Attention by Synchronization in Coupled Oscillator Networks

Anonymous authors
Paper under double-blind review

Abstract

We address transformer attention on energy-constrained physical substrates. Softmax attention requires exponentiation and global reduction, operations with high energy cost on von Neumann hardware and no natural physical analog. We show that Kuramoto synchronization dynamics (which arise in electrical, mechanical, superconducting, and charge-density-wave oscillator arrays, among other physical systems) implement a well-defined attention operation without either. The resulting mechanism, *fixed-query oscillator attention*, replaces softmax's arithmetic with the equilibration of a gradient flow on the sphere: queries are learned anchors fixed on the sphere, and free oscillators evolve under Kuramoto–Lohe dynamics until they settle at positions encoding attention weights via cosine similarity. Because the computation is equilibration, it requires no exponentiation; the only global operation is an affine normalization at readout. The fixed point is provably unique and globally attractive from almost every initial condition, a guarantee that holds across every physical realization. Empirically, at the minimal hardware configuration (oscillator dimension d_{osc}=2), oscillator attention outperforms softmax on keyword spotting (+1.00 pp) and on subject-verb agreement (+5.27 pp on hard sentences, with zero training failures versus one in five for softmax). On causal language modeling, where softmax retains an advantage, oscillator attention closes the gap as d_{osc} grows: from +11.09 PPL at d_{osc}=2 to +2.98 PPL at d_{osc}=32 on WikiText-2, and from +2.39 PPL at d_{osc}=2 to +0.57 PPL at d_{osc}=32 on TinyStories. The main objective of this work is not to replace softmax in software but to provide a mathematically grounded blueprint for accurate attention on physical substrates.

1 Introduction

The attention mechanism is the computational core of modern transformers (Vaswani et al., 2017), but its energy cost is unavoidable on von Neumann hardware. Computing all-pairs query-key similarities and applying a global exponential normalization scales quadratically with sequence length, dominating inference cost on always-on edge devices such as wearables, embedded sensors, and autonomous systems. The sustained memory traffic exceeds the energy budgets of energy-harvested hardware (Tay et al., 2022), keeping transformer-grade inference off the edge not for lack of model quality but for lack of power.

The goal of this paper is not to replace softmax in software, where it is well-optimized, understood, and hard to improve upon. Rather, the goal is to design attention mechanisms that physical systems can implement natively, without approximating a digital algorithm in analog hardware. This requires starting from what physics does naturally and asking whether that computation can serve as attention. The key observation is that attention is fundamentally a *consensus operation*: each token settles to a distribution over its neighbors reflecting pairwise similarity, amplifying close neighbors and attenuating distant ones. Consensus is also what physical oscillator networks, as well as other multi-agent systems, compute when they equilibrate. The Kuramoto model, introduced in 1975 to describe synchronization in populations of coupled nonlinear oscillators (Kuramoto, 1975), has since been shown to govern the dynamics of remarkably diverse physical systems: electrical circuits, mechanical pendulums, superconducting Josephson junction arrays (Wiesenfeld et al., 1998), and charge-density-wave oscillator arrays (Brown et al., 2025). In each, coupled units find a shared phase configuration that balances their pairwise interactions. The energy cost of this computation is determined by the substrate, not by floating-point arithmetic. If attention is consensus, and physical oscillator networks compute consensus, then physical oscillator networks can, in principle, compute attention.

We develop this *physical intelligence* perspective (the view that a useful computation can be a property of a dynamical class rather than a specific device) by introducing *fixed-query oscillator attention*, grounded in the Lohe model (Lohe, 2009), a high-dimensional generalization of the Kuramoto equations for oscillators on a unit sphere. The mechanism uses two kinds of oscillators in different roles, generalizing the query-key structure of softmax attention: anchors are fixed reference points on the sphere (the queries), learned during training, while free oscillators evolve under input-dependent coupling weights derived from key-like projections, settling toward whichever anchors their couplings favor. At equilibrium, cosine similarities between settled oscillators and anchors yield row-stochastic attention weights via affine normalization, without exponentiation. The analysis is substrate-independent: wherever Kuramoto dynamics arise naturally, they can compute attention. This universality extends, perhaps surprisingly, to neuroscience: Kuramoto dynamics are a canonical model of neural oscillations (Breakspear et al., 2010), and the fixed-query structure (reference phases anchoring the network while input-driven oscillators synchronize to them) resembles theoretical accounts of how cortical areas use sustained oscillatory templates for selective attention (Singer & Gray, 1995; Engel et al., 2001). We do not claim a direct biological model, but if neural oscillations compute something like attention, the mechanism described here may point toward biologically plausible sequence modeling architectures.

This paper makes five contributions. First, we introduce fixed-query oscillator attention as a blueprint for physically realizable attention: we establish its closed-form fixed point for efficient training, prove that every design choice is dictated by a physical constraint of the substrate, and identify the dynamical class (coupled oscillators with attractive interactions on the sphere) whose physical realizations can implement it. Second, we prove that when the weighted anchor sum is nonzero, the dynamics have exactly two equilibria (one stable and one unstable), and every trajectory except those starting at the unstable equilibrium converges to the stable one. We characterize the probability of slow convergence for trajectories starting near the unstable equilibrium, and show that with positive coupling weights, the weighted anchor sum vanishes with probability that decays exponentially in the oscillator dimension. Third, we demonstrate that the physical mechanism is a viable attention substrate: oscillator attention outperforms softmax on keyword spotting (+1.00 pp) and on subject-verb agreement at a constrained-capacity configuration with 14 total oscillators (+5.27 pp on hard sentences, with zero training failures versus one in five for softmax) at the minimal hardware configuration (oscillator dimension d_{osc}=2, one oscillator per token), and the gap on causal language modeling closes from +11.09 PPL to +2.98 PPL on WikiText-2 and from +2.39 PPL to +0.57 PPL on TinyStories as the oscillator dimension grows from 2 to 32, following a predictable power-law decay. Fourth, ablations with a frozen random value projection confirm that the oscillator dynamics are the decisive computational element in both bidirectional tasks: on keyword spotting, the oscillator with frozen W_V retains a 24.26 pp gain over random-phase attention, isolating attention dynamics as the source; on subject-verb agreement, freezing W_V leaves accuracy nearly unchanged (−0.24 pp), confirming that learned value transformations do not drive the result. Fifth, we propose that the language modeling gap arises from a dimensional bottleneck: the oscillator attention pattern is constrained to a low-dimensional manifold whose size grows with the oscillator dimension, while softmax operates with the full feature dimension throughout. At low oscillator dimension, this constraint discards directions of similarity that softmax can use; higher oscillator dimension progressively recovers them. We support this view with the observed scaling behavior, with the gap following an approximate ${d_{\text{osc}}}^{-1/2}$ law in the oscillator dimension, providing a practical oscillator-budget design rule for target tasks. Separately, we show that readout sharpening provides a software-side optimization that improves accuracy at fixed d_{osc} (+1.16 pp on KWS at p=2, $\sim$ 0.75 PPL on causal language modeling).

2 Fixed-query oscillator attention

We introduce fixed-query oscillator attention as a drop-in replacement for the softmax attention module in a standard transformer. The mechanism is substrate-independent: the same equations describe its execution on a mechanical oscillator network, an electrical oscillator circuit, a Josephson junction array, or, for training, in software via the closed-form fixed point. The starting point is the Lohe model (Lohe, 2009), which describes the dynamics of n oscillators $x_1, \ldots, x_n$ on the unit sphere $\mathbb{S}^{d_{\text{osc}}-1} \subset \mathbb{R}^{d_{\text{osc}}}$ evolving under pairwise coupling:

$$\dot{x}_i = \Omega_i x_i + \left(I - x_i x_i^\top\right) \sum_{j=1}^{n} w_{ij}\, x_j, \tag{1}$$

where $d_{\text{osc}} \geq 2$ is the oscillator dimension, $\Omega_i \in \mathbb{R}^{d_{\text{osc}} \times d_{\text{osc}}}$ is a skew-symmetric matrix encoding the natural frequency of oscillator i, $w_{ij} \geq 0$ is the coupling weight between oscillators i and j, and the projection $(I - x_i x_i^\top)$ ensures that the trajectories evolve on the sphere. When d_{osc}=2, writing $x_i = (\cos\theta_i, \sin\theta_i)$, Equation (1) reduces to the scalar Kuramoto model (Kuramoto, 1975; Strogatz, 2000) $\dot{\theta}_i = \omega_i + \sum_j w_{ij} \sin(\theta_j - \theta_i)$, a canonical model of coupled electrical and mechanical oscillators (Appendix A shows this derivation). We set $\Omega_i = 0$ throughout; non-zero Ω_i could encode positional information and are a direction for future work.

Fixed-query oscillator attention specializes (1) to the transformer setting. Let $e_1, \ldots, e_T \in \mathbb{R}^{d_{\text{model}}}$ be input token embeddings with T the sequence length. The mechanism partitions the oscillators into two asymmetric roles, with one of each kind per input token. *Anchor oscillators* $r_j \in \mathbb{S}^{d_{\text{osc}}-1}$, $j = 1, \ldots, T$, are reference points on the sphere. They are learned during training and held fixed at inference: they do not evolve during the dynamics. The anchors play the role that learned queries play in softmax attention, but they are now positions on the sphere rather than free vectors. *Free oscillators* $z_i \in \mathbb{S}^{d_{\text{osc}}-1}$, $i = 1, \ldots, T$, are the dynamical variables. They evolve under the influence of the anchors via the Lohe equation (1), pulled by strictly positive, input-dependent coupling weights

$$w_{ij} = \sigma\left(\frac{(Fe_i)^\top (Ge_j)}{\sqrt{d_h}}\right), \tag{2}$$

where $F, G \in \mathbb{R}^{d_h \times d_{\text{model}}}$ are learned projection matrices and $\sigma : \mathbb{R} \to \mathbb{R}_{>0}$ is any strictly positive function (we use softplus, $\sigma(x) = \log(1 + e^x)$, an element-wise positive nonlinearity that avoids the global reduction required by softmax). In the analogy with softmax attention, the anchors r_j play the role of queries (fixed reference points), the free oscillators z_i play the role of keys (input-dependent states settled by the dynamics), and F, G are projection matrices analogous to W_Q, W_K in standard attention; the value projection W_V acts on the value aggregation as usual. The free oscillators follow the Lohe dynamics (1):

$$\dot{z}_i = \left(I - z_i z_i^\top\right) \underbrace{\sum_{j=1}^{T} w_{ij}\, r_j}_{h_i}, \tag{3}$$

where $h_i = \sum_j w_{ij} r_j \in \mathbb{R}^{d_{\text{osc}}}$ is the *weighted anchor sum*. Since z_i is constrained to the unit sphere, it settles at the unit vector in the direction of h_i (not at h_i itself). The dynamics (3) therefore converge to the point on $\mathbb{S}^{d_{\text{osc}}-1}$ most aligned with h_i (Section 3, Theorem 2, proves uniqueness and global stability):

$$z_i^* = \frac{h_i}{\|h_i\|}. \tag{4}$$

Cosine similarities between settled free oscillators and anchors are read out as normalized attention weights:

$$a_{ij} = \frac{1 + z_i^{*\top} r_j}{\sum_{l=1}^{T} \left(1 + z_i^{*\top} r_l\right)}, \tag{5}$$

which is a linear normalization of shifted cosine similarities; the only sum is the affine denominator $\sum_l (1 + z_i^{*\top} r_l)$, with no exponentiation.[1] For autoregressive applications, attention is restricted to past tokens by setting $w_{ij} = 0$ for $j > i$ in (2) and restricting the sum in (5) to $l \leq i$; this is standard causal masking and is orthogonal to the mechanism. The module operates with H independent attention heads of head dimension $d_h = d_{\text{model}}/H$; we have described one head, indexing all per-head quantities by (h) below. The output of head h is the weighted combination of value vectors $v_j^{(h)} = W_V^{(h)} e_j$:

$$o_i^{(h)} = \sum_{j=1}^{T} a_{ij}^{(h)}\, v_j^{(h)}. \tag{6}$$

[1] A power $p \geq 1$ in the numerator and denominator is a natural sharpening generalization that improves performance over p=1 on the tasks we evaluate (Section 4.3). We use p=1 throughout the headline experiments because it is the hardware-native readout; $p > 1$ is available as a software-side improvement when digital post-processing per oscillator is acceptable.

