (\lambda_1 + \lambda_2 - 2\lambda_k) b_k e_k \right) + O(\varepsilon^2).$$

Using

$$W_2^0 u = \frac{u-v}{2}, \qquad W_2^0 e_k = e_k \quad (k \geq 3),$$

it follows that

$$W_2^0 \sigma\big(W_1^0 H(x_\varepsilon; V^0)\big) = \varepsilon \left(\frac{\lambda_1 - \lambda_2}{4} a(u-v) - \frac{1}{4} \sum_{k=3}^{d} (\lambda_1 + \lambda_2 - 2\lambda_k) b_k e_k \right) + O(\varepsilon^2).$$

Since

$$\mathsf{P}_{x_\varepsilon} = \mathsf{P}_s + O(\varepsilon) = \mathsf{P}_u + O(\varepsilon)$$

in the matrix norm, and the vector being projected is already of order ε, the final projection gives

$$\mathcal{F}(x_\varepsilon; p^0) = -\varepsilon \left(\frac{\lambda_1 - \lambda_2}{4} av + \frac{1}{4} \sum_{k=3}^{d} (\lambda_1 + \lambda_2 - 2\lambda_k) b_k e_k \right) + O(\varepsilon^2). \tag{24}$$

Thus the full tangent linearization is

$$D_x \mathcal{F}(\pm u; p^0) h = -\frac{\lambda_1 - \lambda_2}{4} av - \sum_{k=3}^{d} \frac{\lambda_1 + \lambda_2 - 2\lambda_k}{4} b_k e_k, \qquad h = av + \sum_{k=3}^{d} b_k e_k. \tag{25}$$

For every $k \geq 3$, the eigenvalue ordering gives

$$\lambda_1 + \lambda_2 - 2\lambda_k = (\lambda_1 - \lambda_2) + 2(\lambda_2 - \lambda_k) \geq \lambda_1 - \lambda_2 > 0.$$

It follows that

$$\langle h, D_x \mathcal{F}(\pm u; p^0) h \rangle \leq -\frac{\lambda_1 - \lambda_2}{4} \|h\|^2.$$

In particular, every tangent eigenvalue is strictly negative, without any positivity assumption on $\lambda_1, \ldots, \lambda_d$.

More explicitly, the linearized perturbation equation has solutions

$$h(t) = a(0) e^{-\frac{\lambda_1 - \lambda_2}{4} t} v + \sum_{k=3}^{d} b_k(0) e^{-\frac{\lambda_1 + \lambda_2 - 2\lambda_k}{4} t} e_k.$$

Hence

$$\|h(t)\| \leq e^{-\frac{\lambda_1 - \lambda_2}{4} t} \|h(0)\|, \qquad t \geq 0.$$

Both reference equilibria are therefore hyperbolic sinks in the full sphere.

Local coordinates and the implicit function theorem. We first show that small perturbations of $p^0 = (V^0, W_1^0, W_2^0)$ preserve the existence of equilibria near $\pm u$. More precisely, we seek exact solutions of the nonlinear equilibrium equation

$$\mathcal{F}(x_\pm(p); p) = 0, \qquad x_\pm(p^0) = \pm u.$$

We use the implicit function theorem to find these equilibria. Their stability will be proved separately below.

To apply the theorem in Euclidean coordinates, we replace the sphere-valued unknown x by its $d-1$ local coordinates y. For each sign, the map from the sphere to these coordinates is

$$\chi_\pm^{-1} : \{x \in \mathbb{S}^{d-1} : \pm\langle x, u \rangle > 0\} \longrightarrow B_1(0) \subset \mathbb{R}^{d-1},$$

given by

$$x \longmapsto y = \big(\langle x, v \rangle, \langle x, e_3 \rangle, \ldots, \langle x, e_d \rangle\big).$$

Thus y simply records the components of x along the fixed directions $v, e_3, \ldots, e_d$. In particular, $x = \pm u$ corresponds to $y = 0$ in the respective chart. When $d = 2$, there is only one coordinate, $y = \langle x, v \rangle$.

Conversely, the map from coordinates back to the sphere is

$$\chi_\pm : B_1(0) \subset \mathbb{R}^{d-1} \longrightarrow \mathbb{S}^{d-1},$$

defined by

$$y \longmapsto x = \chi_{\pm}(y) := \pm\sqrt{1 - \|y\|^2}\, u + y_1 v + \sum_{k=3}^{d} y_{k-1} e_k.$$

The sum is absent when $d = 2$. The square-root term recovers the component along u from the constraint $\|x\| = 1$ and the chosen sign of $\langle x, u\rangle$. These two maps are smooth inverses on the corresponding hemisphere, and $\chi_{\pm}(0) = \pm u$.

We now express the dynamics in these coordinates. Differentiating $y_1 = \langle x, v\rangle$ and $y_{k-1} = \langle x, e_k\rangle$ along a solution gives

$$\dot{y} = f_{\pm}(y; p), \tag{26}$$

where, with $x = \chi_{\pm}(y)$,

$$f_{\pm}(y; p) := \big(\langle v, \mathcal{F}(x; p)\rangle, \langle e_3, \mathcal{F}(x; p)\rangle, \ldots, \langle e_d, \mathcal{F}(x; p)\rangle\big).$$

Only the first component is present when $d = 2$. This is the same nonlinear flow, now written on the open Euclidean ball $B_1(0) \subset \mathbb{R}^{d-1}$.

In particular, the equilibrium equations are equivalent:

$$f_{\pm}(y; p) = 0 \quad \Longleftrightarrow \quad \mathcal{F}(\chi_{\pm}(y); p) = 0.$$

Indeed, if all the displayed components of $\mathcal{F}(x; p)$ vanish, this vector must be a multiple of u. Since it is tangent to the sphere at x and $\langle x, u\rangle \neq 0$, that multiple must be zero. The converse is immediate.

The equilibrium problem is therefore reduced to $d - 1$ nonlinear equations for the $d - 1$ coordinates y. At the reference parameters,

$$f_{\pm}(0; p^0) = 0.$$

Moreover, $D\chi_{\pm}(0)$ identifies the standard coordinate basis of $\mathbb{R}^{d-1}$ with the orthonormal tangent basis $v, e_3, \ldots, e_d$. Hence the full tangent linearization equation 25 gives

$$D_y f_{\pm}(0; p^0) = -\frac{1}{4}\,\mathrm{diag}(\lambda_1 - \lambda_2, \lambda_1 + \lambda_2 - 2\lambda_3, \ldots, \lambda_1 + \lambda_2 - 2\lambda_d)\,.$$

When $d = 2$, only the first diagonal entry is present. Every displayed entry inside the diagonal matrix is at least $\lambda_1 - \lambda_2 > 0$, so this Jacobian is invertible.

Applying the implicit function theorem with y as the unknown and all $3d^2$ entries of $p = (V, W_1, W_2)$ as parameters gives unique nearby C^1 branches $y_{\pm}(p)$ satisfying

$$f_{\pm}(y_{\pm}(p); p) = 0, \qquad y_{\pm}(p^0) = 0.$$

After intersecting the two parameter neighborhoods, choose $\varepsilon > 0$ so that both branches are defined on

$$\mathcal{P}_{\varepsilon} := \Big\{p \in \big(\mathbb{R}^{d\times d}\big)^3 : \|p - p^0\| < \varepsilon\Big\},$$

where $\|\cdot\|$ is the parameter norm in the theorem. All entries of V, W_1, W_2 may vary independently in this ball; no symmetry or block-structure constraint is imposed on the perturbed matrices.

Mapping these coordinate equilibria back to the sphere, set

$$x_{\pm}(p) := \chi_{\pm}(y_{\pm}(p)).$$

Then

$$\mathcal{F}(x_{\pm}(p); p) = 0, \qquad x_{\pm}(p^0) = \pm u.$$

These are exact equilibria of the nonlinear system, not merely solutions of its linearization. The implicit function theorem establishes their existence and local uniqueness; we next prove their attraction.

Uniform contraction and fixed forward-invariant neighborhoods. We first show that the continued equilibria remain linearly stable, and then establish an energy estimate for the full nonlinear equation.

At $(y, p) = (0, p^0)$, the coordinate Jacobian is symmetric and satisfies

$$\langle h, D_y f_{\pm}(0; p^0) h\rangle \leq -\frac{\lambda_1 - \lambda_2}{4}\|h\|^2.$$

By continuity, we can choose $0 < r < 1/2$ and shrink ε so that

$$\langle h, D_y f_\pm(y;p)h\rangle \le -\frac{\lambda_1-\lambda_2}{8}\|h\|^2 \tag{27}$$

for both signs, all $y \in \overline{B_r(0)}$, $p \in \mathcal{P}_\varepsilon$, and $h \in \mathbb{R}^{d-1}$. This is a bound on the symmetric part of the Jacobian; the perturbed Jacobian itself need not be symmetric.