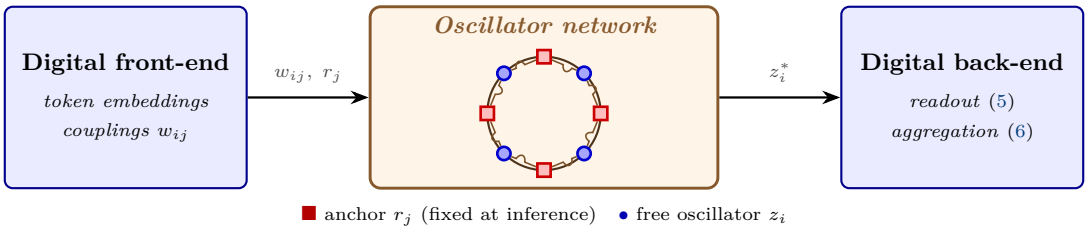

Figure 1: **Fixed-query oscillator attention.** Coupling weights w_{ij} and anchor positions r_j are computed digitally and loaded into the oscillator array. Free oscillators z_i evolve on the sphere under (3), pulled toward the fixed anchors r_j by springs encoding w_{ij}. The settled positions z_i^* are read out as attention weights (5) by the digital back-end. The equilibration runs in physical dynamics, not in von Neumann arithmetic; the energy savings come from this stage, while the digital front-end and back-end retain conventional costs.

Each head maintains its own $r_j^{(h)}$, $F^{(h)}, G^{(h)}, W_V^{(h)}$, producing output $o_i^{(h)}$ as in (6). The H head outputs are projected by $W_O \in \mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$, following the standard transformer block (Vaswani et al., 2017).

Training and inference share a single computational object. During training, the closed-form fixed point (4) is used directly: it is differentiable in all parameters and costs $O(T \cdot d_{\text{osc}})$ per token (a weighted sum over T anchors in $\mathbb{R}^{d_{\text{osc}}}$, then a normalization). At hardware inference, the oscillator array runs Equation (3) physically until it settles. Both paths converge to the same z_i^*, so there is no algorithmic gap between training and hardware inference, only a finite-time approximation gap that Section 4.4 characterizes empirically. Figure 1 shows the resulting hybrid pipeline: coupling weights and anchor positions are computed digitally in the front-end, the oscillator network equilibrates the dynamics (3), and the digital back-end reads out attention weights via (5) and feeds them to the remaining transformer block. In a purely digital setting there may be no reason to prefer oscillator attention over softmax; the latter is simpler, faster, and better understood. The oscillator mechanism is designed for physical substrates, and we evaluate it against softmax solely as a calibration of how much the physically constrained mechanism costs in accuracy relative to the unconstrained digital baseline. This cost is modest: on bidirectional tasks oscillator attention matches or exceeds softmax, and on causal language modeling the gap closes predictably with the oscillator dimension.

Remark 1 (*Substrate-dictated design choices*). Every architectural choice in the mechanism is constrained by what physical oscillator substrates can implement, rather than by software optimization. *Positive coupling ($w_{ij} > 0$)*: passive coupling elements (resistive, capacitive, inductive) are positive-valued; any coupling function $\sigma : \mathbb{R} \to \mathbb{R}_{>0}$ ensures this, and Theorem 2 shows that this guarantees a unique fixed point. *Sphere geometry ($z_i, r_j \in \mathbb{S}^{d_{\text{osc}}-1}$)*: physical oscillators have a natural state space defined by their dynamics: $\mathbb{S}^1$ for phase oscillators, $\mathbb{S}^{d_{\text{osc}}-1}$ for higher-dimensional substrates. *Fixed query anchors*: driving oscillators with a fixed reference signal — injection locking in electrical systems (Todri-Sanial et al., 2022), entrainment in mechanical and biological ones — is the standard way to steer oscillator networks toward chosen targets, and is what implements r_j in hardware. *Cosine readout, linear normalization*: inner products on the sphere are the natural observable from oscillator hardware; ${z_i^*}^\top r_j$ is what one measures. The linear normalization in the readout requires only division, which is cheap to perform digitally on the back-end; the exponentiation in softmax is precisely the operation that has no efficient physical analog, and the one we are replacing. □

3 Theoretical analysis

For a generic Kuramoto or Lohe network, multiple stable equilibria typically coexist, and which one is reached depends sensitively on initial conditions. The fixed-query design eliminates this ambiguity: because anchors r_j are external forcing terms rather than free oscillators, each free oscillator z_i minimizes a potential with exactly one minimum on $\mathbb{S}^{d_{\text{osc}}-1}$. Theorem 2 establishes uniqueness and global stability, and Propositions 3 and 4 characterize the two ways finite-time convergence can fail in practice, both controlled by the oscillator dimension d_{osc}. Throughout, we treat each free oscillator z_i independently: under the fixed-query dynamics (3), the evolution of z_i depends only on the anchors r_j and couplings w_{ij} (not on the other free oscillators) and is governed by the gradient flow induced by the weighted anchor sum $h_i = \sum_j w_{ij}\, r_j$.

Theorem 2 (Uniqueness and global stability)**.** *Given anchors $r_1, \dots, r_T \in \mathbb{S}^{d_{\text{osc}}-1}$ and weights $w_{i1}, \dots, w_{iT}$, let $h_i = \sum_{j=1}^{T} w_{ij}\, r_j$ and assume that $\|h_i\| > 0$. Then the gradient flow*

$$\dot{z}_i = (I - z_i z_i^\top)\, h_i, \qquad z_i \in \mathbb{S}^{d_{\text{osc}}-1}, \tag{7}$$

has exactly two equilibria, $z_i^* = h_i / \|h_i\|$ and $-z_i^*$. The equilibrium z_i^* is asymptotically stable with basin of attraction $\mathbb{S}^{d_{\text{osc}}-1} \setminus \{-z_i^*\}$; the equilibrium $-z_i^*$ is unstable.

Proof. Equilibria of (7) satisfy $(I - z_i z_i^\top)\, h_i = 0$. Combined with $\|z_i\| = 1$, this yields exactly two solutions $z_i = \pm z_i^*$. Consider the energy function $V(z_i) = -z_i^\top h_i$. Differentiating along the trajectories of (7),

$$\dot{V} = -h_i^\top \dot{z}_i = -h_i^\top (I - z_i z_i^\top)\, h_i = -\big(\|h_i\|^2 - (z_i^\top h_i)^2\big) \le 0,$$

where the last inequality follows from the Cauchy–Schwarz inequality applied to $z_i^\top h_i$ with $\|z_i\| = 1$. Equality holds if and only if $(z_i^\top h_i)^2 = \|h_i\|^2$ or, equivalently, if and only if $z_i = \pm z_i^*$. By LaSalle's invariance principle on the compact, positively invariant manifold $\mathbb{S}^{d_{\text{osc}}-1}$, every trajectory converges to the set $\{z_i^*, -z_i^*\}$ (Khalil, 2002). For any initial condition $z_i(0) \neq -z_i^*$, the Cauchy–Schwarz inequality gives $V(z_i(0)) < V(-z_i^*) = \|h_i\|$. Since V is non-increasing along the trajectories of (7), $V(z_i(t)) < \|h_i\|$ for all $t \ge 0$, ruling out convergence to $-z_i^*$. Hence every trajectory with $z_i(0) \neq -z_i^*$ converges asymptotically to z_i^*. □

Theorem 2 guarantees asymptotic convergence under the hypothesis $\|h_i\| > 0$, but practical inference stops the dynamics after a finite integration window, and the hypothesis itself can be violated when the anchor sum collapses. Two distinct failure modes can therefore compromise the practical implementation of the proposed scheme: (i) *degenerate anchor positions*, where $\|h_i\|$ is anomalously small and the gradient flow (7) has a vanishing driving force; and (ii) *antipodal initialization*, where the trajectory starts close to the unstable equilibrium $-z_i^*$. The softplus activation in (2) keeps all couplings w_{ij} strictly positive, ensuring that h_i is a positive combination of the anchors; this does not by itself rule out $\|h_i\| = 0$ (e.g., on $\mathbb{S}^1$ with $r_2 = -r_1$ and $w_1 = w_2$), but such exact cancellations are a measure-zero event for generic anchor configurations. More importantly, $\|h_i\|$ can still be small even when nonzero, slowing the dynamics. The two propositions below quantify both failure modes and show their probabilities decay exponentially with d_{osc}.

Proposition 3 (Degenerate positions)**.** *Let anchors $r_1, \dots, r_T$ be drawn independently and uniformly from $\mathbb{S}^{d_{\text{osc}}-1}$, $w_i = [w_{i1}, \dots, w_{iT}]^\top \in \mathbb{R}^T_{>0}$ be a vector of positive weights, and $h_i = \sum_{j=1}^{T} w_{ij}\, r_j$. Then*

$$\mathbb{E}\big[\|h_i\|^2\big] = \|w_i\|^2,$$

and for every $\varepsilon \in (0,1)$ and for some absolute constant $c > 0$

$$\Pr\big(\|h_i\|^2 \le \varepsilon\, \|w_i\|^2\big) \le \exp\big(-c\,(d_{\text{osc}} - 1)(1-\varepsilon)^2\big).$$

Proof. For r_j drawn uniformly from $\mathbb{S}^{d_{\text{osc}}-1}$,

$$\mathbb{E}[r_j] = 0, \qquad \mathbb{E}[\|r_j\|^2] = 1.$$

The first follows from antipodal symmetry of the sphere, the second from $\|r_j\| = 1$. Expanding $\|h_i\|^2 = h_i^\top h_i$ as a double sum over independent indices $j, k \in \{1, \dots, T\}$,

$$\mathbb{E}[\|h_i\|^2] = \mathbb{E}\Big[\sum_{j,k} w_{ij} w_{ik}\, r_j^\top r_k\Big] = \sum_j w_{ij}^2\, \mathbb{E}[\|r_j\|^2] + \sum_{j \neq k} w_{ij} w_{ik}\, \mathbb{E}[r_j^\top r_k] = \|w_i\|^2,$$

where the second equality uses linearity of expectation and isolates the diagonal and off-diagonal terms, and the third uses $\mathbb{E}[\|r_j\|^2] = 1$ together with independence and zero mean of the r_j ($\mathbb{E}[r_j^\top r_k] = \mathbb{E}[r_j]^\top \mathbb{E}[r_k] = 0$).

Stack the anchors into the column vector $\bar{r} = [r_1^\top, \dots, r_T^\top]^\top \in \mathbb{R}^{T d_{\text{osc}}}$, noting that $\mathbb{E}[\bar{r}] = 0$. Lévy's concentration of measure on $\mathbb{S}^{d_{\text{osc}}-1}$ (Boucheron et al., 2013, Thm. 5.7) states that for every 1-Lipschitz function $f : \mathbb{R}^{d_{\text{osc}}} \to \mathbb{R}$,

$$\Pr\big(\big|f(r_j) - \mathbb{E}[f(r_j)]\big| \ge t\big) \le 2\exp\big(-c_0 (d_{\text{osc}} - 1) t^2\big)$$

for an absolute constant $c_0 > 0$. Since this bound holds for all 1-Lipschitz functions, it holds in particular for all 1-Lipschitz convex functions. Thus, following (Adamczak, 2015, Def. 2.2), the random vector r_j satisfies the convex concentration property with constant $K = 1/\sqrt{c_0(d_{\text{osc}} - 1)}$. Since the r_j are independent, the joint vector $\bar{r}$ inherits the convex concentration property with the same constant K (Boucheron et al., 2013, Chap. 6). By Theorem 2.5 of (Adamczak, 2015), there exists an absolute constant $c_1 > 0$ such that, for every symmetric matrix $A \in \mathbb{R}^{Td_{\text{osc}} \times Td_{\text{osc}}}$ and every $t > 0$,

$$\Pr\big(\big|\bar{r}^\top A\bar{r} - \mathbb{E}[\bar{r}^\top A\bar{r}]\big| \geq t\big) \leq 2\exp\bigg(-c_1 \min\Big(\tfrac{t^2}{2K^4\,\|A\|_{\text{F}}^2}, \tfrac{t}{K^2\,\|A\|_2}\Big)\bigg). \tag{8}$$

We apply this to $A = w_i w_i^\top \otimes I_{d_{\text{osc}}} \in \mathbb{R}^{Td_{\text{osc}} \times Td_{\text{osc}}}$, the symmetric block matrix with $d_{\text{osc}} \times d_{\text{osc}}$ blocks $[A]_{jk} = w_{ij} w_{ik}\, I_{d_{\text{osc}}}$ for $j, k \in \{1, \ldots, T\}$. By block matrix multiplication,

$$\bar{r}^\top A\bar{r} = \sum_{j,k} r_j^\top [A]_{jk}\, r_k = \sum_{j,k} w_{ij} w_{ik}\, r_j^\top r_k = \|h_i\|^2 .$$

Each block $[A]_{jk} = w_{ij} w_{ik}\, I_{d_{\text{osc}}}$ has d_{osc} nonzero entries, each equal to $w_{ij} w_{ik}$, so $\|[A]_{jk}\|_{\text{F}}^2 = d_{\text{osc}}\,(w_{ij} w_{ik})^2$ and, summing,

$$\|A\|_{\text{F}}^2 = \sum_{j,k} d_{\text{osc}}\,(w_{ij} w_{ik})^2 = d_{\text{osc}} \Big(\sum_j w_{ij}^2\Big)\Big(\sum_k w_{ik}^2\Big) = d_{\text{osc}}\, \|w_i\|^4 .$$

For the spectral norm, $A = w_i w_i^\top \otimes I_{d_{\text{osc}}}$ is a Kronecker product. Its eigenvalues are pairwise products of eigenvalues of $w_i w_i^\top$ and $I_{d_{\text{osc}}}$. The matrix $w_i w_i^\top$ has eigenvalue $\|w_i\|^2$ once and zero $T-1$ times, while $I_{d_{\text{osc}}}$ has eigenvalue 1 with multiplicity d_{osc}. Therefore the nonzero eigenvalues of A are $\|w_i\|^2$ with multiplicity d_{osc} and, because A is symmetric, the spectral norm equals the largest eigenvalue: $\|A\|_2 = \|w_i\|^2$.

Substituting $K^2 = 1/(c_0(d_{\text{osc}} - 1))$, $\|A\|_{\text{F}}^2 = d_{\text{osc}}\, \|w_i\|^4$, $\|A\|_2 = \|w_i\|^2$, and $\mathbb{E}[\bar{r}^\top A\bar{r}] = \|w_i\|^2$ into (8):

$$\frac{t^2}{2K^4\,\|A\|_{\text{F}}^2} = \frac{c_0^2\,(d_{\text{osc}} - 1)^2\, t^2}{2\, d_{\text{osc}}\, \|w_i\|^4}, \qquad \frac{t}{K^2\,\|A\|_2} = \frac{c_0\,(d_{\text{osc}} - 1)\, t}{\|w_i\|^2}.$$

Hence

$$\Pr\big(\big|\|h_i\|^2 - \|w_i\|^2\big| \geq t\big) \leq 2\exp\bigg(-c_2 \min\Big(\tfrac{(d_{\text{osc}}-1)^2\, t^2}{d_{\text{osc}}\, \|w_i\|^4}, \tfrac{(d_{\text{osc}}-1)\, t}{\|w_i\|^2}\Big)\bigg), \tag{9}$$

where $c_2 > 0$ absorbs c_0, c_0^2, c_1, and the factor $1/2$.

To conclude, fix $\varepsilon \in (0, 1)$ and set $t = (1-\varepsilon)\, \|w_i\|^2$. The event $\|h_i\|^2 \leq \varepsilon\, \|w_i\|^2$ is equivalent to $\|w_i\|^2 - \|h_i\|^2 \geq (1-\varepsilon)\, \|w_i\|^2$, hence is contained in $\big\{\big|\|h_i\|^2 - \|w_i\|^2\big| \geq t\big\}$. Substituting into (9),

$$\Pr\big(\|h_i\|^2 \leq \varepsilon\, \|w_i\|^2\big) \leq 2\exp\bigg(-c_2 \min\Big(\tfrac{(d_{\text{osc}}-1)^2\,(1-\varepsilon)^2}{d_{\text{osc}}}, (d_{\text{osc}} - 1)(1-\varepsilon)\Big)\bigg).$$

For $d_{\text{osc}} \geq 2$ and $\varepsilon \in (0, 1)$, the bounds $(d_{\text{osc}} - 1)/d_{\text{osc}} \geq 1/2$ and $1 - \varepsilon \geq (1-\varepsilon)^2$ give

$$\min\Big(\tfrac{(d_{\text{osc}}-1)^2\,(1-\varepsilon)^2}{d_{\text{osc}}}, (d_{\text{osc}} - 1)(1-\varepsilon)\Big) \geq \tfrac{1}{2}\,(d_{\text{osc}} - 1)(1-\varepsilon)^2,$$

yielding

$$\Pr\big(\|h_i\|^2 \leq \varepsilon\, \|w_i\|^2\big) \leq 2\exp\big(-c\,(d_{\text{osc}} - 1)(1-\varepsilon)^2\big)$$

with $c = c_2/2$. □

The proposition shows that as d_{osc} grows, $\|h_i\|$ concentrates sharply around its mean $\|w_i\|$, so degenerate positions ($\|h_i\| \ll \|w_i\|$) become exponentially rare. In autoregressive settings where oscillator i has only a subset of active anchors (those $j \leq i$ under causal masking), the bound applies with $\|w_i\|^2 = \sum_j w_{ij}^2$ counting only those active anchors. Early positions therefore have a small absolute scale $\|w_i\|$ (few active

anchors), but the concentration is in d_{osc} rather than in the number of active anchors, so what matters for the gradient flow — the relative magnitude $\|h_i\| / \|w_i\|$ — is controlled regardless of position. More anchors do not help; higher-dimensional anchors do. This bound assumes uniformly distributed anchors and is therefore an initialization-time guarantee; Table 6 confirms empirically that the concentration survives training.

The second failure mode is independent of $\|h_i\|$ and instead concerns the initial condition. Theorem 2 guarantees convergence to z_i^* for every $z_i(0) \neq -z_i^*$, but trajectories initialized close to the unstable equilibrium $-z_i^*$ leave its neighborhood only slowly, and a finite integration window may end before the trajectory has reached a useful neighborhood of z_i^*. To quantify how often this happens, we model an initialization in which the free oscillator is independent of the dynamics, with $z_i(0)$ drawn uniformly from $\mathbb{S}^{d_{\text{osc}}-1}$. This corresponds to a cold start (e.g., a random phase set by injection-lock noise) and provides a calibrated upper bound on how often, under uninformative initialization, a trajectory begins close enough to $-z_i^*$ to fail convergence within a finite integration window.

Proposition 4 (Antipodal initialization)**.** *Let $z_i(0)$ be drawn uniformly from $\mathbb{S}^{d_{\text{osc}}-1}$, independently of the anchors and weights, and let $z_i^* \in \mathbb{S}^{d_{\text{osc}}-1}$ be the stable equilibrium of (7). For every $\alpha \in (0, \pi)$,*

$$\Pr\big(\angle(z_i(0), -z_i^*) < \alpha\big) = \frac{1}{\sqrt{\pi}} \frac{\Gamma\big(\frac{d_{\text{osc}}}{2}\big)}{\Gamma\big(\frac{d_{\text{osc}}-1}{2}\big)} \int_0^\alpha \sin^{d_{\text{osc}}-2}(\theta)\, d\theta. \tag{10}$$

For every $\alpha \in (0, \pi/2)$, the probability in (10) decays exponentially in d_{osc}. Moreover, for each fixed $d_{\text{osc}} \geq 2$,

$$\Pr\big(\angle(z_i(0), -z_i^*) < \alpha\big) \sim C(d_{\text{osc}})\, \alpha^{d_{\text{osc}}-1} \qquad \text{as } \alpha \to 0,$$

for a constant $C(d_{\text{osc}}) > 0$.

Proof. The probability that a uniformly distributed point on $\mathbb{S}^{d_{\text{osc}}-1}$ falls within angular distance α of a fixed point equals the ratio of the spherical cap area to the total sphere surface area; see, e.g., Lecture 8 of (Ball, 1997). Slicing $\mathbb{S}^{d_{\text{osc}}-1}$ by hyperplanes perpendicular to the axis through the cap's pole, the slice at polar angle θ is a $(d_{\text{osc}} - 2)$-sphere of radius $\sin\theta$. Integrating the slice areas over $\theta \in [0, \alpha]$ gives the cap area

$$\text{Area}\big(\mathbb{S}^{d_{\text{osc}}-2}\big) \int_0^\alpha \sin^{d_{\text{osc}}-2}(\theta)\, d\theta,$$

where $\text{Area}\big(\mathbb{S}^{d_{\text{osc}}-2}\big) = 2\pi^{(d_{\text{osc}}-1)/2} / \Gamma\big((d_{\text{osc}} - 1)/2\big)$ (where Γ is the Euler Gamma function) is the surface area of the $(d_{\text{osc}} - 2)$-dimensional unit sphere. The total surface area of $\mathbb{S}^{d_{\text{osc}}-1}$ is $2\pi^{d_{\text{osc}}/2}/\Gamma(d_{\text{osc}}/2)$. Taking the ratio, we recover (10):

$$\begin{aligned}\Pr\big(\angle(z_i(0), -z_i^*) < \alpha\big) &= \frac{\text{cap area}}{\text{total area}} = \frac{2\pi^{(d_{\text{osc}}-1)/2}/\Gamma((d_{\text{osc}} - 1)/2)}{2\pi^{d_{\text{osc}}/2}/\Gamma(d_{\text{osc}}/2)} \cdot \int_0^\alpha \sin^{d_{\text{osc}}-2}(\theta)\, d\theta \\ &= \frac{1}{\sqrt{\pi}} \frac{\Gamma\big(\frac{d_{\text{osc}}}{2}\big)}{\Gamma\big(\frac{d_{\text{osc}}-1}{2}\big)} \int_0^\alpha \sin^{d_{\text{osc}}-2}(\theta)\, d\theta.\end{aligned}$$

For any $\alpha \in (0, \pi/2)$ and for all $\theta \in [0, \alpha]$, we have $\sin\theta \leq \sin\alpha < 1$. Thus,

$$\int_0^\alpha \sin^{d_{\text{osc}}-2}(\theta)\, d\theta \leq \alpha \sin^{d_{\text{osc}}-2}(\alpha),$$

which decays exponentially in d_{osc}. Stirling's approximation (Abramowitz & Stegun, 1964, 6.1.47) gives $\Gamma(d_{\text{osc}}/2)/\Gamma((d_{\text{osc}}-1)/2) \sim \sqrt{d_{\text{osc}}/2}$ as $d_{\text{osc}} \to \infty$, so the prefactor in (10) grows only as $\sqrt{d_{\text{osc}}}$. Multiplying by the exponential decay of the integral, the probability decays exponentially in d_{osc} for every fixed $\alpha \in (0, \pi/2)$. For small values of α, the expansion $\sin\theta = \theta + O(\theta^3)$ gives

$$\int_0^\alpha \sin^{d_{\text{osc}}-2}(\theta)\, d\theta = \frac{\alpha^{d_{\text{osc}}-1}}{d_{\text{osc}} - 1} + O\big(\alpha^{d_{\text{osc}}+1}\big).$$

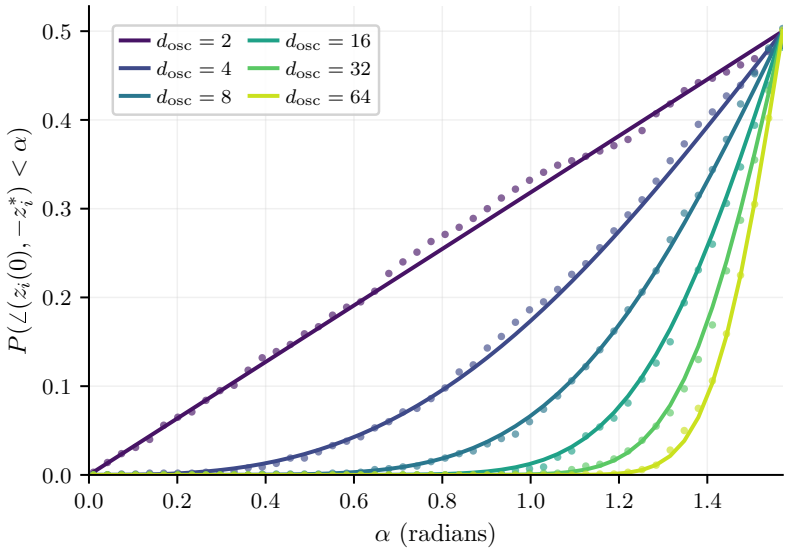

Figure 2: **Validation of Proposition 4.** Markers show the empirical fraction of N=10000 uniform samples on $\mathbb{S}^{d_{\text{osc}}-1}$ falling within angular distance α of a fixed pole, and lines show the closed-form prediction (10). The two agree within sampling noise across all d_{osc} and α, confirming that the probability of initializing $z_i(0)$ near the unstable equilibrium $-z_i^*$ decays sharply with d_{osc}.

The remaining factors — the prefactor $\Gamma(d_{\text{osc}}/2)/(\sqrt{\pi}\,\Gamma((d_{\text{osc}}-1)/2))$ and the $1/(d_{\text{osc}}-1)$ from the leading term — depend only on d_{osc}, so

$$\Pr\big(\angle(z_i(0), -z_i^*) < \alpha\big) \sim C(d_{\text{osc}})\,\alpha^{d_{\text{osc}}-1} \qquad \text{as } \alpha \to 0,$$

where $C(d_{\text{osc}}) > 0$. □

The proposition makes two distinct scaling claims that should not be conflated. The exponential-in-d_{osc} decay holds for every fixed $\alpha \in (0, \pi/2)$ and describes how the unstable hemisphere shrinks as the sphere dimension grows. The small-α asymptotic $\sim C(d_{\text{osc}})\,\alpha^{d_{\text{osc}}-1}$ is a separate statement, holding at each fixed $d_{\text{osc}} \geq 2$, describing the polynomial rate at which the cap probability vanishes as the threshold α tightens.