Since $y_\pm(p^0) = 0$ and $f_\pm(0;p^0) = 0$, we may further shrink ε so that

$$\|y_\pm(p)\| < \frac{r}{2}, \qquad \|f_\pm(0;p)\| \le \frac{(\lambda_1-\lambda_2)r}{16}.$$

In particular, the continued equilibria lie inside $B_r(0)$. Evaluating equation 27 there shows that every eigenvalue of $D_y f_\pm(y_\pm(p);p)$ has real part at most $-(\lambda_1-\lambda_2)/8$. Thus the continued equilibria are hyperbolic sinks of the sphere vector field.

To prove nonlinear attraction, we first ensure that trajectories remain in a fixed neighborhood where the Jacobian bound holds. For $\|y\| = r$, the mean-value formula along the segment from 0 to y gives

$$\langle y, f_\pm(y;p)\rangle \le -\frac{\lambda_1-\lambda_2}{8}r^2 + r\|f_\pm(0;p)\| \le -\frac{\lambda_1-\lambda_2}{16}r^2 < 0.$$

The vector field therefore points strictly inward on the boundary of $B_r(0)$. Every solution starting in this ball stays there and exists for all $t \ge 0$.

Now fix p and either sign. We measure the distance to the exact equilibrium $y_\pm(p)$, rather than to the reference point 0. Since $f_\pm(y_\pm(p);p) = 0$ and $B_r(0)$ is convex, the mean-value formula along the segment from $y_\pm(p)$ to $y(t)$, together with equation 27, gives

$$\frac{1}{2}\frac{d}{dt}\|y(t)-y_\pm(p)\|^2 \le -\frac{\lambda_1-\lambda_2}{8}\|y(t)-y_\pm(p)\|^2.$$

Consequently,

$$\|y(t)-y_\pm(p)\| \le e^{-\frac{\lambda_1-\lambda_2}{8}t}\|y(0)-y_\pm(p)\|, \qquad t \ge 0. \tag{28}$$

This estimate holds for solutions of the full nonlinear equation. It proves nonlinear exponential attraction and also rules out any other equilibrium in $B_r(0)$.

Finally, we transfer this conclusion back to the sphere. Set

$$\mathcal{U}_\pm := \chi_\pm(B_r(0)).$$

These are fixed disjoint open neighborhoods of the two branches, independent of p, and they are forward invariant because $B_r(0)$ is forward invariant in coordinates.

For $y \in B_r(0)$ and $h \in \mathbb{R}^{d-1}$, differentiation of the chart gives

$$D\chi_\pm(y)h = \mp\frac{\langle y,h\rangle}{\sqrt{1-\|y\|^2}}u + h_1 v + \sum_{k=3}^{d} h_{k-1}e_k.$$

Orthogonality yields

$$\|D\chi_\pm(y)h\|^2 = \|h\|^2 + \frac{|\langle y,h\rangle|^2}{1-\|y\|^2} \le \frac{\|h\|^2}{1-r^2}.$$

The inverse chart records orthonormal components and is therefore 1-Lipschitz. Since $r < 1/2$, it follows that

$$\|y-z\| \le \|\chi_\pm(y)-\chi_\pm(z)\| \le 2\|y-z\|, \qquad y,z \in B_r(0).$$

Combining these inequalities with equation 28 gives

$$\|x(t)-x_\pm(p)\| \le 2e^{-\frac{\lambda_1-\lambda_2}{8}t}\|x(0)-x_\pm(p)\|, \qquad t \ge 0, \tag{29}$$

whenever $x(0) \in \mathcal{U}_\pm$.

Persistence of the non-eigendirection property. At the reference parameters, equation 20 gives

$$\|\mathsf{P}_{\pm u}(V^0(\pm u))\| = \frac{\lambda_1-\lambda_2}{2}.$$

The maps

$$p \longmapsto \left\| \mathsf{P}_{x_\pm(p)}(V x_\pm(p)) \right\|$$

are continuous. After shrinking ε once more, both branches satisfy

$$\left\| \mathsf{P}_{x_\pm(p)}(V x_\pm(p)) \right\| \geq \frac{\lambda_1 - \lambda_2}{4} > 0, \qquad p \in \mathcal{P}_\varepsilon. \tag{30}$$

For every real matrix V and every unit vector x,

$$\mathsf{P}_x(Vx) = 0 \quad \Longleftrightarrow \quad Vx = \langle x, Vx \rangle x.$$

Thus the positive residual excludes eigendirections even for nonsymmetric perturbed matrices.

Each $\mathcal{U}_\pm$ is a nonempty open subset of $\mathbb{S}^{d-1}$ and therefore has positive surface measure. Every solution starting there converges exponentially to the corresponding non-spectral equilibrium $x_\pm(p)$. Together with the C^1 dependence on all parameter entries, this proves part (ii) and completes the proof. □

Proof of Theorem 4.1.

Proof. We first obtain synchronized equilibria from the first-token theorem. We then compute their tangent linearization and show that all causal diagonal blocks remain strictly stable on one common parameter neighborhood. Triangularity and Lemma C.6 will then give nonlinear attraction.

Write

$$p = (V, W_1, W_2), \qquad p^0 = (V^0, W_1^0, W_2^0), \qquad \Theta = (Q, K, V, W_1, W_2).$$

For $X = (x_1, \ldots, x_L)$, set

$$S_\ell(X; \Theta) := \sum_{j \leq \ell} a_{\ell j}(X; Q, K) V x_j,$$

and

$$F_\ell(X; \Theta) := \mathsf{P}_{x_\ell} W_2 \sigma(W_1 \mathsf{P}_{x_\ell} S_\ell(X; \Theta)). \tag{31}$$

For fixed parameters and finite L, this is a smooth tangent vector field on the compact manifold $(\mathbb{S}^{d-1})^L$. Hence solutions exist for all $t \geq 0$.

Step 1: synchronized equilibria. We first show that the first-token equilibrium branches give exact equilibria of every finite causal prefix.

By Theorem 3.1, there are C^1 branches $x_\pm(p)$ on a neighborhood of p^0 satisfying

$$x_\pm(p^0) = \pm u, \qquad \mathcal{F}(x_\pm(p); p) = 0,$$

where

$$\mathcal{F}(x; p) := \mathsf{P}_x W_2 \sigma(W_1 \mathsf{P}_x(Vx)).$$

After shrinking the parameter neighborhood, the branches are distinct and satisfy

$$\left\| \mathsf{P}_{x_\pm(p)}(V x_\pm(p)) \right\| \geq \frac{\lambda_1 - \lambda_2}{4} > 0.$$

In particular, they are not equilibria of the attention-only dynamics.

Fix either branch and write

$$s = x_\pm(p), \qquad X^* = X^*_{\pm,L}(p) := (s, \ldots, s).$$

Since the attention weights sum to one,

$$S_\ell(X^*; \Theta) = Vs, \qquad F_\ell(X^*; \Theta) = \mathcal{F}(s; p) = 0.$$

Thus X^* is an exact equilibrium for every finite L and every fixed Q, K. The state s depends only on p.

Step 2: the attention expansion. We now compute the first-order change in the attention average near X^*. This is the step that removes Q, K from the linearization.

For tangent perturbations $h_j \in T_s\mathbb{S}^{d-1}$, set

$$x_j^\varepsilon := \frac{s+\varepsilon h_j}{\|s+\varepsilon h_j\|} = s + \varepsilon h_j + O(\varepsilon^2), \qquad X^\varepsilon = (x_1^\varepsilon, \dots, x_L^\varepsilon).$$

In the expansions below, the parameters and finite prefix length are fixed, and the remainders are uniform for h_j in bounded sets.

At synchronization, all logits in row ℓ equal $\langle Qs, Ks\rangle$. Expanding the softmax therefore gives

$$a_{\ell j}(X^\varepsilon;\Theta) = \frac{1}{\ell} + \varepsilon\alpha_{\ell j} + O(\varepsilon^2),$$

where

$$\alpha_{\ell j} = \frac{1}{\ell}\left(\langle Qs, Kh_j\rangle - \frac{1}{\ell}\sum_{k\le\ell}\langle Qs, Kh_k\rangle\right).$$

The query-perturbation term $\langle Qh_\ell, Ks\rangle$ is common to the row and cancels when the logits are normalized. Moreover,

$$\sum_{j\le\ell}\alpha_{\ell j} = 0.$$

Consequently, the first-order contribution from changing the weights is $\big(\sum_{j\le\ell}\alpha_{\ell j}\big)Vs = 0$, and

$$S_\ell(X^\varepsilon;\Theta) = Vs + \frac{\varepsilon}{\ell}\sum_{j\le\ell} Vh_j + O(\varepsilon^2). \tag{32}$$

Thus the linearized attention average is independent of Q, K.

Step 3: the full linearized ODE. We now pass the expansion through the projections and the feed-forward map, and separate the self-token term from the preceding-token terms.

Set

$$H_s := \mathsf{P}_s(Vs).$$

Expanding the inner projection in equation 32 gives

$$\widehat{x}_\ell^\varepsilon := \mathsf{P}_{x_\ell^\varepsilon} S_\ell(X^\varepsilon;\Theta) = H_s + \varepsilon q_\ell + O(\varepsilon^2), \tag{33}$$

where

$$q_\ell := \mathsf{P}_s\left(\frac{1}{\ell}\sum_{j\le\ell} Vh_j\right) - \langle h_\ell, Vs\rangle s - \langle s, Vs\rangle h_\ell.$$

Define the derivative of the feed-forward map at H_s by

$$R_s := W_2\,\mathrm{Diag}\big(\sigma'(W_1H_s)\big)W_1.$$

Here σ' acts componentwise and Diag forms the diagonal matrix from its argument.

The equilibrium condition $\mathcal{F}(s;p) = 0$ implies

$$W_2\sigma(W_1H_s) = \gamma_s s, \qquad \gamma_s := \langle s, W_2\sigma(W_1H_s)\rangle.$$

Taylor expansion of the feed-forward map yields

$$W_2\sigma(W_1\widehat{x}_\ell^\varepsilon) = \gamma_s s + \varepsilon R_s q_\ell + O(\varepsilon^2).$$

Applying the outer projection and using $\langle s, h_\ell\rangle = 0$, we obtain

$$F_\ell(X^\varepsilon;\Theta) = \varepsilon\big(\mathsf{P}_s R_s q_\ell - \gamma_s h_\ell\big) + O(\varepsilon^2). \tag{34}$$

The term $-\gamma_s h_\ell$ comes from the moving outer projection. It must be retained because the unprojected feed-forward output need not vanish for perturbed parameters.

Let $A_1(p)$ denote the first-token tangent linearization at s. The same expansion with $\ell = 1$ gives

$$A_1(p)h = \mathsf{P}_s R_s\big(\mathsf{P}_s(Vh) - \langle h, Vs\rangle s - \langle s, Vs\rangle h\big) - \gamma_s h.$$

Define also

$$B(p)h := \mathsf{P}_s R_s \mathsf{P}_s (Vh).$$

Both operators act on $T_s\mathbb{S}^{d-1}$ and refer to the chosen branch $s = x_\pm(p)$. Separating the term $j = \ell$ in q_ℓ gives

$$A_\ell(p) := A_1(p) - \left(1 - \frac{1}{\ell}\right) B(p), \tag{35}$$

and

$$F_\ell(X^\varepsilon; \Theta) = \varepsilon \left(A_\ell(p) h_\ell + \frac{1}{\ell} \sum_{j<\ell} B(p) h_j \right) + O(\varepsilon^2). \tag{36}$$

Therefore the linearized ODE is

$$\dot{h}_\ell = A_\ell(p) h_\ell + \frac{1}{\ell} \sum_{j<\ell} B(p) h_j, \qquad 1 \le \ell \le L.$$

The diagonal block is $A_\ell(p)$, every preceding-token block is $B(p)/\ell$, and all later-token blocks vanish. These formulas require no symmetry of the perturbed matrices and contain no Q, K.

Step 4: reference diagonal blocks. We next compute the spectrum of each diagonal block at p^0. Set

$$v := \frac{-e_1 + e_2}{\sqrt{2}}.$$

At either $s = u$ or $s = -u$, the reference construction gives

$$W_1^0 H_s = 0, \qquad \gamma_s = 0, \qquad R_s = \frac{1}{2} W_2^0 W_1^0.$$

The first-token calculation gives

$$A_1(p^0) v = -\frac{\lambda_1 - \lambda_2}{4} v, \tag{37}$$

and

$$A_1(p^0) e_k = -\frac{\lambda_1 + \lambda_2 - 2\lambda_k}{4} e_k, \qquad 3 \le k \le d. \tag{38}$$

Since $\mathsf{P}_u(V^0 v) = (\lambda_1 + \lambda_2)v/2$ and $W_1^0 v = 0$, we have $B(p^0)v = 0$. On the remaining coordinates, W_1^0 and W_2^0 act as the identity, so

$$B(p^0) e_k = \frac{\lambda_k}{2} e_k, \qquad 3 \le k \le d. \tag{39}$$

Substitution into equation 35 yields

$$A_\ell(p^0) v = -\frac{\lambda_1 - \lambda_2}{4} v, \tag{40}$$

and

$$A_\ell(p^0) e_k = \left(-\frac{\lambda_1 + \lambda_2}{4} + \frac{\lambda_k}{2\ell} \right) e_k, \qquad 3 \le k \le d. \tag{41}$$

Thus the basis $v, e_3, \ldots, e_d$ diagonalizes every reference block at both synchronized equilibria. When $d = 2$, only the v-direction is present.

Step 5: stability on one parameter neighborhood for all token indices. We now show that the reference blocks have a common negative margin and that this margin survives small parameter perturbations, independently of the token index.

Introduce the family

$$A(p, \alpha) := A_1(p) - (1 - \alpha) B(p), \qquad 0 \le \alpha \le 1,$$

so that $A_\ell(p) = A(p, 1/\ell)$. For $h = av + \sum_{k=3}^d b_k e_k \in T_{\pm u}\mathbb{S}^{d-1}$, Step 4 gives

$$\langle h, A(p^0, \alpha) h \rangle = -\frac{\lambda_1 - \lambda_2}{4} a^2 - \frac{1}{4} \sum_{k=3}^d (\lambda_1 + \lambda_2 - 2\alpha\lambda_k) b_k^2. \tag{42}$$

For $d \geq 3$,

$$\lambda_1 + \lambda_2 - 2\alpha\lambda_k = (1-\alpha)(\lambda_1 + \lambda_2) + \alpha(\lambda_1 + \lambda_2 - 2\lambda_k).$$

Both endpoint values are positive: the first by the additional hypothesis, and the second because $\lambda_k \leq \lambda_2$. Hence, with

$$\kappa := \frac{1}{4}\min\{\lambda_1 - \lambda_2, \lambda_1 + \lambda_2\} > 0 \qquad (d \geq 3),$$

and $\kappa := (\lambda_1 - \lambda_2)/4$ when $d = 2$, we have

$$\langle h, A(p^0, \alpha)h\rangle \leq -\kappa\|h\|^2, \qquad \alpha \in [0,1]. \tag{43}$$

We claim that, on a sufficiently small common parameter ball,

$$\langle h, A(p, \alpha)h\rangle \leq -\frac{\kappa}{2}\|h\|^2, \qquad h \in T_{x_\pm(p)}\mathbb{S}^{d-1}, \quad \alpha \in [0,1]. \tag{44}$$

To account for the moving tangent spaces, fix one branch and argue by contradiction. Failure of this estimate arbitrarily close to p^0 would give $p_n \to p^0$, $\alpha_n \in [0,1]$, and unit vectors $h_n \in T_{x_\pm(p_n)}\mathbb{S}^{d-1}$ such that

$$\langle h_n, A(p_n, \alpha_n)h_n\rangle > -\frac{\kappa}{2}.$$

By compactness, a subsequence satisfies $\alpha_n \to \alpha_0$ and $h_n \to h_0$, with $\|h_0\| = 1$. Since $\langle h_n, x_\pm(p_n)\rangle = 0$ and $x_\pm(p_n) \to \pm u$, we have $h_0 \in T_{\pm u}\mathbb{S}^{d-1}$. The explicit formulas in Step 3 depend continuously on p and h. Passing to the limit therefore gives

$$\langle h_0, A(p^0, \alpha_0)h_0\rangle \geq -\frac{\kappa}{2},$$

contradicting equation 43. Apply the same argument to the other branch and take the smaller parameter radius.

Taking $\alpha = 1/\ell$ in equation 44 shows that the symmetric part of every $A_\ell(p)$ is bounded above by $-\kappa I/2$. Consequently, every eigenvalue of every diagonal block has real part at most $-\kappa/2$. Choose the resulting radius $\varepsilon_* > 0$ no larger than the radius in Theorem 3.1. It depends neither on Q, K nor on the finite prefix length.

Step 6: nonlinear attraction of the full prefix. We finally use causal triangularity to pass from the diagonal-block estimates to stability of the full nonlinear system.

Fix $\|p - p^0\| < \varepsilon_*$, arbitrary fixed Q, K, and finite L. For either branch $s = x_\pm(p)$, use the local chart

$$\chi_s(y) := \sqrt{1 - \|y\|^2}\, s + y, \qquad y \in T_s\mathbb{S}^{d-1}, \quad \|y\| < 1,$$

whose inverse is $y = \mathsf{P}_s x$. Applying this chart to each token identifies a neighborhood of X^* with a neighborhood of 0 in $(T_s\mathbb{S}^{d-1})^L$. The coordinate vector field is

$$G_\ell(Y; \Theta) := \mathsf{P}_s F_\ell(\chi_s(y_1), \ldots, \chi_s(y_L); \Theta)\,.$$

It satisfies $G(0;\Theta) = 0$. Since $D\chi_s(0)$ is the identity on $T_s\mathbb{S}^{d-1}$, its Jacobian $J_L := D_Y G(0;\Theta)$ has the blocks computed in equation 36.