Proposition 3 and Proposition 4 identify two failure modes for finite-time ODE convergence at low d_{osc}. The implementation reported here trains and infers using the analytic fixed point $z_i^* = h_i/\|h_i\|$ (with a standard numerical safeguard at $\|h_i\| \to 0$; see Appendix B), so these failure modes affect only physical hardware deployment, where they can be reduced by three levers. First, d_{osc} is the principal architectural lever (Table 6): degenerate positions drop from 4.0% at $d_{\text{osc}} = 2$ to 0.03% at $d_{\text{osc}} = 32$, and near-antipodal initializations from 11.7% at $d_{\text{osc}} = 2$ to 0.81% at $d_{\text{osc}} = 32$. Second, extending the integration window helps: at $d_{\text{osc}} = 2$, growing $T_{\max}$ from 30 to 5000 recovers convergence from 84.3% to 98.7% of tokens. Third, stronger coupling accelerates the dynamics: the gradient flow scales with $\|h_i\|$, so increasing coupling magnitudes is equivalent to extending the integration window. This is a substrate design choice for physical oscillator arrays.

4 Experiments

All experiments use the analytic fixed point (4) during training, with gradients through the normalization handled automatically by PyTorch autograd. The softmax baseline uses standard scaled dot-product attention with identical architecture and hyperparameters. Query anchors are initialized uniformly on $\mathbb{S}^{d_{\text{osc}}-1}$ and learned jointly with all other parameters. We use p=1 throughout; ablation results appear in Section 4.3. Unless stated otherwise, all results report mean ± std over 5 seeds.

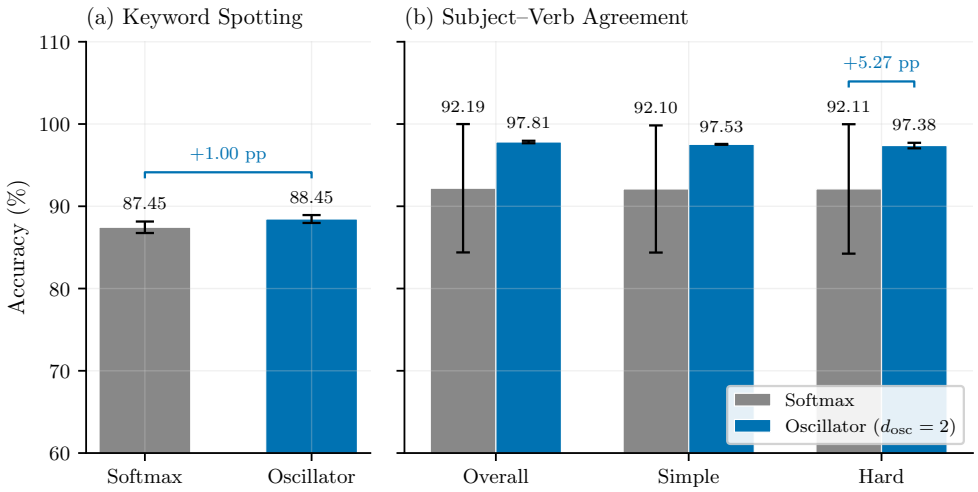

Figure 3: **Bidirectional task accuracy.** (a) KWS: oscillator (d_{osc}=2) outperforms softmax by +1.00 pp. (b) SVA at the minimum-hardware configuration (d_{model}=32, 1 head, 1 layer): 0/5 training failures for oscillator versus 1/5 for softmax (78.14% hard on the failing seed). Mean hard accuracy is 97.38% (oscillator) versus 92.11% (softmax), a +5.27 pp gap driven primarily by the single softmax failure; the four successful softmax seeds match oscillator within 1–2 pp.

4.1 Bidirectional tasks: keyword spotting and subject-verb agreement

On bidirectional tasks, oscillator attention at d_{osc}=2 matches or exceeds softmax. We verify this on two tasks: acoustic classification (keyword spotting) and syntactic agreement (subject-verb agreement). Figure 3 summarizes accuracy on both bidirectional tasks side by side; a detailed analysis and comparison follow.

4.1.1 Keyword spotting

For keyword spotting (KWS), we train transformers with d_{model}=32, 2 heads, 1 layer on 10-class Google Speech Commands (Warden, 2018) (log-mel spectrograms, 40 bins, T=49 frames, 5 seeds). The result is $88.45 \pm 0.48\%$ versus $87.45 \pm 0.70\%$ for softmax, a +1.00 pp accuracy gain with notably tighter variance across seeds (Table 1). This is consistent with the implicit regularization that the bounded cosine readout and sphere constraint impose; the oscillator mechanism is less sensitive to random seed than softmax.

A natural question is whether the oscillator dynamics actually determine the classification, or whether the answer is already encoded in the value vectors regardless of the attention pattern. To test this, we freeze W_V at random initialization and train only the oscillator parameters (anchor vectors and coupling projections), removing any possibility that the value projection learns to compensate for weak attention structure. The frozen-W_V model reaches $88.26 \pm 0.17\%$, only 0.19 pp below the full model: the oscillator dynamics account for essentially all of the classification performance, with the learned value projection adding only a marginal refinement. Two further controls isolate the contribution of the dynamics themselves. Zero attention (uniform weights, mean pooling) achieves 35%, near chance for a 10-class problem, confirming that embedding structure alone is insufficient. Random oscillator positions (free oscillators z_i drawn uniformly on the sphere with no equilibration to the fixed point $h_i/\|h_i\|$) achieve 64.0%, reflecting the contribution of learned token embeddings combined with attention from random query positions.

The +24.26 pp lift from random phases to the analytic fixed point (with frozen W_V) is attributable to the structured attention geometry produced by the oscillator dynamics (learned anchors, coupling weights, and equilibration). The frozen-W_V result also resolves the question of whether the ODE is necessary at inference. With random W_V, classification probability at random oscillator initialization averages 64.0% correct, already above chance but with the wrong class leading on hard examples. After ODE convergence the accuracy reaches 88.26%, with probability trajectories that start at the wrong class and flip to the correct class during

Condition	Val acc (%)	Δ vs softmax
Softmax	87.45 ± 0.70	—
Oscillator d_{osc}=2 (full model)	88.45 ± 0.48	+1.00 pp
Oscillator, frozen W_V	88.26 ± 0.17	+0.81 pp
Random oscillator positions	64.0	−23.5 pp
Zero attention (mean pooling)	35.0	−52.5 pp

Table 1: **KWS results.** 10-class Google Speech Commands, d_{model}=32, 2 heads, 1 layer. Full model and frozen-W_V ablation compared to the softmax baseline; random oscillator positions and zero attention isolate the contribution of the oscillator dynamics.

Condition	Overall (%)	Hard (%)
Softmax	92.19 ± 7.80	92.11 ± 7.87
Oscillator d_{osc}=2 (full)	97.81 ± 0.15	97.38 ± 0.33
Oscillator, frozen W_V	97.73 ± 0.14	97.14 ± 0.46

Table 2: **SVA results at minimum-hardware configuration** (d_{model}=32, 1 head, 1 layer, d_{ff}=64, 14 oscillators per attention layer at d_{osc}=2). The "hard" split contains sentences whose distractor disagrees with the subject. The frozen-W_V ablation holds W_V at random initialization, training only the oscillator parameters. At a larger standard architecture (d_{model}=64, 2 heads, 2 layers), both methods reach 98%.

settling. These observations are consistent with the dimensional bottleneck argument of Section 4.2: on tasks with low effective attention rank, the attention geometry is the binding constraint, not the value dimension.

4.1.2 Subject-verb agreement

We turn next to a task of very different structure: syntactic agreement in natural-language sentences. For subject-verb agreement (SVA), we train on a synthetic Linzen-style dataset (Linzen et al., 2016) of 40K/4K/4K sentences (e.g., "The keys on the table are/is"), with subject, distractor, and verb indices recorded for attention analysis. We use synthetically generated sentences to ensure precise control over subject, distractor, and verb positions; the goal is attention analysis rather than linguistic benchmark performance. At the standard transformer architecture (d_{model}=64, 2 heads, 2 layers), both softmax and oscillator attention reach 98% accuracy. The task is fully within the capacity of standard transformers at this scale.

To examine oscillator attention at minimum hardware scale, we evaluate a constrained configuration: d_{model}=32, single attention head, single transformer layer, feed-forward dimension $d_{\text{ff}} = 64$. This setup uses only 14 total oscillators per attention layer at d_{osc}=2 (7 position–head pairs each contributing one free oscillator and one anchor). At this configuration, oscillator attention has zero training failures across 5 seeds, all reaching 97.38 ± 0.33% on hard distractor sentences (Table 2). Softmax shows 1/5 catastrophic failures: one seed collapsing to 78.14% hard; the remaining four softmax seeds reach 94.19–96.79% hard. Reported as means, this is 97.38 ± 0.33% for oscillator versus 92.11 ± 7.87% for softmax. The +5.27 pp gap is driven primarily by the single softmax failure. Two findings follow: at this constrained capacity, oscillator attention matches softmax on every successful training run (the four good softmax seeds are close to oscillator's 97.38 ± 0.33%); and oscillator exhibits no catastrophic-failure mode at all (0/5 vs 1/5 for softmax), suggesting that the bounded sphere geometry may provide implicit regularization. To verify that the oscillator dynamics — not the value projection — drive this configuration's accuracy, we train the oscillator with W_V frozen at random initialization (third row of Table 2). The frozen-W_V model reaches 97.14 ± 0.46% hard accuracy, indistinguishable from the full model (−0.24 pp). The attention pattern alone, without learned value transformations, suffices. This parallels the KWS finding above: freezing W_V leaves accuracy essentially unchanged on both bidirectional tasks, confirming that oscillator attention dynamics — not learned value transformations — drive classification.

Attention	Split	a_{vs}	a_{vd}	Gap
Softmax	correct	0.59 ± 0.16	0.24 ± 0.15	+0.35
Softmax	wrong	0.23 ± 0.14	0.06 ± 0.06	+0.17
Oscillator	correct	0.20 ± 0.11	0.08 ± 0.04	+0.12
Oscillator	wrong	0.26 ± 0.09	0.07 ± 0.04	+0.19

Table 3: **Verb-to-subject vs. verb-to-distractor attention at the minimum-hardware configuration.** Values are means across all hard distractor sentences (n=2415); ± values are between-seed variation. Both mechanisms maintain positive subject preference on correct predictions. The oscillator gap is similar (slightly larger) on wrong predictions, indicating that classification errors are not driven by attention misallocation. Note: the softmax wrong row pools across seeds; the catastrophic-failure seed (1/5, 78.14% hard accuracy) contributes 252 of the wrong sentences and shows attention collapse; the four successful seeds show error patterns similar to oscillator.

The SVA task also reveals a structural difference in attention geometry. Table 3 shows verb-to-subject versus verb-to-distractor attention weights aggregated across all hard test sentences (n=2415, 5 seeds). The subject-distractor gap is consistently positive for oscillator across seeds and across correct/wrong predictions, showing that the oscillator reliably prioritizes the subject position; the gap is in fact slightly larger on wrong predictions (+0.19) than on correct ones (+0.12). The errors are structural: 150 of 151 wrong predictions (99.34%) involve the words *fish* or *deer* as the subject, number-invariant nouns whose singular and plural forms are identical in English (*one fish, two fish*; *one deer, two deer*). The oscillator correctly attends to the subject, but the subject's embedding carries no morphological information distinguishing singular from plural. The resulting classification is driven to near-zero logit magnitude ($\overline{|\ell|} = 0.92$ on wrong vs. 6.45 on correct), landing on the wrong side of the decision boundary on close calls. Softmax's error structure differs by seed. The four softmax seeds that train successfully ($\geq 95\%$ hard accuracy) show error patterns similar to oscillator: gap +0.23 on wrong predictions, with 77% of these errors having close-call logits ($|\ell| < 0.5$). The catastrophic-failure seed (one of five, 78.14% hard accuracy) shows true attention collapse: gap −0.08 on wrong predictions, attention favoring the distractor over the subject. The pooled drop reported in Table 3 (subject focus +0.35 → +0.17, correct → wrong) is dominated by this seed. When training succeeds, both mechanisms fail on similar intrinsically ambiguous cases; the +5.27 pp accuracy gap at this configuration is driven primarily by softmax's training-instability mode rather than by a difference in error mechanism.

Figure 4 shows verb attention distributions across all hard distractor sentences (n=2415). Softmax distributions are broad due to high between-seed variability, with the catastrophic-failure seed pulling the subject-attention distribution toward lower values. Oscillator distributions are tighter and show a consistent subject preference, with anchor positions spread broadly across the unit circle, indicating the mechanism uses the available representational geometry rather than collapsing to a degenerate fixed point.

4.2 Causal language modeling

We evaluate causal language modeling on two datasets: WikiText-2 (word-level, vocabulary 10K, T=50, a standard LM benchmark) and TinyStories (Eldan & Li, 2023) (∼2M short stories, vocabulary 8K, simpler narrative structure). The same transformer architecture is used throughout (d_{model}=128, 4 heads, 2 layers). Table 4 reports validation perplexity at $d_{\mathrm{osc}} \in \{2, 8, 32\}$; both datasets show the oscillator–softmax gap shrinking as d_{osc} grows. On WikiText-2, the gap decreases from +11.09 PPL at d_{osc}=2 to +5.65 PPL at d_{osc}=8 and +2.98 PPL at d_{osc}=32. On TinyStories, whose simpler narrative structure demands less simultaneous disambiguation, the gap is smaller at every d_{osc} and closes faster: +2.39 PPL at d_{osc}=2, +1.24 PPL at d_{osc}=8, and +0.57 PPL at d_{osc}=32, approaching practical parity. Replacing the analytic fixed point with RK45 ODE integration and finite horizon $T_{\max} = 30$ at inference recovers the same performance within 0.13 PPL (Section 4.4). Figure 5 illustrates the oscillator dynamics on a TinyStories sentence: the active token's trajectory converges toward its fixed point under learned coupling springs anchored to context tokens.

The root cause of the oscillator-versus-softmax PPL gap is a dimensional bottleneck. Both softmax and oscillator attention produce scalar attention weights per token pair; the difference is the dimensionality available

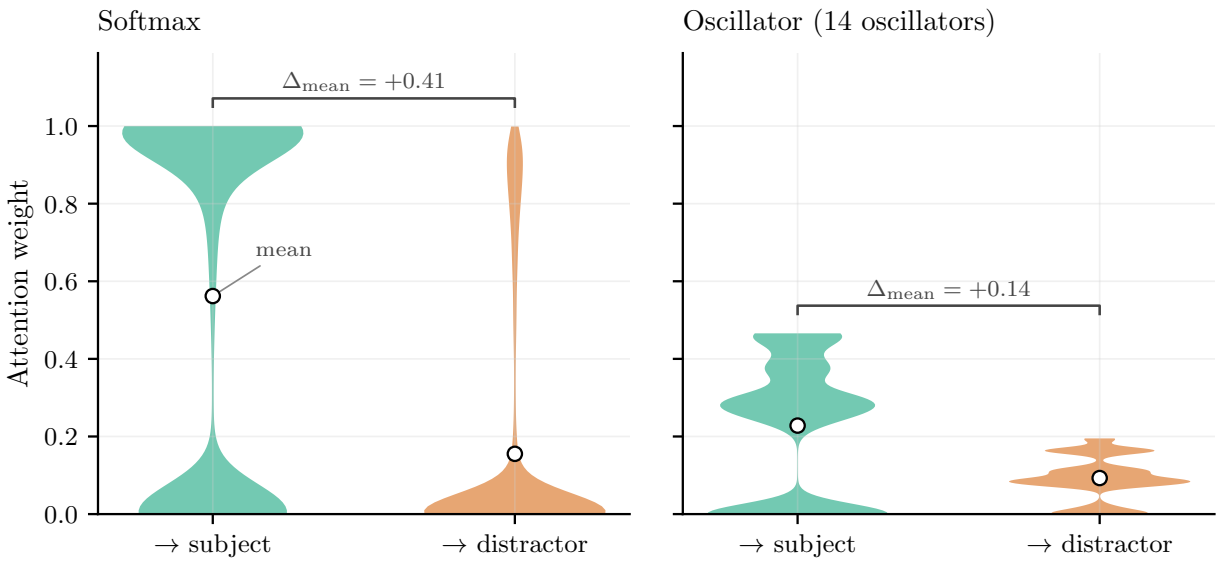

Figure 4: **Verb attention distributions across all hard test sentences (n=2415) at the minimum-hardware configuration, aggregated over 5 seeds.** Softmax shows broad distributions reflecting high seed-level variability (one of five seeds failed to converge above 80%; the displayed distribution mixes stable and unstable runs). Oscillator shows tighter distributions with a consistent positive subject preference ($a_{\text{vs}} > a_{\text{vd}}$), reflecting the implicit regularization of the bounded sphere geometry.

Dataset	d_{osc}	Oscillator PPL	Softmax PPL	Δ
WikiText-2	2	110.67 ± 0.51	99.58	+11.09
WikiText-2	4	107.20 ± 0.55	99.58	+7.62
WikiText-2	8	105.23 ± 0.32	99.58	+5.65
WikiText-2	16	103.78 ± 0.49	99.58	+4.20
WikiText-2	32	102.56 ± 0.37	99.58	+2.98
TinyStories	2	10.93 ± 0.05	8.54	+2.39
TinyStories	8	9.78 ± 0.06	8.54	+1.24
TinyStories	32	9.11 ± 0.02	8.54	+0.57

Table 4: **Causal LM perplexity as a function of the oscillator dimension.** Validation perplexity on WikiText-2 ($d_{\text{osc}} \in \{2, 4, 8, 16, 32\}$) and TinyStories ($d_{\text{osc}} \in \{2, 8, 32\}$), compared to a softmax baseline. All models use the same backbone (d_{model}=128, 4 heads, 2 layers) and analytic fixed-point inference. The oscillator-versus-softmax gap closes monotonically as d_{osc} grows. Mean and standard deviation across 5 seeds.

to express the attention pattern across many pairs. Softmax computes $\alpha_{ij}^{\text{sm}} \propto \exp\big((Fe_i)^\top (Ge_j)/\sqrt{d_h}\big)$, so the attention pattern at position i can vary along all d_h directions of the query-key product. Oscillator attention computes $\alpha_{ij} \propto 1 + z_i^{*\top} r_j$ with the settled oscillator constrained to the anchor span: $z_i^* = h_i / \|h_i\|$ where $h_i = \sum_l w_{il}\, r_l$. The attention pattern at position i therefore lives in the d_{osc}-dimensional manifold spanned by the anchors. When d_{osc} is small, the attention matrix has bounded effective rank: rows $\{\alpha_{ij}\}_j$ are tied to one another through the shared anchor geometry, and rotations of h_i that softmax could distinguish collapse to the same oscillator state. As d_{osc} grows, the anchor span enlarges and the attention rows become more independent. On tasks where one dimension of similarity is sufficient (KWS; SVA at the standard architecture), the dimensional constraint is not binding and d_{osc}=2 achieves parity with softmax. At constrained capacity (SVA minimum-hardware configuration), the sphere constraint provides additional implicit regularization, manifested as zero training failures (versus 1/5 for softmax at the same configuration). On tasks requiring simultaneous tracking of many contextual relationships (causal LM), higher d_{osc} is needed. We verified that

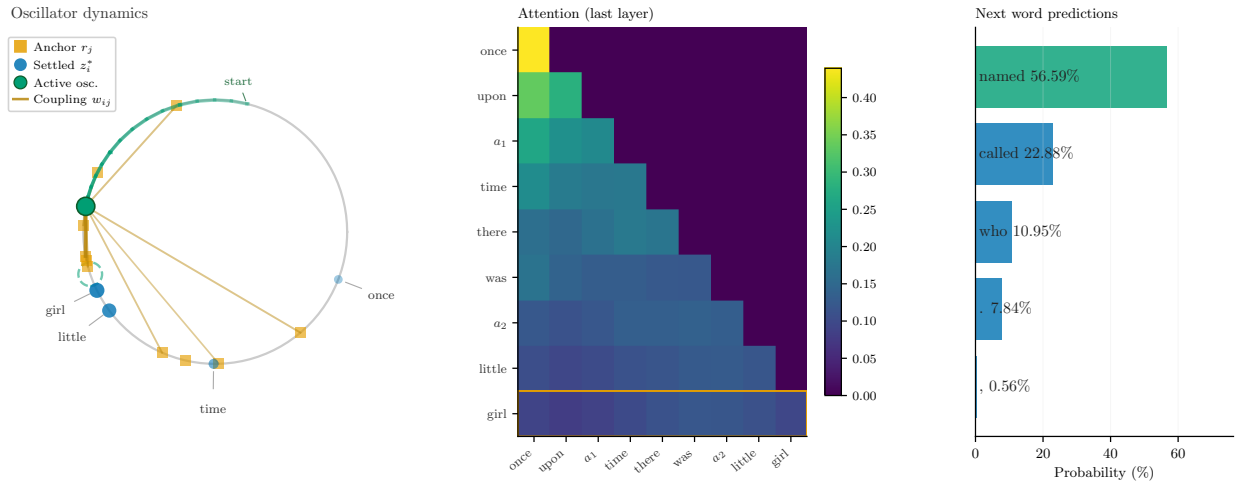

Figure 5: **Oscillator dynamics on a TinyStories sentence.** *Left:* token oscillators on the unit circle; amber squares mark anchor positions r_j, dark markers show settled free-oscillator fixed points z_i^*, the active free oscillator z_i traces a trajectory toward its fixed point $z_i^* = h_i/\|h_i\|$ along the trail, and lines indicate coupling weights w_{ij}. *Top right:* attention heatmap (last layer, head average). *Bottom right:* next-word probability distribution; the prediction "named" receives 56.59% probability.

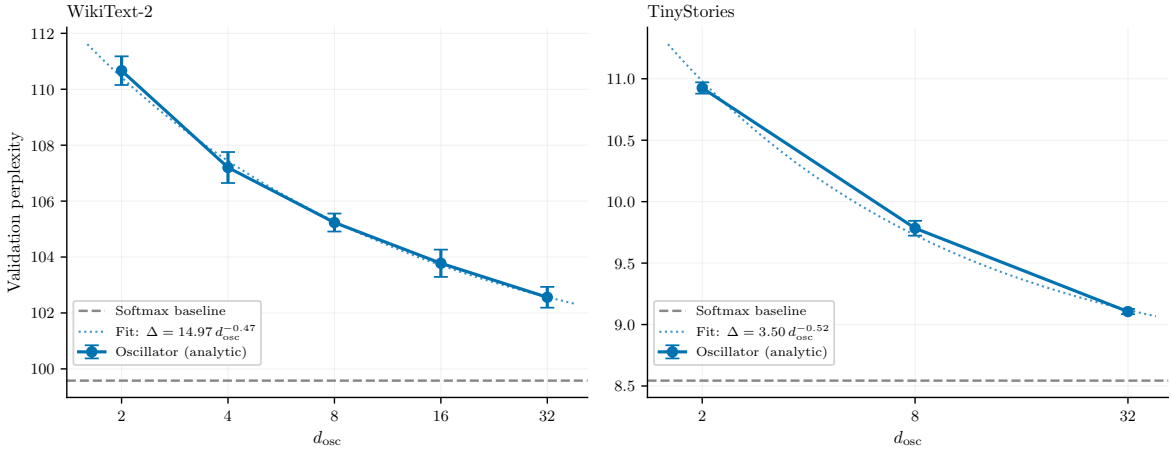

Figure 6: **Perplexity gap vs. oscillator dimension.** Validation perplexity at $d_{\text{osc}} \in \{2, 8, 32\}$ on WikiText-2 (left) and TinyStories (right). Markers: oscillator attention with analytic fixed-point inference; horizontal dashed line: softmax baseline. Error bars indicate seed-to-seed standard deviation. Dashed curves show power-law fits $\Delta \approx C \cdot {d_{\text{osc}}}^{-\alpha}$ with $C \approx 14.97$, $\alpha \approx 0.47$ on WikiText-2 and $C \approx 3.50$, $\alpha \approx 0.52$ on TinyStories.

the gap persists under three alternative interventions — learned position-dependent phase offsets, per-head coupling amplification, and independent scaling of attention heads and d_{model} — none of which close the gap independently of d_{osc} (Appendix C). We conjecture that parity is reached when d_{osc} matches the effective rank of the softmax attention matrix on the task, a question that we leave as the subject of future investigation.