By Lemma C.5, J_L is block lower triangular and

$$\operatorname{spec}(J_L) = \bigcup_{\ell=1}^{L} \operatorname{spec}(A_\ell(p)).$$

Step 5 therefore places every eigenvalue of J_L strictly in the left half-plane. Lemma C.6 gives local exponential stability of $Y = 0$.

The product chart and its inverse are Lipschitz on a smaller neighborhood. Transferring the coordinate estimate back to the sphere gives open neighborhoods $U_{\pm,L}(\Theta)$ of $X^*_{\pm,L}(p)$ and constants $C_{\Theta,L}, c_{\Theta,L} > 0$ such that

$$\|X(t) - X^*_{\pm,L}(p)\| \leq C_{\Theta,L}e^{-c_{\Theta,L}t}\|X(0) - X^*_{\pm,L}(p)\|, \qquad t \geq 0,$$

for every $X(0) \in U_{\pm,L}(\Theta)$, where $\|X - Z\|^2 = \sum_{j=1}^{L} \|x_j - z_j\|^2$. The two neighborhoods can be chosen disjoint, with constants valid for both signs.

Each neighborhood is open in the full product manifold and hence has positive product surface measure. Thus initially distinct tokens can converge to the same state $x_\pm(p)$, which remains non-spectral by Step 1.

The parameter restriction involves only $p = (V, W_1, W_2)$. Consequently, the conclusion holds on the full open set

$$\mathcal{O} := \left\{(Q, K, V, W_1, W_2) : Q, K \in \mathbb{R}^{d\times d},\ \|p - p^0\| < \varepsilon_*\right\}.$$

The attracting neighborhoods and convergence constants may depend on the parameters and on L, whereas ε_* does not. This completes the proof. □

Remark C.1 (Why the causal result has an additional spectral condition)**.** Theorem 3.1 requires only the ordering $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_d$. For a causal prefix, the transverse diagonal rates in equation 41 tend to $-(\lambda_1 + \lambda_2)/4$ as the token index increases. When $d \geq 3$, the condition $\lambda_1 + \lambda_2 > 0$ provides a strict margin uniform over all indices. This permits negative lower eigenvalues and is weaker than $\lambda_2 > 0$. There are no transverse directions when $d = 2$, so no additional sign condition is needed.

The condition concerns this particular reference construction. Theorem 4.2 uses different feed-forward matrices and applies to arbitrary real value matrices at every direction with nonzero attention residual. In either causal result, convergence is local near a full synchronized configuration; first-token convergence alone does not impose convergence of arbitrarily initialized later tokens.

Proof of Theorem 4.2. Let

$$\mathcal{F}(x; p) := \mathsf{P}_x W_2 \sigma(W_1 \mathsf{P}_x(Vx))$$

denote the first-token vector field, and write F_ℓ for the ℓ-th component of the full causal vector field. We suppress parameter arguments when convenient. These fields are smooth and tangent to the corresponding spheres. For fixed parameters and finite L, compactness of $(\mathbb{S}^{d-1})^L$ therefore gives global existence of solutions.

Define

$$c_\pm := \frac{\sum_{i=1}^{d} z_i^2 \Phi(\pm z_i)}{\|z\|^2}, \qquad c_0 := \min\{c_+, c_-\},$$

where Φ is the standard Gaussian distribution function. Since $z \neq 0$ and $\Phi(t) > 0$ for every finite t, we have $c_\pm > 0$. The definition of W_2^0 gives

$$W_2^0 \sigma(\pm z) = c_\pm(\pm s).$$

As $\mathsf{P}_{\pm s}(V^0(\pm s)) = \pm H_s$, the outer projection annihilates these radial outputs, and hence

$$\mathcal{F}(\pm s; p^0) = 0.$$

At the reference parameters, the feed-forward output and its state derivative take values in $\operatorname{Ran} W_2^0 = \operatorname{span}\{s\}$. The latter is therefore annihilated by $\mathsf{P}_{\pm s}$, so only the derivative of the outer projection contributes to the tangent linearization. Using $D\mathsf{P}_r[h]\, r = -h$ for $h \in T_r\mathbb{S}^{d-1}$, we obtain

$$D_x\mathcal{F}(\pm s; p^0)h = -c_\pm h, \qquad h \in T_{\pm s}\mathbb{S}^{d-1}.$$

The implicit function theorem, applied in local sphere coordinates, yields C^1 equilibrium branches $x_\pm(p)$ on a common neighborhood of p^0 such that

$$\mathcal{F}(x_\pm(p); p) = 0, \qquad x_\pm(p^0) = \pm s.$$

The parameter varies in the full space $(\mathbb{R}^{d\times d})^3$, so all matrix entries may vary independently.

At the reference points,

$$\left\|\mathsf{P}_{\pm s}(V^0(\pm s))\right\| = \|H_s\| > 0.$$

By continuity, we may shrink the parameter neighborhood so that the branches remain distinct and

$$\left\|\mathsf{P}_{x_\pm(p)}(Vx_\pm(p))\right\| \geq \frac{1}{2}\|H_s\| > 0.$$

This proves equation 13. The branches and these parameter restrictions are independent of Q, K, L, since the first-token field contains neither Q nor K.

We next establish stability on one parameter neighborhood for every finite causal prefix. Fix either branch and write $r = x_{\pm}(p)$. Since the attention weights sum to one,

$$\sum_{j\le\ell} a_{\ell j}(r,\dots,r;Q,K)Vr = Vr.$$

Thus $X^* := (r,\dots,r)$ is an equilibrium of the full system.

At X^*, all logits in a given attention row coincide, so $a_{\ell j} = 1/\ell$. Moreover, the derivatives of the weights sum to zero. Since all value vectors equal Vr, their contribution to the derivative of the attention average vanishes. Consequently, for $h_j \in T_r\mathbb{S}^{d-1}$,

$$D_X\left(\sum_{j\le\ell} a_{\ell j}(X;Q,K)Vx_j\right)\Bigg|_{X=X^*}[h_1,\dots,h_L] = \frac{1}{\ell}\sum_{j\le\ell} Vh_j.$$

In particular, the full linearization at synchronization is independent of Q, K.

For $\alpha \in [0,1]$, let $A(p,\alpha)$ denote the tangent linearization at $x = r$ of the vector field

$$x \longmapsto \mathsf{P}_x W_2\sigma\big(W_1\mathsf{P}_x V\big(\alpha x + (1-\alpha)r\big)\big),$$

with r held fixed when differentiating with respect to x. This field vanishes at r. The preceding attention identity shows that $A(p,1/\ell)$ is exactly the ℓ-th diagonal block of the full tangent Jacobian: when all other token perturbations vanish, the variation of the attention average is Vh_ℓ/ℓ.

At $p = p^0$, the range argument used above applies for every $\alpha \in [0,1]$. The feed-forward output at $r = \pm s$ is $c_{\pm}r$, while its state derivative is annihilated by P_r. Hence

$$A(p^0,\alpha) = -c_{\pm}I \qquad \text{for all } \alpha \in [0,1].$$

Choose a smooth local orthonormal tangent frame near each of $\pm s$. In these frames, $A(p,\alpha)$ depends continuously on (p,α). Compactness of $[0,1]$ and $c_0 > 0$ therefore give a common radius $\varepsilon > 0$, contained in the preceding parameter neighborhood, such that both branches satisfy

$$\langle h, A(p,\alpha)h\rangle \le -\frac{c_0}{2}\|h\|^2, \qquad h \in T_{x_{\pm}(p)}\mathbb{S}^{d-1}, \quad \alpha \in [0,1], \quad \|p-p^0\| < \varepsilon.$$

Every eigenvalue of every diagonal block $A(p,1/\ell)$ thus has real part at most $-c_0/2$. The family $A(p,\alpha)$ contains neither Q nor K, and $1/\ell \in [0,1]$ for all $\ell \ge 1$. This choice of ε is consequently independent of Q, K, L.

Fix $\|p-p^0\| < \varepsilon$, arbitrary Q, K, and a finite L. By causality, the full tangent Jacobian J_L at X^* is block lower triangular, with diagonal blocks $A(p,1/\ell)$. Lemma C.5 gives

$$\operatorname{spec}(J_L) = \bigcup_{\ell=1}^{L}\operatorname{spec}(A(p,1/\ell)) \subset \left\{\mu \in \mathbb{C} : \operatorname{Re}\mu \le -\frac{c_0}{2}\right\}.$$

Hence X^* is a hyperbolic sink.