Figure 6 shows that the perplexity gap on both datasets follows a power-law decay with remarkably similar exponents: $\Delta \approx 14.97 \cdot {d_{\text{osc}}}^{-0.47}$ on WikiText-2 and $\Delta \approx 3.50 \cdot {d_{\text{osc}}}^{-0.52}$ on TinyStories, same shape, different scale, with the simpler dataset's smaller prefactor reflecting its lower attention complexity. The power-law shape is consistent across both datasets, suggesting a general phenomenon. The WikiText-2 fit uses five values of d_{osc} ($\{2, 4, 8, 16, 32\}$); TinyStories uses three ($\{2, 8, 32\}$). Residuals on the WikiText-2 points are small and unbiased (max log-residual 0.030, no systematic deviation), consistent with the power-law form. We conjecture

Task / d_{osc}	p=1	p=2	p=4
KWS (d_{osc}=2, acc %)	88.45 ± 0.48	89.61 ± 0.20	89.28 ± 0.82
TinyStories (d_{osc}=2, PPL)	10.86 ± 0.04	10.40 ± 0.04	10.16 ± 0.11
TinyStories (d_{osc}=8, PPL)	9.78 ± 0.06	9.18 ± 0.02	8.98 ± 0.07

Table 5: **Readout sharpening.** First row is KWS accuracy (%) at d_{osc}=2; softmax baseline is 87.45 ± 0.70%. Remaining rows are TinyStories validation PPL at d_{osc}=2 and d_{osc}=8. Sharpening improves performance over p=1 in all settings; p=2 is best on KWS, p=4 is best on TinyStories.

d_{osc}	Frac < 0.01	App. antipodal	Degenerate	nfev/tok
2	84.3%	11.7%	4.0%	632
8	94.9%	3.4%	0.55%	485
32	98.3%	0.81%	0.03%	526

Table 6: **ODE convergence (RK45, $T_{\max} = 30$, 5 random initializations per token).** Convergence improves with d_{osc}: both apparent-antipodal failures (initialization near the unstable equilibrium) and degenerate-position failures ($\|h_i\|$ small) become rare, as predicted by Propositions 3 and 4.

that the perplexity gap obeys a scaling law of the form $\Delta(d_{\text{osc}}) \approx C\, {d_{\text{osc}}}^{-\alpha}$ for task-dependent constants $C, \alpha > 0$. A precise theoretical bridge between the two mechanisms is hindered by the fact that softmax and oscillator attention compute attention weights through structurally different operations (exponentiated similarity versus shifted cosine on the sphere), so the apparent convergence in our data is an empirical observation, not a theoretical guarantee.

4.3 Readout sharpening

The readout exponent p plays a role analogous to inverse temperature in softmax: larger p sharpens the attention distribution toward the strongest cosine matches, while smaller p flattens it toward a uniform distribution. As a software-side optimization, p consistently improves performance across all tasks tested (Table 5). On KWS at d_{osc}=2, raising p from 1 to 2 gives 89.61 ± 0.20% versus 88.45 ± 0.48% at p=1, a +1.16 pp gain from sharpening (+2.16 pp over softmax baseline 87.45 ± 0.70%). On causal language modeling, raising p from 1 to 4 improves perplexity by approximately 0.75 at both d_{osc}=2 and d_{osc}=8. Both directions agree: a modestly sharpened readout improves task accuracy by suppressing weak background alignments that would otherwise dilute the attention signal. Both mechanisms ultimately approach a hard-attention limit as their sharpening parameter grows: the readout exponent for oscillator attention, the inverse temperature for softmax. These results confirm sharpening as a deployment-mode choice: p=1 for substrate-native attention, p>1 when digital post-processing is available.

4.4 ODE convergence verification

Theorem 2 guarantees convergence in continuous time; here we verify it numerically in finite time and characterize its dependence on d_{osc}. We run the Lohe ODE using an adaptive RK45 integrator (`scipy.solve_ivp`, $rtol = 10^{-6}$, $T_{\max} = 30$) on the trained TinyStories models and measure convergence error $\|z^*_{\text{ODE}} - z^*_i\|_2$ against the analytic fixed point. Convergence improves strongly with d_{osc} (Table 6), consistent with Proposition 4. The "apparent antipodal" classification in Table 6 is a finite-time threshold determined by $T_{\max}$=30 (tokens that have not converged within the integration window), not the asymptotic $\alpha = \pi/2$ event of Proposition 4; the empirical count is the more conservative quantity. Residual failures at each d_{osc} reflect initializations near the unstable equilibrium where convergence within the fixed budget $T_{\max}$=30 is insufficient, not convergence to a wrong attractor. Panel (a) of Figure 7 confirms that all cases eventually converge given sufficient integration time: extending $T_{\max}$ from 30 to 5000 for d_{osc}=2 recovers convergence monotonically (84.3% → 91.3% → 96.0% → 97.3% → 98.7%). Sequential initialization reduces mean error 4.6× at d_{osc}=2 at

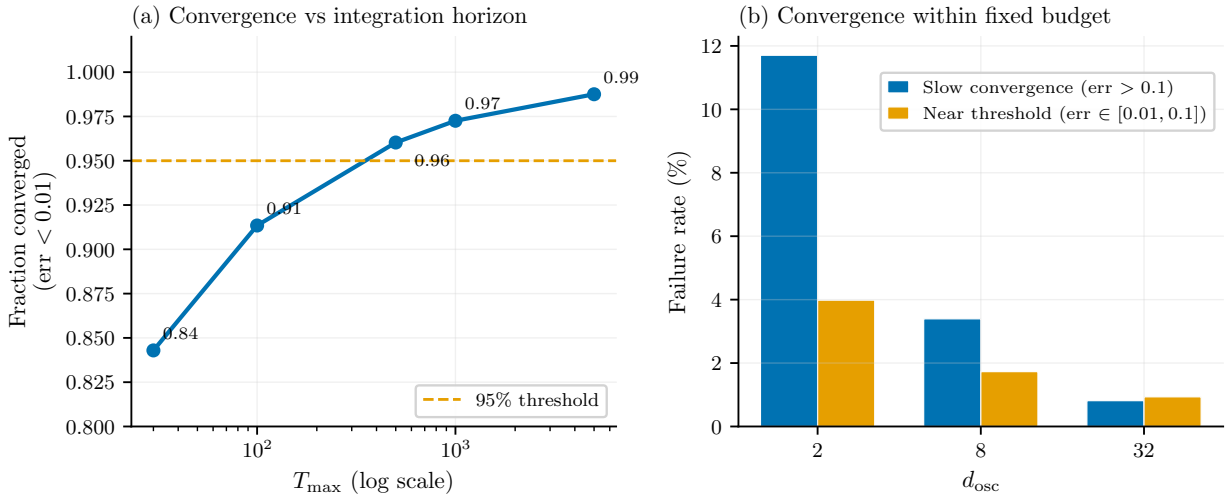

Figure 7: **ODE convergence verification.** (a) At d_{osc}=2, the fraction of tokens converging (err < 0.01) grows with integration horizon $T_{\max}$, reaching 98.7% at $T_{\max}$=5000; all tokens eventually converge given sufficient time, consistent with Theorem 2. (b) At fixed budget $T_{\max}$=30, convergence failure rates decrease strongly with d_{osc}: both apparent-antipodal failures (slow escape from the unstable equilibrium) and degenerate-position failures ($\|h_i\|$ small) become exponentially rare, consistent with Propositions 3 and 4.

no additional cost; at d_{osc}=32 the advantage is negligible because random initialization already avoids the slow-convergence region with high probability.

Beyond pointwise convergence, RK45 ODE integration on the trained WikiText-2 d_{osc}=2 model (finite horizon $T_{\max} = 30$, 5 random initialization seeds) recovers the analytic fixed-point perplexity within 0.13 PPL. Sequential initialization (starting each token's free oscillator from the previous token's fixed point) closes the residual gap to 0.03 PPL. The strong driving forces produced by the trained coupling weights ensure near-perfect convergence within $T_{\max}$=30. Physical hardware running continuous-time dynamics could close the gap further; the equilibration trajectory provably converges to the analytic fixed point.

5 Discussion

The experiments reveal a coherent picture. On bidirectional tasks where the attention structure is simple — grouping voiced frames in KWS, identifying a single grammatical dependency in SVA — scalar Kuramoto dynamics at d_{osc}=2 are sufficient and achieve parity with softmax. The sphere constraint and bounded cosine readout provide beneficial implicit regularization, visible as tighter variance across seeds and more geometrically structured attention maps. On causal language modeling, where each token must simultaneously track many independent contextual relationships, the d_{osc}-dimensional anchor subspace constrains the rank of the attention pattern; higher d_{osc} relaxes this constraint, and the gap closes following the power law $\Delta \sim 14.97 \cdot {d_{\text{osc}}}^{-0.47}$. These results should be read as evidence of viability, not competition. Existing efforts to reduce the cost of attention on digital hardware — linear attention (Katharopoulos et al., 2020), Performer (Choromanski et al., 2021), Linformer (Wang et al., 2020), sparse and local attention variants (Child et al., 2019; Beltagy et al., 2020), and relative position encodings such as RoPE (Su et al., 2024) and ALiBi (Press et al., 2022) — reduce asymptotic cost or impose inductive biases but remain within the softmax framework, addressing sequence length scaling rather than the energy cost of the attention operation itself. The von Neumann memory hierarchy remains the dominant energy sink at the small sequence lengths characteristic of edge devices, where their asymptotic advantages have not yet materialized. Our mechanism addresses a different axis: the goal is to establish that a physically realizable attention mechanism can achieve accuracy within a predictable and controllable range of softmax. The d_{osc} scaling law provides the control: a system designer can choose d_{osc} to match the accuracy requirements of the task. The mechanism is also orthogonal to algorithmic choices:

sparsity patterns and positional encodings can be combined with oscillator attention, since they modify which weights w_{ij} are nonzero rather than how attention weights are computed from token embeddings.

The hardware implications are concrete. At d_{osc}=2, the mechanism requires T scalar oscillators per head: 98 total for the KWS setting (2 heads, T=49). At d_{osc}=32, approximately 6,200 oscillators are needed for the language modeling setting. These phase-oscillator counts assume substrates with native d_{osc}-dimensional order parameters. At d_{osc}=2, hardware realizations are well-developed across multiple substrates: mechanical oscillators, Josephson junctions, MEMS resonators, and electrical LC tanks. For $d_{\text{osc}} > 2$, oscillator hardware is less mature: some substrates with intrinsic multi-dimensional order (e.g., vector laser modes, three-dimensional spin systems) appear in principle compatible with Lohe dynamics on $\mathbb{S}^{d_{\text{osc}}-1}$, but established hardware demonstrations at higher d_{osc} remain open. The d_{osc}=2 case is therefore the immediate hardware-deployment target; higher d_{osc} is realized in software in this work and is an open hardware question. Across these substrates, energy per oscillator is bounded by device physics, so the energy advantage scales with the oscillator count. The scaling law $\Delta(d_{\text{osc}}) \sim 14.97 \cdot {d_{\text{osc}}}^{-0.47}$ provides a task-complexity-to-oscillator-budget design rule for whichever substrate is used. Other physical-computing approaches for attention exist but are tied to specific substrates: spiking transformers (Zhu et al., 2023; Lv et al., 2023; Yao et al., 2023) achieve 2–8× energy reduction but retain the switching costs of transistor circuits; photonic transformer accelerators (Zhu et al., 2024) use optical interference for dynamic matrix products in attention but still require digital softmax and optical-to-electrical conversion between layers; memristive accelerators implement in-memory matrix multiplications but incur ADC overhead at every layer. To our knowledge, no existing physical-computing approach for attention is substrate-independent.

The dynamical-systems lineage of the mechanism deserves comment. The Lohe model (Lohe, 2009) generalizes Kuramoto to $\mathbb{S}^{d_{\text{osc}}-1}$; its synchronization has been studied for symmetric all-to-all coupling (see e.g. (Chandra et al., 2019)). When the dynamics admit a gradient-flow structure with a Lyapunov function, multi-agent consensus on $\mathbb{S}^n$ ($n \geq 2$) achieves almost global convergence on any connected graph (Olfati-Saber, 2006; Markdahl et al., 2018; Lipton et al., 2021), a property that fails on the circle ($n = 1$) and is generally lost for non-gradient dynamics (heterogeneous frequencies, asymmetric or matrix-weighted couplings), where multistability is the norm. The fixed-query asymmetric setting (where anchors are external forcing terms rather than free participants) inherits the gradient-flow structure in its simplest form: a single oscillator under linear forcing h_i, with Lyapunov function $V(z_i) = -z_i^\top h_i$ on $\mathbb{S}^{d_{\text{osc}}-1}$. This admits a unique global minimum on the sphere, yielding the uniqueness guarantee of Theorem 2. Earlier work has used oscillator networks for pattern recognition via phase-locking in PLL circuits (Hoppensteadt & Izhikevich, 2000) and has analyzed the broader control-theoretic structure of coupled-oscillator networks (Menara et al., 2022; Qin et al., 2023), but not as a transformer attention primitive. Kuramoto oscillators have also been applied to associative memory (Ogranovich et al., 2026), combinatorial optimization on CDW arrays (Brown et al., 2025), and as a modeling primitive for neural dynamics (Miyato et al., 2025).

Beyond engineered hardware, the framework connects to theories of biological neural computation. The binding-by-synchrony hypothesis (Breakspear et al., 2010; Fries, 2005) proposes that cortical oscillations coordinate distributed representations through phase alignment, with stable phase relationships encoding which features belong together. The fixed-query mechanism is a formalization of this idea: query anchors play the role of sustained reference oscillations (analogous to gamma-band templates in cortical attention), and free oscillators play the role of stimulus-driven populations that phase-lock selectively to those references. The positive-coupling constraint that guarantees convergence parallels the predominantly excitatory long-range projections in cortex. Whether this analogy can be made precise enough to inform biologically plausible learning architectures (in which the coupling weights w_{ij} are updated by a Hebbian or spike-timing-dependent plasticity rule rather than backpropagation) is an open question we regard as promising.

Within the deep-learning literature, the most directly comparable framework is the modern Hopfield network (Hopfield, 1982; Ramsauer et al., 2021): both mechanisms produce attention as fixed-point retrieval from a continuous energy-based system over a finite set of reference patterns. The two differ in both the structure of the energy landscape and the form of retrieval. Modern Hopfield retrieval is an analytic one-step update (a softmax over similarities to stored patterns), corresponding to a free-energy landscape that admits multiple attractors (one per stored pattern), with retrieval converging to the basin containing the query. Oscillator attention's energy landscape is sharper: under strictly positive coupling weights, Theorem 2 guarantees that

the fixed-query gradient flow on $\mathbb{S}^{d_{\text{osc}}-1}$ has a single stable attractor (with a single unstable equilibrium on a measure-zero set), so retrieval converges to a unique fixed point regardless of initialization. The retrieval itself is realized by physical continuous-time dynamics rather than a discrete software step: trajectories trace an explicit gradient flow whose convergence rate is set by the substrate's physical time constant, not by a floating-point computation. The combination (energy landscape collapsed to a single attractor, retrieval realized as a physical gradient flow) is what makes oscillator attention substrate-realizable while preserving the fixed-point-retrieval semantics the Hopfield literature established as a useful frame for attention. Fixed query anchors are also related to the prototype or codebook attention of Set Transformers (Lee et al., 2019), where a fixed set of learned seed vectors attends to the input. The key difference is that in Set Transformers the seeds are free parameters updated by standard softmax attention, while our anchors are fixed during inference and the free oscillators converge to them via physical dynamics, with convergence guaranteed by Theorem 2. Taken together, this positions oscillator attention as a candidate primitive for physical computing: a mechanism with clean theoretical guarantees, predictable scaling behavior, and a path to hardware realization that does not depend on approximating digital arithmetic in analog devices.

6 Conclusion

Fixed-query oscillator attention replaces the exponentiation and in-loop reduction of softmax with the physical equilibration of a coupled oscillator network on $\mathbb{S}^{d_{\text{osc}}-1}$. The fixed point is provably unique and globally attractive from almost every initial condition whenever the weighted anchor sum is nonzero, with degenerate cases vanishing exponentially in d_{osc}. At d_{osc}=2, the mechanism outperforms softmax on KWS (+1.00 pp) and on SVA at the minimum-hardware configuration (+5.27 pp on hard sentences and no training failures). On causal language modeling, the gap closes as $\Delta(d_{\text{osc}}) \sim {d_{\text{osc}}}^{-0.47}$, reaching +2.98 PPL at d_{osc}=32.

The broader claim is substrate independence. The Kuramoto synchronization phenomenon arises naturally in electrical circuits, mechanical oscillators, superconducting junctions, quantum materials, and neural tissue. Wherever it arises, it can compute attention, not because of engineering effort to approximate a digital algorithm in analog hardware, but because the mathematics of equilibration and the mathematics of attention are the same mathematics. This is what we mean by physical intelligence: a computation that is a property of a dynamical class rather than a specific device. Softmax remains the right choice for digital hardware. For any physical substrate where Kuramoto-Lohe dynamics can be realized, oscillator attention provides a principled alternative with theoretical guarantees and characterized scaling behavior. The specific energy and fabrication tradeoffs across different substrates are subjects for further work.

References

Milton Abramowitz and Irene A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, 1964.

Radosław Adamczak. A note on the Hanson–Wright inequality for random vectors with dependencies. *Electronic Communications in Probability*, 20:Paper No. 72, 13, 2015.

Keith Ball. An elementary introduction to modern convex geometry. In Silvio Levy (ed.), *Flavors of Geometry*, volume 31 of *Mathematical Sciences Research Institute Publications*, pp. 1–58. Cambridge University Press, 1997.

I. Beltagy, M. E. Peters, and A. Cohan. Longformer: The long-document transformer. *arXiv:2004.05150*, 2020.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

M. Breakspear, S. Heitmann, and A. Daffertshofer. Generative models of cortical oscillations: Neurobiological implications of the Kuramoto model. *Frontiers in Human Neuroscience*, 4:190, 2010. doi: 10.3389/fnhum.2010.00190.

J. O. Brown, T. Guo, F. Pasqualetti, and A. A. Balandin. Charge-density-wave oscillator networks for solving combinatorial optimization problems. *Physical Review Applied*, 24:024040, 2025. doi: 10.1103/zmlj-6nn7.

S. Chandra, M. Girvan, and E. Ott. Continuous versus discontinuous transitions in the D-dimensional generalized Kuramoto model: Odd D is different. *Physical Review X*, 9(1):011002, 2019. doi: 10.1103/PhysRevX.9.011002.

R. Child, S. Gray, A. Radford, and I. Sutskever. Generating long sequences with sparse transformers. *arXiv:1904.10509*, 2019.

K. Choromanski, V. Likhosherstov, D. Dohan, X. Song, A. Gane, T. Sarlos, P. Hawkins, J. Davis, A. Mohiuddin, L. Kaiser, D. Belanger, L. Colwell, and A. Weller. Rethinking attention with Performers. In *International Conference on Learning Representations*, 2021.

R. Eldan and Y. Li. TinyStories: How small can language models be and still speak coherent English? *arXiv:2305.07759*, 2023.

Andreas K. Engel, Pascal Fries, and Wolf Singer. Dynamic predictions: oscillations and synchrony in top-down processing. *Nature Reviews Neuroscience*, 2:704–716, 2001.

P. Fries. A mechanism for cognitive dynamics: Neuronal communication through neuronal coherence. *Trends in Cognitive Sciences*, 9(10):474–480, 2005. doi: 10.1016/j.tics.2005.08.011.

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

F. C. Hoppensteadt and E. M. Izhikevich. Pattern recognition via synchronization in phase-locked loop neural networks. *IEEE Transactions on Neural Networks*, 11(3):734–738, 2000. doi: 10.1109/72.846744.

A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, pp. 5156–5165, 2020.

H. K. Khalil. *Nonlinear Systems*. Prentice Hall, 2002. ISBN 0130673897.

Y. Kuramoto. Self-entrainment of a population of coupled non-linear oscillators. In H. Araki (ed.), *Int. Symposium on Mathematical Problems in Theoretical Physics*, volume 39 of *Lecture Notes in Physics*, pp. 420–422. Springer, 1975. ISBN 978-3-540-07174-7.

J. Lee, Y. Lee, J. Kim, A. Kosiorek, S. Choi, and Y. W. Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International Conference on Machine Learning*, pp. 3744–3753, 2019.

T. Linzen, E. Dupoux, and Y. Goldberg. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535, 2016. doi: 10.1162/tacl_a_00115.

M. Lipton, R. Mirollo, and S. H. Strogatz. The Kuramoto model on a sphere: Explaining its low-dimensional dynamics with group theory and hyperbolic geometry. *Chaos*, 31(9):093113, 2021.

M. A. Lohe. Non-Abelian Kuramoto models and synchronization. *Journal of Physics A: Mathematical and Theoretical*, 42(39):395101, 2009. doi: 10.1088/1751-8113/42/39/395101.

C. Lv, T. Li, J. Xu, C. Gu, Z. Ling, C. Zhang, X. Zheng, and X. Huang. SpikeBERT: A language Spikformer trained with two-stage knowledge distillation from BERT. *arXiv:2308.15122*, 2023.

J. Markdahl, J. Thunberg, and J. Gonçalves. Almost global consensus on the n-sphere. *IEEE Transactions on Automatic Control*, 63(6):1664–1675, 2018.

T. Menara, G. Baggio, D. S. Bassett, and F. Pasqualetti. Functional control of oscillator networks. *Nature Communications*, 13:4721, 2022. doi: 10.1038/s41467-022-31733-2.

T. Miyato, S. Löwe, A. Geiger, and M. Welling. Artificial Kuramoto oscillatory neurons. In *International Conference on Learning Representations*, 2025.

A. Ogranovich, T. Guo, A. R. Venkatakrishnan, M. R. Shapiro, F. Bullo, and F. Pasqualetti. Oscillator-based associative memory with exponential capacity:
theory, algorithms, and hardware implementation. *IEEE Transactions on Control of Network Systems*, 2026. Submitted.

R. Olfati-Saber. Swarms on sphere: A programmable swarm with synchronous behaviors like oscillator networks. In *IEEE Conf. on Decision and Control*, pp. 5060–5066, 2006.

O. Press, N. A. Smith, and M. Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*, 2022.

Y. Qin, A. M. Nobili, D. S. Bassett, and F. Pasqualetti. Vibrational stabilization of cluster synchronization in oscillator networks. *IEEE Open Journal of Control Systems*, 2:439–453, 2023. doi: 10.1109/OJCSYS.2023.3331195.

H. Ramsauer, B. Schäfl, J. Lehner, P. Seidl, M. Widrich, T. Adler, L. Gruber, M. Holzleitner, D. Kreil, M. Kopp, G. Klambauer, J. Brandstetter, and S. Hochreiter. Hopfield networks is all you need. In *International Conference on Learning Representations*, 2021.

Wolf Singer and Charles M. Gray. Visual feature integration and the temporal correlation hypothesis. *Annual Review of Neuroscience*, 18:555–586, 1995.

S. H. Strogatz. From Kuramoto to Crawford: Exploring the onset of synchronization in populations of coupled oscillators. *Physica D: Nonlinear Phenomena*, 143(1):1–20, 2000.

J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063.

Y. Tay, M. Dehghani, D. Bahri, and D. Metzler. Efficient transformers: A survey. *ACM Computing Surveys*, 55(6):1–28, 2022. doi: 10.1145/3530811.

A. Todri-Sanial, S. Carapezzi, C. Delacour, M. Abernot, T. Gil, E. Corti, S. F. Karg, J. Núñez, M. Jiménez, M. J. Avedillo, and B. Linares-Barranco. How frequency injection locking can train oscillatory neural networks to compute in phase. *IEEE Transactions on Neural Networks and Learning Systems*, 33(5): 1996–2009, 2022. doi: 10.1109/TNNLS.2021.3107771.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

S. Wang, B. Z. Li, M. Khabsa, H. Fang, and H. Ma. Linformer: Self-attention with linear complexity. *arXiv:2006.04768*, 2020.

P. Warden. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv:1804.03209*, 2018.

K. Wiesenfeld, P. Colet, and S. H. Strogatz. Frequency locking in Josephson arrays: Connection with the Kuramoto model. *Physical Review E*, 57(2):1563–1569, 1998.

M. Yao, J. Hu, Z. Zhou, L. Yuan, Y. Tian, B. Xu, and G. Li. Spike-driven Transformer. In *Advances in Neural Information Processing Systems*, 2023.

H. Zhu, J. Gu, H. Wang, Z. Jiang, Z. Zhang, R. Tang, C. Feng, S. Han, R. T. Chen, and D. Z. Pan. Lightening-transformer: A dynamically-operated optically-interconnected photonic transformer accelerator. In *2024 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pp. 686–703. IEEE, 2024.

R.-J. Zhu, Q. Zhao, and J. K. Eshraghian. SpikeGPT: Generative pre-trained language model with spiking neural networks. *arXiv:2302.13939*, 2023.

A Derivation of the Kuramoto equation from the Lohe model at $d_{\text{osc}} = 2$

For completeness, we show how the scalar Kuramoto equation arises as the $d_{\text{osc}} = 2$ case of the Lohe model (1). The derivation is short but its content is worth flagging: a single phase angle $\theta_i \in [0, 2\pi)$ fully captures the dynamics of a unit vector $x_i \in \mathbb{S}^1$, because all the tangent motion projects onto one direction.