Applying Lemma C.6 in local coordinates on $(\mathbb{S}^{d-1})^L$ gives an open neighborhood U of X^* and constants $C, c > 0$ such that

$$\|X(t) - X^*\| \le Ce^{-ct}\|X(0) - X^*\|, \qquad t \ge 0, \qquad X(0) \in U,$$

where $\|X-Y\|^2 = \sum_{j=1}^{L}\|x_j - y_j\|^2$. Here the coordinate estimate transfers to the product Euclidean distance because the coordinate map and its inverse are Lipschitz on a smaller neighborhood.

Applying this argument to both branches gives open attracting neighborhoods, which may be chosen disjoint and have positive product surface measure. Their limiting states are non-spectral by the residual bound, since

$$\mathsf{P}_x(Vx) = 0 \quad\Longleftrightarrow\quad Vx = \langle x, Vx\rangle x \qquad (x \in \mathbb{S}^{d-1}).$$

The parameter radius is independent of Q, K, L, whereas the attracting neighborhoods and convergence constants may depend on the parameters and on L, as asserted. □

Corollary C.2 (Admissible value matrices)**.** *A real matrix $V^0 \in \mathbb{R}^{d\times d}$ admits a choice of feed-forward matrices with robust non-spectral synchronized sinks as in Theorem 4.2 if and only if V^0 is not a scalar multiple of the identity.*

More precisely, for every unit vector s satisfying $\mathsf{P}_s(V^0 s) \neq 0$, one may choose

$$W_1^0 = I_d, \qquad W_2^0 = \frac{s\big(\mathsf{P}_s(V^0 s)\big)^\top}{\|\mathsf{P}_s(V^0 s)\|^2}.$$

Proof. Suppose first that V^0 is not a scalar multiple of the identity. There must exist a unit vector s such that $\mathsf{P}_s(V^0 s) \neq 0$. Indeed, otherwise every nonzero vector would be an eigenvector of V^0. Applying this observation to the coordinate vectors e_i gives $V^0 e_i = \lambda_i e_i$. Applying it to $e_i + e_j$ then forces $\lambda_i = \lambda_j$ for all i, j, contradicting the assumption.

For such an s, the choice $W_1^0 = I_d$ gives $z = H_s = \mathsf{P}_s(V^0 s) \neq 0$. The conclusion follows directly from Theorem 4.2.

Conversely, if $V^0 = cI_d$, then

$$\mathsf{P}_x(V^0 x) = 0 \qquad \text{for every } x \in \mathbb{S}^{d-1}.$$

Since $\sigma(0) = 0$, the first-token field is identically zero for every choice of W_1, W_2. It has no hyperbolic sink, and every direction is spectral. Thus the claimed non-spectral synchronized-sink construction is impossible. □

Remark C.3 (Prescribed directions and the role of the feed-forward map)**.** The target direction is encoded in the range of the reference output matrix. The identities

$$W_2^0 \sigma(W_1^0 H_s) = c_+ s, \qquad \mathsf{P}_s W_2^0 = 0$$

explain the two parts of the contraction mechanism: the radial output produces a restoring term through the moving outer projection, and the range condition removes the remaining tangent response at the reference point. Stability persists on an open parameter neighborhood, which includes full-rank output matrices.

Any invertible W_1^0 is admissible when $H_s \neq 0$. The same choice $W_2^0 = sz^\top/\|z\|^2$ also works with $\sigma(t) = t$, yielding radial coefficients equal to 1 at both targets. The construction therefore establishes a mechanism of the composed feed-forward and projection maps; it does not require GELU's nonlinearity or assert that training selects the prescribed attractors.

Remark C.4 (Extension to other activations)**.** Theorems 3.1, 4.1, and 4.2 are stated and proved for exact GELU. Their conclusions also hold for any fixed componentwise activation $\sigma \in C^1(\mathbb{R})$ satisfying

$$\sigma(0) = 0, \qquad a := \sigma'(0) > 0, \qquad t\sigma(t) > 0 \quad \text{for every } t \neq 0.$$

For the first construction, the reference tangent eigenvalues are

$$-\frac{a}{2}(\lambda_1 - \lambda_2), \qquad -\frac{a}{2}(\lambda_1 + \lambda_2 - 2\lambda_k), \quad 3 \le k \le d.$$

For the causal diagonal family, they become

$$-\frac{a}{2}(\lambda_1 - \lambda_2), \qquad -\frac{a}{2}(\lambda_1 + \lambda_2 - 2\alpha\lambda_k), \quad 3 \le k \le d, \quad \alpha \in [0, 1].$$

The same spectral hypotheses therefore give a strict negative margin uniform over all token indices. For the prescribed-direction construction, replace the GELU-specific coefficients by

$$c_+ := \frac{\sum_i z_i \sigma(z_i)}{\|z\|^2}, \qquad c_- := -\frac{\sum_i z_i \sigma(-z_i)}{\|z\|^2}.$$

The sign condition gives $c_\pm > 0$, and the reference synchronized Jacobian remains $-c_\pm I_{L(d-1)}$. The same implicit-function and stability arguments then apply. With only C^1 regularity, use $\sigma(\xi) = a\xi + o(\|\xi\|)$ and $o(\varepsilon)$ remainders in the activation linearizations in place of the displayed $O(\varepsilon^2)$ remainders. For each such fixed activation, the parameter radius is still independent of Q, K, L; the radius and convergence constants may depend on σ.

C.5 TECHNICAL LEMMAS

The following lemmas connect the diagonal-block calculations with the nonlinear conclusions. The first identifies the spectrum of a causal linearization; the second turns a strictly stable spectrum into local exponential attraction, including in sphere coordinates.

Lemma C.5 (Triangular linearization). *Let $n, L \geq 1$, and consider a C^1 triangular system*

$$\dot{x}_\ell = F_\ell(x_1, \ldots, x_\ell), \qquad x_\ell \in \mathbb{R}^n, \qquad 1 \leq \ell \leq L,$$

defined on a neighborhood of an equilibrium $X^* = (x_1^*, \ldots, x_L^*)$. Define the full vector field

$$G(X) := \big(F_1(x_1), F_2(x_1, x_2), \ldots, F_L(x_1, \ldots, x_L)\big).$$

For $j \leq \ell$, set

$$C_{\ell j} := D_{x_j} F_\ell(x_1^*, \ldots, x_\ell^*).$$

Then

$$DG(X^*) = \begin{pmatrix} C_{11} & 0 & \cdots & 0 \\ C_{21} & C_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ C_{L1} & C_{L2} & \cdots & C_{LL} \end{pmatrix}.$$

The diagonal block $C_{\ell\ell}$ is exactly the linearization of the ℓ-th equation with preceding variables fixed at $x_1^*, \ldots, x_{\ell-1}^*$. Moreover,

$$\det(\zeta I_{nL} - DG(X^*)) = \prod_{\ell=1}^{L} \det(\zeta I_n - C_{\ell\ell}),$$

and hence

$$\operatorname{spec}(DG(X^*)) = \bigcup_{\ell=1}^{L} \operatorname{spec}(C_{\ell\ell}).$$

Proof. The ℓ-th vector field does not depend on x_j for $j > \ell$, so the corresponding derivatives vanish. This gives the block lower triangular form. By the definition of a partial derivative, $C_{\ell\ell}$ is the derivative with respect to x_ℓ while all preceding variables are fixed at their equilibrium values. The determinant of a block triangular matrix is the product of the determinants of its diagonal blocks. Applying this fact to $\zeta I_{nL} - DG(X^*)$ proves the characteristic-polynomial factorization and the spectral identity. □

Lemma C.6 (Linearized stability implies local nonlinear attraction). *Let $\Omega \subset \mathbb{R}^n$ be open, let $G \in C^1(\Omega; \mathbb{R}^n)$, and let $X^* \in \Omega$ satisfy $G(X^*) = 0$. Suppose every eigenvalue of $DG(X^*)$ has strictly negative real part. Then X^* is a locally exponentially stable equilibrium of*

$$\dot{X} = G(X).$$

In particular, there exist an open neighborhood $U \subset \Omega$ of X^* and constants $C, c > 0$ such that every solution with $X(0) \in U$ exists for all $t \geq 0$, stays in Ω, and satisfies

$$\|X(t) - X^*\| \leq Ce^{-ct}\|X(0) - X^*\|, \qquad t \geq 0.$$

Thus U is contained in the basin of attraction of X^*. The same conclusion applies to a C^1 system on a finite-dimensional smooth manifold after passing to local coordinates at the equilibrium.