Let $d_{\text{osc}} = 2$ and parametrize each oscillator by its phase angle:

$$x_i = (\cos\theta_i,\, \sin\theta_i) \in \mathbb{S}^1, \qquad \theta_i \in [0, 2\pi).$$

The tangent space to $\mathbb{S}^1$ at x_i is one-dimensional and spanned by

$$\tau_i := (-\sin\theta_i,\, \cos\theta_i) = \frac{\partial x_i}{\partial \theta_i}.$$

By the chain rule, the time derivative of x_i is

$$\dot{x}_i = \frac{\partial x_i}{\partial \theta_i}\,\frac{\mathrm{d}\theta_i}{\mathrm{d}t} = \tau_i\,\dot{\theta}_i,$$

which lies along the tangent direction. The projector $(I - x_i x_i^\top)$ restricts any vector to the line spanned by τ_i:

$$(I - x_i x_i^\top)\, v = (\tau_i^\top v)\,\tau_i \qquad \text{for any } v \in \mathbb{R}^2.$$

Apply this to the right-hand side of (1) with $\Omega_i = \omega_i J$ where $J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ is the 90° rotation. The first term satisfies $\Omega_i x_i = \omega_i \tau_i$. The second term projects the coupling onto the tangent:

$$(I - x_i x_i^\top) \sum_j w_{ij} x_j = \sum_j w_{ij}\,(\tau_i^\top x_j)\,\tau_i.$$

Using $\tau_i^\top x_j = -\sin\theta_i \cos\theta_j + \cos\theta_i \sin\theta_j = \sin(\theta_j - \theta_i)$, the full Lohe equation (1) becomes

$$\dot{\theta}_i\,\tau_i = \omega_i\,\tau_i + \Big(\sum_j w_{ij} \sin(\theta_j - \theta_i)\Big)\,\tau_i.$$

Both sides are scalar multiples of τ_i, so the equation reduces to its scalar coefficient:

$$\dot{\theta}_i = \omega_i + \sum_j w_{ij} \sin(\theta_j - \theta_i),$$

which is the classical Kuramoto model (Kuramoto, 1975; Strogatz, 2000). On $\mathbb{S}^1$, the projector $(I - x_i x_i^\top)$ restricts motion to the tangent line at x_i, which is one-dimensional; the dynamics therefore reduce to a single scalar ODE for the phase angle θ_i. The same reduction does not occur in higher dimensions: for $d_{\text{osc}} \geq 3$ the tangent space to $\mathbb{S}^{d_{\text{osc}}-1}$ at x_i is $(d_{\text{osc}} - 1)$-dimensional, and no scalar parametrization captures the full state.

B Experimental details

This appendix provides the hyperparameters, training procedures, and infrastructure needed to reproduce all experiments. Code is available at the URL listed in Appendix D.

Training and inference modes. All models are trained with the analytic fixed point $z_i^* = h_i / \|h_i\|$ as the inference computation, with gradients flowing through the normalization via PyTorch autograd. The normalization is implemented with PyTorch's `F.normalize(h, dim=-1, eps=`10^{-8}`)`, which lower-bounds $\|h_i\|$ by $\varepsilon = 10^{-8}$ in the denominator to avoid the numerical singularity at $\|h_i\| = 0$; on the measure-zero degenerate set bounded by Proposition 3, the resulting attention is approximately uniform. The analytic fixed point and the equilibration trajectory of the ODE (3) converge to the same point under the hypotheses of

d_{ff}	Softmax hard (%)	Oscillator hard (%)	Δ
72	91.07 ± 10.03	96.76 ± 0.46	+5.69 pp
64	92.11 ± 7.87	97.38 ± 0.33	+5.27 pp
56	83.64 ± 9.69	92.42 ± 9.85	+8.78 pp
32	92.28 ± 11.17	94.41 ± 4.40	+2.13 pp
16	95.32 ± 2.14	92.23 ± 10.65	−3.09 pp

Table 7: **SVA hard-sentence accuracy across d_{ff} values** ($d_{\text{model}} = 32$, $n_h = 1$, $n_\ell = 1$, $d_{\text{osc}} = 2$, 5 seeds each). Canonical result at $d_{\text{ff}} = 64$ (Table 2) shown for reference. At $d_{\text{ff}} = 72$ the oscillator is stable while softmax is bimodal; at $d_{\text{ff}} \leq 56$ both models show training instability within the 20-epoch horizon.

Theorem 2, so no algorithmic gap is introduced between training and hardware inference. ODE inference at test time (the RK45 verification reported at the end of Section 4.4) uses `scipy.integrate.solve_ivp` with `method='RK45'`, $rtol = atol = 10^{-6}$, and $T_{\max} = 30$. The KWS frozen-W_V ablation and SVA frozen-W_V ablation (Tables 1 and 2) use the same analytic fixed-point inference ($z_i^* = h_i/\|h_i\|$) as all other experiments.

Antipodal validation (Figure 2). For each $d_{\text{osc}} \in \{2, 4, 8, 16, 32, 64\}$, $N = 10000$ samples are drawn uniformly from $\mathbb{S}^{d_{\text{osc}}-1}$ by sampling $z \sim \mathcal{N}(0, I_{d_{\text{osc}}})$ and normalizing $z/\|z\|$. The empirical fraction within angular distance α of a fixed pole is computed for α on a uniform grid of 50 points in $[0.01, \pi/2]$. The closed-form prediction is computed via `scipy.integrate.quad` on the integrand $\sin^{d_{\text{osc}}-2}(\theta)$ on $[0, \alpha]$, multiplied by the prefactor $\Gamma(d_{\text{osc}}/2)/(\sqrt{\pi}\,\Gamma((d_{\text{osc}}-1)/2))$ from (10).

KWS (Tables 1, 5). $d_{\text{model}} = 32$, $n_h = 2$, $n_\ell = 1$, $d_{\text{ff}} = 128$. Input is log-mel spectrograms with 40 bins, 25 ms windows with 10 ms hop, $T = 49$ frames per utterance. No positional encoding; the spectrogram's temporal structure is encoded implicitly through the coupling weights. Training: AdamW optimizer with weight decay 10^{-4}, $lr = 10^{-3}$, batch 64, 30 epochs, cosine learning-rate schedule, gradient clipping at 1.0. The full-model, softmax-baseline, frozen-W_V, and p-ablation runs all use 5 seeds.

Readout sharpening ablation (Table 5): same architecture as the main KWS experiment (d_{model}=32, n_h=2, n_ℓ=1), d_{osc}=2 throughout. Each value of $p \in \{1, 2, 4\}$ is trained with 5 seeds. Identical training setup to the main KWS runs except for the readout exponent.

SVA (Tables 2, 3). Synthetic Linzen-style sentences (Linzen et al., 2016) of the form "*The keys on the table are/is*" with 40K training, 4K validation, and 4K test examples; subject, distractor, and verb positions are recorded for attention analysis. Sinusoidal positional encoding (Vaswani et al., 2017) added to token embeddings before the first attention layer. Training: AdamW with weight decay 10^{-4}, $lr = 5 \times 10^{-4}$, batch 64, 20 epochs.

Standard architecture: $d_{\text{model}} = 64$, $n_h = 2$, $n_\ell = 2$, $d_{\text{ff}} = 256$. Both softmax and oscillator reach 98% overall (5 seeds each). At this scale the task is fully within transformer capacity and the comparison does not stress the attention mechanism.

Minimum-hardware configuration (Table 2): $d_{\text{model}} = 32$, $n_h = 1$, $n_\ell = 1$, $d_{\text{ff}} = 64$. This is the smallest configuration tested that reveals a statistically meaningful difference. Both softmax and oscillator use 5 seeds; the frozen-W_V ablation uses 5 seeds. Total oscillators per attention layer: $n_h \times T \times d_{\text{osc}} = 1 \times 7 \times 2 = 14$ for typical hard sentences (T=7 tokens).

Architectural robustness sweep: To verify that the minimum-hardware result is not fragile to the specific d_{ff} choice, we trained additional models at four d_{ff} values (5 seeds each) while holding all other hyperparameters fixed. Table 7 reports hard-sentence accuracy. At $d_{\text{ff}} = 72$, the oscillator is stable ($96.76 \pm 0.46\%$) while softmax shows bimodal outcomes ($91.07 \pm 10.03\%$), a +5.69 pp advantage. At $d_{\text{ff}} \leq 56$, both models encounter bimodal training outcomes within the 20-epoch horizon, consistent with the measure-zero initialization risk near the unstable equilibrium at severely constrained capacity. The oscillator advantage at the hardware-minimum ($d_{\text{ff}} = 64$, Table 2) is robust across 5 seeds.

WikiText-2 (Table 4). $d_{\text{model}} = 128$, $n_h = 4$, $n_\ell = 2$, $d_{\text{ff}} = 512$, causal masking, sinusoidal positional encoding. Word-level tokenization with vocabulary size 10K. Training: AdamW with weight decay 10^{-4}, $lr = 5 \times 10^{-4}$, batch 64, 30 epochs. Analytic-fixed-point training uses 5 seeds at each of $d_{\text{osc}} \in \{2, 4, 8, 16, 32\}$. The $d_{\text{osc}} = 2$ trained model is evaluated under three inference modes (analytic, RK45 random init, RK45 sequential init); RK45 random-init uses 5 random initialization seeds.

TinyStories (Table 4). Same architecture as WikiText-2. Word-level tokenization with vocabulary size 8K. Training: AdamW with weight decay 10^{-4}, $lr = 5 \times 10^{-4}$, batch 256, 5 epochs. All three $d_{\text{osc}} \in \{2, 8, 32\}$ values use 5 seeds.

Readout sharpening ablation (Table 5): TinyStories with same architecture as the headline runs in Table 4, varying d_{osc} and p. The d_{osc}=8 and d_{osc}=2 rows use 5 seeds. Same optimizer, learning rate, batch size, and tokenization as the main TinyStories runs.

ODE convergence verification (Table 6, Figure 7). RK45 via `scipy.solve_ivp`, $rtol = atol = 10^{-6}$, $T_{\max} = 30$, evaluated on 100 validation sequences from each trained TinyStories model with 5 random initializations of $z(0)$ per token. Convergence is measured by the final-time error $\text{err}_i = \|z_{\text{ODE}}(T_{\max}) - z_i^*\|_2$ between the integrated trajectory and the analytic fixed point. The "Frac < 0.01" column reports tokens with $\text{err}_i < 0.01$ (converged); "apparent-antipodal" tokens have $\text{err}_i > 0.1$ (slow convergence consistent with initialization near $-z_i^*$); "degenerate" tokens have $\|h_i\| < 0.01$ (small driving force).

Compute. All experiments were run on a Mac Studio (M4 Max) using PyTorch with the MPS backend. Antipodal validation and ODE convergence verification run on CPU.

C Confirming the dimensional bottleneck in causal language modeling

We tested three modifications that could plausibly close the oscillator–softmax perplexity gap without increasing d_{osc}. Each is consistent with the dimensional-bottleneck diagnosis of Section 4.2: none closes the gap independently of d_{osc}. All experiments use the WikiText-2 architecture (d_{model}=128, n_h=4, n_ℓ=2, d_{osc}=2) and identical optimizer/training-budget settings as the headline runs; ablations L3 and L4 report mean $\pm$ std over 5 seeds, and L5 also uses 5 seeds.

Learned position-dependent anchor phase offsets. The analytic fixed-point mechanism carries no explicit notion of token distance: the anchor r_j depends only on the content of token j, not on the offset $i-j$ from the query position. We added a learned rotation $\delta(i-j)$ applied to each anchor before computing the fixed point, exposing relative position directly to the coupling weights. Training the augmented model produced no significant change in PPL (-0.42 ± 0.25 relative to the d_{osc}=2 mean baseline). Position encoding already enters via the input embeddings; routing it through the anchor coupling adds no further capacity.

Learned per-head, per-position coupling amplification. A learned scalar $\beta_{h,i}$ multiplies the coupling magnitude for each (head, query-position) pair, allowing the model to soften or sharpen the effective coupling constant on a per-position basis. After training, $\beta_{h,i}$ spans the range $[0.10, 2.68]$ (mean 0.99 ± 0.32 across positions, seed 0), showing that the optimizer does identify positions that warrant distinct treatment. PPL is negligibly affected (-0.06 ± 0.44 over 5 seeds): the model recognizes per-position differences but a scalar amplification cannot enlarge the d_{osc}-dimensional anchor subspace that governs attention rank. The bottleneck is dimensional, not gain-related.

Attention-head and d_{model} scaling. Increasing n_h from 4 to 8 at fixed d_{model}=128 (so each head receives d_h=16 instead of 32) degraded oscillator attention (111.13 ± 0.72 vs. 110.22) with negligible effect on softmax (98.81 ± 0.35 vs. 99.58), consistent with the smaller per-head dimensionality. Scaling n_h and d_{model} together (8 heads at d_{model}=256) improved both architectures (oscillator: 104.42 ± 0.49, softmax: 92.49 ± 0.46, vs. 110.22 and 99.58 at baseline), but the oscillator–softmax gap did not close (11.93 PPL at the larger configuration vs. 10.64 at baseline).

D Code availability

Code, models, and demo will be released at a public GitHub URL upon acceptance.