Proof. This is the standard linearized-stability theorem for a C^1 vector field at an equilibrium with a Hurwitz Jacobian. It gives local exponential stability and therefore the stated estimate on a sufficiently small neighborhood. For a manifold system, apply the theorem to a coordinate representation. Local coordinate changes conjugate the Jacobian at an equilibrium and preserve its spectrum; on smaller coordinate neighborhoods they also preserve local exponential estimates up to multiplicative constants. □

D Further Experimental Details

This appendix gives the protocols behind Section 6, the reference-point experiments that are not in the main text, and the complete tables. All simulation experiments were run on local device CPU; and the pretrained-models runs need one GPU with at least 24 GB.

D.1 Protocol to the simulation experiments

Integration. We integrate equations (6) and (7) with the classical fourth-order Runge–Kutta method and renormalize every token after each step, which keeps the states on $(\mathbb{S}^{d-1})^L$. The step is $h = 0.02$. For every trajectory figure we repeat the integration with step $h/2$ and report the largest angle between saved states at equal times; it is below 10^{-6} degrees in Figure 2 and below 2×10^{-4} degrees in Figures 6 and 7. Plotted curves are the saved RK4 states, without smoothing, interpolation or jitter.

Equilibria and certificates. Candidate limits are refined by damped Newton iterations on the first-token field $\mathcal{F}(x;p) = \mathsf{P}_x W_2 \sigma(W_1 H(x;V))$ in moving orthonormal tangent charts. The diagonal family of the proof of Theorem 4.2 is evaluated in closed form:

$$A(p,\alpha)h = \mathsf{P}_r R_r\big(\alpha \mathsf{P}_r(Vh) - \langle h, Vr\rangle r - \langle r, Vr\rangle h\big) - \gamma_r h, \qquad h \in T_r\mathbb{S}^{d-1}, \tag{45}$$

with $R_r := W_2 \operatorname{Diag}\big(\sigma'(W_1 H(r;V))\big) W_1$ and $\gamma_r := \langle r, W_2\sigma(W_1 H(r;V))\rangle$. This is $A_1(p) - (1-\alpha)B(p)$ from the proof of Theorem 4.1, written at a general synchronized point r; we evaluate it as a $(d-1)\times(d-1)$ matrix in an orthonormal basis of $T_r\mathbb{S}^{d-1}$. The prefix abscissa is $\mu_L = \max_{\ell \le L} \mu(A(p,1/\ell))$ by Lemma C.5; we compare it with the spectral abscissa of a central-difference Jacobian of the full causal field (step 10^{-6}, retraction by normalization) and report $\sup_\alpha \mu(A(p,\alpha))$ over 101 equispaced $\alpha \in [0,1]$ as a certificate for all L and all Q, K.

Display. Each trajectory figure uses one PCA basis, fitted once to the pooled saved states of the GELU and SiLU flows and held fixed across its three panels. The basis is a camera only: no reported number is computed in it.

D.2 The rank-one augmented point of Figure 2

Table 4 lists the detailed constructed values of Figure 2. Here $b_\alpha^\top h = n^\top \operatorname{Diag}(\sigma'(z)) W_1\big(\alpha \mathsf{P}_s V h - \langle h, Vs\rangle s - \langle s, Vs\rangle h\big)$ and b_α^- is b_α with $\sigma'(-z)$ in place of $\sigma'(z)$, so that $A(p,\alpha) = -c_+ I + \rho\, w\, b_\alpha^\top$ at $+s$ and $A(p,\alpha) = -c_- I + \rho\, w\, (b_\alpha^-)^\top$ at $-s$.

D.3 Theorems reference points

Theorem 4.2 at its reference point. When $W_2 = W_2^0 = s z^\top / \|z\|^2$, the MLP output is $W_2^0 \sigma(W_1 h) = c_\sigma(h)\, s$ with $c_\sigma(h) := z^\top \sigma(W_1 h)/\|z\|^2$ (so that $c_\sigma(H_s) = c_+$), and $\dot{x}_\ell = c_\sigma(\widehat{x}_\ell)\, \mathsf{P}_{x_\ell} s$. Writing $x_\ell = \cos\theta_\ell\, s + \sin\theta_\ell\, \omega_\ell$ with a unit vector $\omega_\ell \perp s$ gives $\dot{\omega}_\ell = 0$ and $\dot{\theta}_\ell = -c_\sigma(\widehat{x}_\ell)\sin\theta_\ell$: every token moves on the fixed great circle through its initial state and s, and the activation only changes the clock c_σ. Figure 6 confirms this numerically. This is why any figure drawn at this reference point shows the same paths for every admissible activation.

Theorems 3.1 and 4.1 at their reference point. Here, the input of the MLP vanishes at $\pm u$, so the linearization depends on σ only through $a = \sigma'(0)$, which equals $1/2$ for GELU and SiLU (Remark C.4). Figure 7 shows the resulting picture in $d = 6$: identical certified rates, and paths that differ away from u, where the second-order terms of the activations ($t^2/\sqrt{2\pi}$ for GELU, $t^2/4$ for SiLU) act. The difference stays small because GELU and SiLU agree to first order at 0 and have the same asymptotes; this is why Figure 2 uses the rank-one term instead. Initial tokens are $x_\ell(0) = \cos\phi_\ell(\cos\theta_\ell e_1 + \sin\theta_\ell e_2) + \sin\phi_\ell\, \omega_\ell$ with $\theta_\ell \sim \mathcal{U}[30°, 85°]$, $\phi_\ell \sim \mathcal{U}[10°, 60°]$ and ω_ℓ a random unit vector orthogonal to e_1, e_2. The sinks $\pm u$ are local: at the exact reference point W_2^0 maps the (e_1, e_2) channel to e_1 only, so e_1 is also a (non-hyperbolic) equilibrium for later tokens, and a token that starts close to e_1 can stall there. In two of four draws all tokens reach u (in the other two, one token stalls at e_1); we show the first.

Table 4: Values of the rank-one augmented point of Figure 2. An entry spanning both columns holds for GELU and SiLU.

Quantity	GELU	SiLU
d, L, β	16, 16, 1	
V	i.i.d. $\mathcal{N}(0, 4/d)$ entries; $\lambda_1 = 2.047$, real and simple	
Q, K	i.i.d. $\mathcal{N}(0, 1/d)$ entries; $T = Q^\top K$	
s	unit; $\|H_s\| = 2.15$; 67° from e_1	
W_1^0, W_2^0	I_d; $sz^\top/\|z\|^2$ with $z = W_1^0 H_s$	
W_2	$W_2^0 + \rho\, wn^\top$ with $\rho = 3$	
n, w	unit; $n \perp \mathrm{GELU}(\pm z), \mathrm{SiLU}(\pm z)$; $w \in T_s\mathbb{S}^{d-1}$	
c_+, c_-	0.315, 0.685	0.366, 0.634
$\rho\, b_\alpha^\top w$ at $\alpha = 0, 1$	$-0.231, -0.094$	$-0.143, -0.047$
$\rho\, (b_\alpha^-)^\top w$ at $\alpha = 0, 1$	0.281, 0.152	0.193, 0.105
μ_L at $+s$, every L	-0.315	-0.366
μ_L at $-s$, $L = 1, 16, \infty$	$-0.533, -0.412, -0.404$	$-0.529, -0.446, -0.441$
$x_\ell(0)$	$(c + 0.2\,\xi_\ell)/\|c + 0.2\,\xi_\ell\|$ with $\xi_\ell \sim \mathcal{N}(0, I_d)$	
c	unit, at 60° from s	
Median pairwise angle of $x_\ell(0)$	50°	
Integration	$t \in [0, 150]$, $h = 0.02$	
Draws of (c, ξ)	four: three reach $+s$, one reaches $-s$ (certified)	
Shown	the first $+s$ draw	

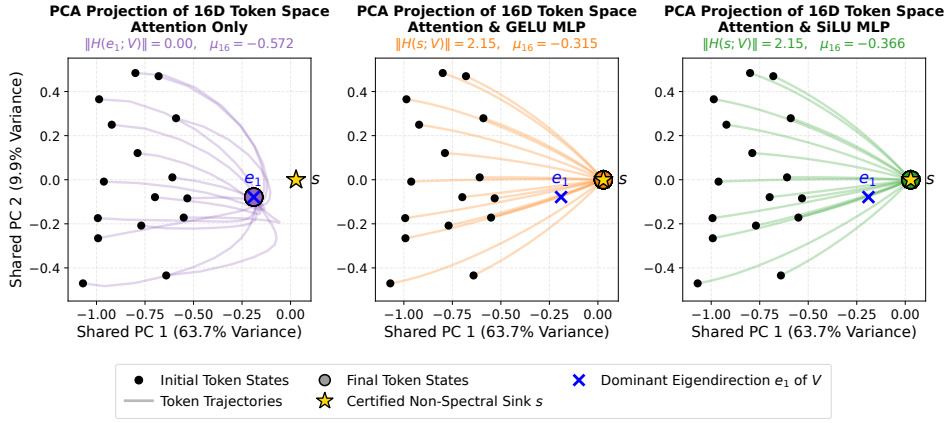


Figure 6: PCA projection of $d = 16$ causal token dynamics at the reference point of Theorem 4.2 ($\rho = 0$ in equation (14)), with the same V, Q, K, s and initial tokens as Figure 2. The certified limits are unchanged, but GELU and SiLU now follow the same great-circle arcs toward s, at different speeds.

D.4 Pretrained Models

Models. Table 5 lists the four checkpoints, loaded from the Hugging Face Hub with PyTorch and Transformers; the resolved checkpoint revision and the library versions are recorded with every run. Weights and activations are in bfloat16, the models are in evaluation mode (no dropout), and nothing is generated: each condition is a single forward pass without key–value cache.

Input. The records of the WikiText-2 (raw) validation split (Merity et al., 2017) are read in order; empty records and headings are skipped, and the first record with at least 40 whitespace-separated words is the prompt. This rule was fixed before any run; the threshold only excludes short fragments. The prompt is tokenized by each model's own tokenizer and truncated to 128 tokens; there is no padding. Position 0 acts as an attention sink with an outlier norm (Xiao et al., 2024) and is excluded together with any special token. The displayed prompt reaches the 128-token limit for every tokenizer,

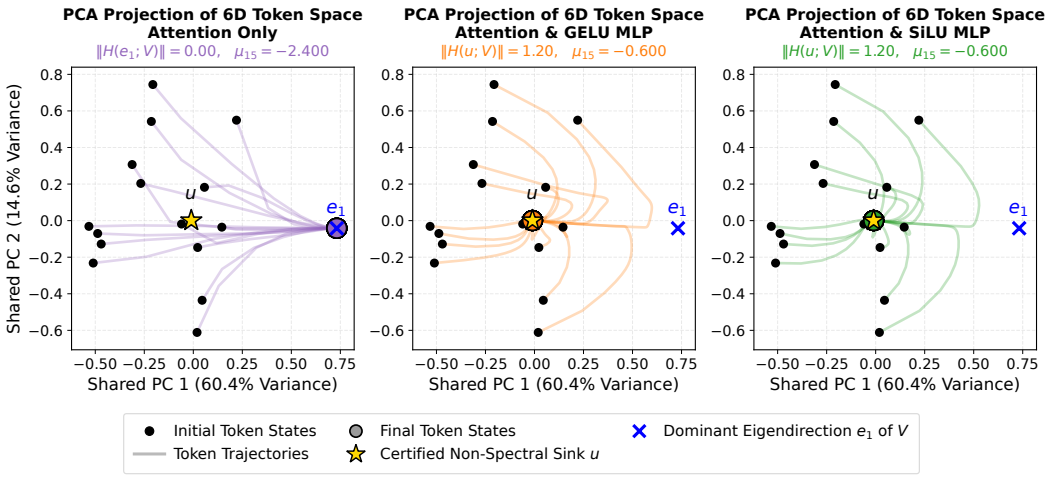


Figure 7: PCA projection of $d = 6$ causal token dynamics ($L = 15$) at the reference point of Theorems 3.1 and 4.1, $V^0 = 4\,\mathrm{diag}(1, 0.4, -0.5, -1, -1.5, -2)$, Gaussian Q, K, $\beta = 0.7$. Attention-only dynamic selects e_1; with the MLP all tokens synchronize at the non-spectral sink $u = (e_1 + e_2)/\sqrt{2}$. The Jacobian at u depends on σ only through $a = \sigma'(0) = 1/2$, so GELU and SiLU share the certified abscissa $-\frac{a}{2}(\lambda_1 - \lambda_2) = -0.600$.

Table 5: Pretrained models of Figure 3. K: number of layers; d: hidden size; MLP: hidden width and activation; block: whether attention and MLP act one after the other on the residual stream or both read the same normalized input.

Model	Hub identifier	K	d	MLP	normalization	positions	block
Phi-2	`microsoft/phi-2`	32	2560	10240, GELU	LayerNorm	rotary (partial)	parallel
GPT-2 XL	`openai-community/gpt2-xl`	48	1600	6400, GELU	LayerNorm	learned absolute	sequential
Llama-2-7B	`meta-llama/Llama-2-7b-hf`	32	4096	11008, gated SiLU	RMSNorm	rotary	sequential
Mistral-7B	`mistralai/Mistral-7B-v0.1`	32	4096	14336, gated SiLU	RMSNorm	rotary	sequential

so 127 tokens remain; over the 64 prompts used for the mean cosine below, 56 to 127 tokens remain. The whole truncated prompt is processed, and all its remaining tokens enter the mean cosine below; only the display is restricted, to the next 32 tokens, because more paths make the panels unreadable. Because attention is causal, the states of these tokens do not depend on the later ones, so they are exactly the states of a 33-token prompt. In the top row of Figure 3 all 32 tokens are drawn; their last-layer states are so close that the dots overlap (in two nearby groups for Phi-2). The token ids of the displayed tokens are stored with each run.

Intervention. In the unmodified condition, nothing is changed that hooks record at every layer k the residual stream $X^{(k)} \in \mathbb{R}^{L\times d}$ and the outputs $\Delta^{(k)}_{\mathrm{attn}}$, $\Delta^{(k)}_{\mathrm{mlp}}$ of the attention and MLP branches ($X^{(k+1)} = X^{(k)} + \Delta^{(k)}_{\mathrm{attn}} + \Delta^{(k)}_{\mathrm{mlp}}$). In the Attention-only condition, a forward hook on the MLP module of every decoder layer replaces its output by a zero tensor of the same shape, so that $X^{(k+1)} = X^{(k)} + \Delta^{(k)}_{\mathrm{attn}}$ at every layer, with $\Delta^{(k)}_{\mathrm{attn}}$ the output of the attention branch; the output bias of the MLP is removed with it. All weights, the other biases, the normalizations (including the pre-MLP normalization, whose output is no longer used), the positional encodings and the attention are unchanged.

Recorded states. A forward pre-hook on every decoder layer stores its input, $X^{(k)}$ for $k = 0, \ldots, K-1$, and a forward hook on the last layer stores its output $X^{(K)}$, before the final normalization and the output head. Each token state is divided by its Euclidean norm, $x^{(k)}_\ell \mapsto x^{(k)}_\ell / \|x^{(k)}_\ell\|$, which places it on $\mathbb{S}^{d-1}$ as in Section 2; every later step uses these unit vectors.

Groups and hulls. Groups are computed in the hidden dimension, on the 32 unit vectors $X^{(K)}$ of one panel. If the graph that links two tokens whose angle is below 15° is connected, the panel has one group. Otherwise we run spherical k-means (assignment by largest cosine, centres renormalized, at most 50 iterations, 4 random initializations, fixed seed) for $k = 2, \ldots, 6$ and keep the k with the largest mean cosine silhouette distance $(1 - \cos)$. Groups are numbered by decreasing size and colored from a fixed palette without blue. When a panel has more than one group, each group with at least three tokens is outlined by the convex hull of its projected last-layer states (dashed line, light fill). The groups only color the display: two hulls may overlap in the projection while the groups are separated in $\mathbb{R}^d$.

Projection. We use PCA, fitted once per model on 96 unit vectors: the 32 input states $X^{(0)}$ (identical in both conditions) and the 32 last-layer states $X^{(K)}$ of each condition. Every state, at every layer and in both conditions, is projected onto the two leading principal components, so the two panels of a column share the same basis and axis ranges; the axis labels give the share of variance explained by each component. No nonlinear embedding (such as t-SNE or UMAP) is used.

Mean pairwise cosine. The number above each panel is computed in $\mathbb{R}^d$ from the recorded states, not from the projection. For one prompt with token states $x_1, \ldots, x_L$ (all positions except position 0, at most 127 tokens), $\bar{c} = \binom{L}{2}^{-1} \sum_{i<j} \langle x_i, x_j \rangle / (\|x_i\| \|x_j\|)$, in double precision. It is evaluated on $X^{(K)}$ for the first 64 records selected by the rule above and averaged over them.

D.5 2D Minimal Dynamics

We isolate the fundamental mechanism in an analytically tractable two-dimensional space, explicitly visualizing how the MLP generates non-eigenvector attractors with positive-measure basins.

Experimental Setup. We consider the continuous-time limit of the first-token dynamics on the unit circle $\mathbb{S}^1$. We instantiate the value matrix as a non-diagonal matrix $V = \begin{pmatrix} 5.0 & 1.5 \\ 0.5 & 2.0 \end{pmatrix}$, in order to capture a more general setting beyond the idealized case. For completeness, we also include an experiment that follows the exact configuration specified in Theorem 3.1, where V is taken to be diagonal, $V = \mathrm{diag}(\lambda_1, \lambda_2)$, with a strict spectral gap ($\lambda_1 = 2.0, \lambda_2 = 1.0$). The two-layer MLP is still configured with $W_1 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ and $W_2 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$, utilizing a standard GELU activation.

We parameterize each token $x \in \mathbb{S}^1$ by an angle $\theta \in [0, 2\pi)$ such that $x = (\cos\theta, \sin\theta)$, and initialize $L = 30$ uniformly distributed token representations on $\mathbb{S}^1$. To avoid artifacts from theoretically unstable equilibria (i.e., the minor eigenvector at $\pm e_2$), we apply a small angular offset of $\epsilon = 0.01$ radians. The system is integrated forward for up to $T = 1500$ and $T = 120$ steps for the MLP-coupled and attention-only cases, respectively, using a common step size of $\Delta t = 0.02$. We independently simulate and track the trajectories under two regimes: (1) attention-only dynamics, and (2) attention coupled with the GELU MLP.

We also record intermediate states at selected time steps. Specifically, for the attention-only dynamics, we track the trajectories at steps $t \in \{40, 80, 120\}$ and at $t \in \{100, 500, 1500\}$ for the MLP-coupled system. Follow Theorem 3.1, to find the basins of attraction, we discretize the angular domain into a high-resolution grid ($N = 5000$ points) and compute the vector field $\dot{\theta}$ at each coordinate. The stable MLP attractors are precisely the points θ_* where the system reaches an equilibrium ($\dot{\theta} = 0$) and is locally asymptotically stable.

Results and Discussion. As visualized in Figure 8 (with non-diagonal matrix V), the convergence behavior of the token trajectories reveals a stark contrast between the two regimes, perfectly mirroring our theoretical predictions:

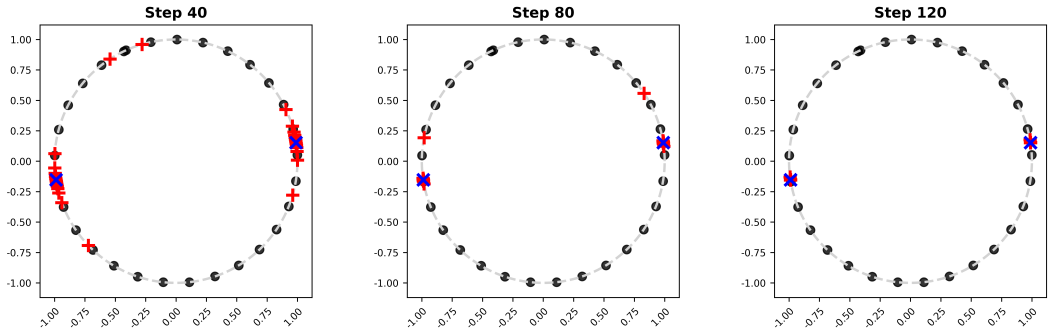
- **Attention-Only Collapse:** In the absence of the MLP, the nonlocal alignment mechanism is strictly constrained by the spectral properties of V. Within the first 40 integration steps, almost initial tokens start gradually collapsing onto the dominant eigendirections ($\pm e_1$, corresponding to nearly angles of 8.8° and 188.8°). At step 120, the result confirms that

pure attention uniformly drives the system toward a dimensional collapse precisely aligned with the principal components of the value matrix V.

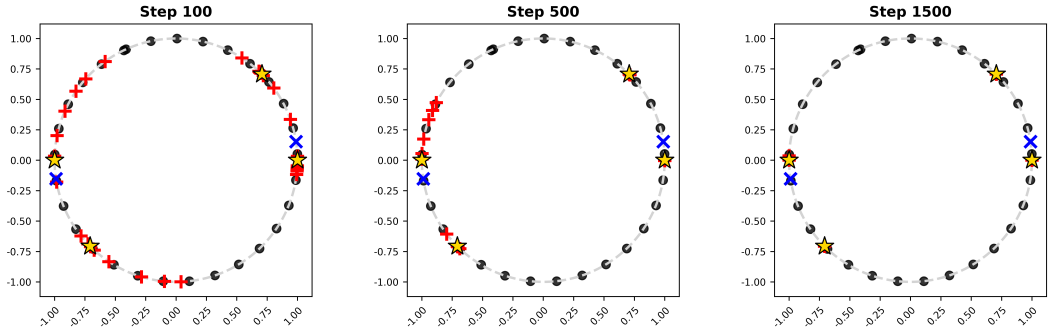
- **MLP-Coupled Expand Clustering:** When the forward pass is coupled with the GELU MLP, the asymptotic behavior is fundamentally altered. Rather than collapsing to $\pm e_1$, the trajectories are attracted to newly formed equilibria dynamically generated by the MLP's deformation of the vector field. Specifically, our numerical finding method confirms the creation of stable attractors at four angles (e.g., $\theta \approx 0$, $\theta \approx \pi/4$, $\theta \approx \pi$ and $\theta \approx 5\pi/4$). The simulation shows trajectories actively bypassing the dominant eigenvectors to converge into these new basins of attraction.

The discrepancy in the number of integration steps reflects a difference in convergence rates between the two regimes. Particularly, the attention-only dynamics exhibit significantly faster convergence toward their attractors, typically reaching a stable configuration within a small number of steps. In contrast, coupling with the GELU MLP introduces additional transformation that slow down the effective dynamics, leading to more gradual trajectory evolution and delayed convergence. Hence, we allocate a larger integration horizon for the MLP-coupled system to ensure that its asymptotic behavior is fully captured and that comparisons between the two regimes are made at convergence rather than at only intermediate transient states.

Moreover, Figure 9 (with diagonal matrix V) strongly aligns with theoretical proof of Theorem 3.1: under attention-only dynamics, token states collapse to the dominant eigendirections ($\pm e_1$, i.e., $0°$ and $180°$), whereas with the MLP, trajectories instead converge to newly induced stable equilibria at nontrivial angles (e.g., $\theta = \pi/4$ and $\theta = 5\pi/4$).

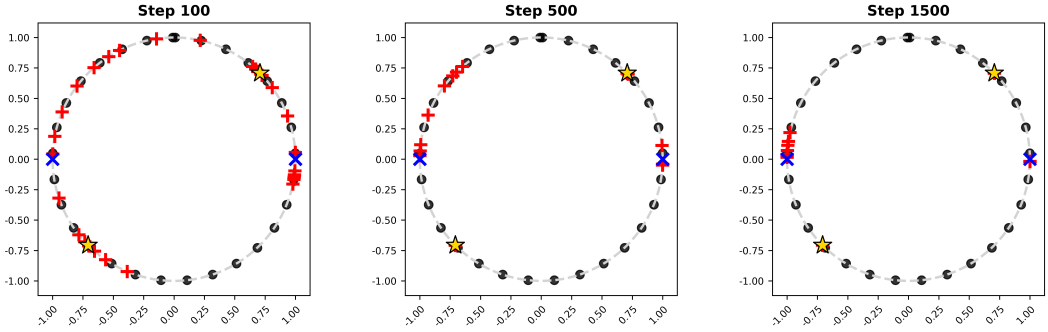


(a) Attention-Only dynamics

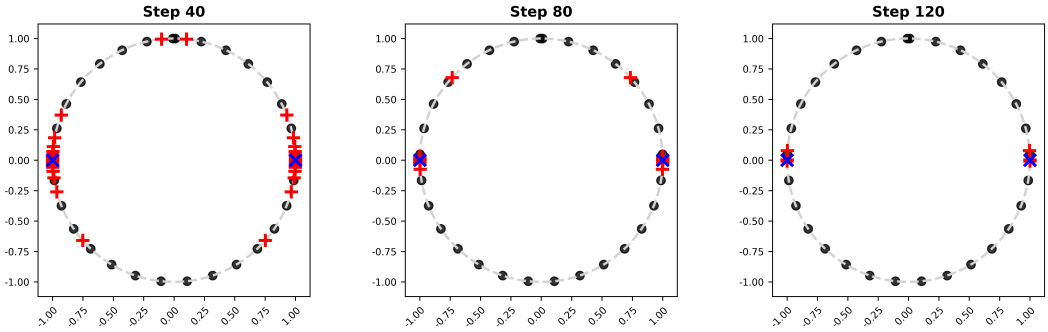


(b) Attention + GELU MLP dynamics

Figure 8: Phase portraits of token dynamics on $\mathbb{S}^1$ with non-diagonal matrix V. **(a):** Attention-only collapses to dominant eigenvectors ($\pm e_1$). **(b):** Attention + MLP introduces new stable non-eigenvector attractors. **Legend:** Initial tokens states (**Black dots**), current tokens states (**Red +**), MLP-induced attractors (**Gold stars**), dominant eigenvectors (**Blue X**).



(a) Attention + MLP dynamics



(b) Attention-Only dynamics

Figure 9: Phase portraits of token dynamics on $\mathbb{S}^1$ with diagonal matrix V. **(a):** Attention + MLP introduces new stable non-eigenvector attractors. **(b):** Attention-only collapses to dominant eigenvectors ($\pm e_1$). Visualization follows the same legend as Figure 8.